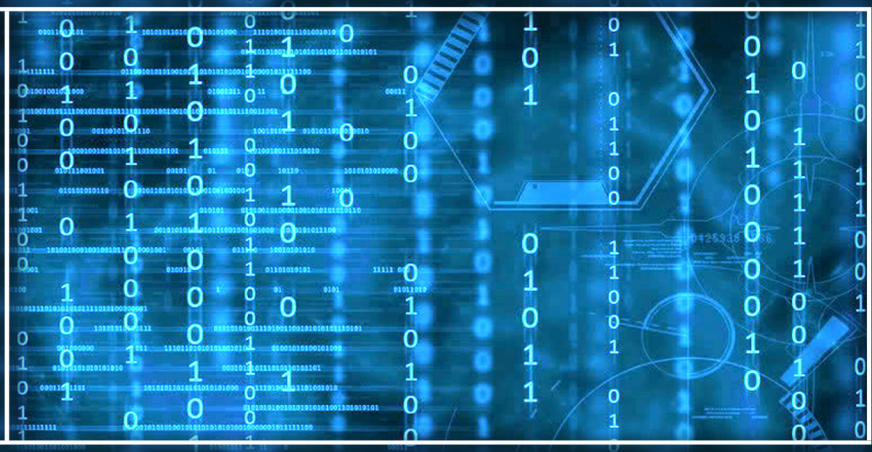
IJACSA
WHERE WISDOM SHARES
International Journal of Advanced Computer Science and Applications

Volume 7 Issue 4
April 2016



ISSN 2156-5570(Online)
ISSN 2158-107X(Print)

www.ijacsa.thesai.org

Editorial Preface

From the Desk of Managing Editor…

It may be difficult to imagine that almost half a century ago we used computers far less sophisticated than current home desktop computers to put a man on the moon. In that 50 year span, the field of computer science has exploded.

Computer science has opened new avenues for thought and experimentation. What began as a way to simplify the calculation process has given birth to technology once only imagined by the human mind. The ability to communicate and share ideas even though collaborators are half a world away and exploration of not just the stars above but the internal workings of the human genome are some of the ways that this field has moved at an exponential pace.

At the International Journal of Advanced Computer Science and Applications it is our mission to provide an outlet for quality research. We want to promote universal access and opportunities for the international scientific community to share and disseminate scientific and technical information.

We believe in spreading knowledge of computer science and its applications to all classes of audiences. That is why we deliver up-to-date, authoritative coverage and offer open access of all our articles. Our archives have served as a place to provoke philosophical, theoretical, and empirical ideas from some of the finest minds in the field.

We utilize the talents and experience of editor and reviewers working at Universities and Institutions from around the world. We would like to express our gratitude to all authors, whose research results have been published in our journal, as well as our referees for their in-depth evaluations. Our high standards are maintained through a double blind review process.

We hope that this edition of IJACSA inspires and entices you to submit your own contributions in upcoming issues. Thank you for sharing wisdom.

Thank you for Sharing Wisdom!

Managing Editor
IJACSA
Volume 7 Issue 4 April 2016
ISSN 2156-5570 (Online)
ISSN 2158-107X (Print)

Editorial Board

Reviewer Board Members

- **Aamir Shaikh**
- **Abbas Al-Ghaili**
  Mendeley
- **Abbas Karimi**
  Islamic Azad University Arak Branch
- **Abdelghni Lakehal**
  Université Abdelmalek Essaadi Faculté Polydisciplinaire de Larache Route de Rabat, Km 2 - Larache BP. 745 - Larache 92004. Maroc.
- **Abdul Razak**
- **Abdul Karim ABED**
- **Abdur Rashid Khan**
  Gomal Unversity
- **Abeer ELkorany**
  Faculty of computers and information, Cairo
- **ADEMOLA ADESINA**
  University of the Western Cape
- **Aderemi A. Atayero**
  Covenant University
- **Adi Maaita**
  ISRA UNIVERSITY
- **Adnan Ahmad**
- **Adrian Branga**
  Department of Mathematics and Informatics, Lucian Blaga University of Sibiu
- **agana Becejski-Vujaklija**
  University of Belgrade, Faculty of organizational
- **Ahmad Saifan**
  yarmouk university
- **Ahmed Boutejdar**
- **Ahmed AL-Jumaily**
  Ahlia University
- **Ahmed Nabih Zaki Rashed**
  Menoufia University
- **Ajantha Herath**
  Stockton University Galloway
- **Akbar Hossain**
- **Akram Belghith**
  University Of California, San Diego
- **Albert S**
  Kongu Engineering College
- **Alcinia Zita Sampaio**
  Technical University of Lisbon
- **Alexane Bouënard**
  Sensopia
- **ALI ALWAN**
  International Islamic University Malaysia
- **Ali Ismail Awad**
  Luleå University of Technology
- **Alicia Valdez**
- **Amin Shaqrah**
  Taibah University
- **Amirrudin Kamsin**
- **Amitava Biswas**
  Cisco Systems
- **Anand Nayyar**
  KCL Institute of Management and Technology, Jalandhar
- **Andi Wahju Rahardjo Emanuel**
  Maranatha Christian University
- **Anews Samraj**
  Mahendra Engineering College
- **Anirban Sarkar**
  National Institute of Technology, Durgapur
- **Anthony Isizoh**
  Nnamdi Azikiwe University, Awka, Nigeria
- **Antonio Formisano**
  University of Naples Federico II
- **Anuj Gupta**
  IKG Punjab Technical University
- **Anuranjan misra**
  Bhagwant Institute of Technology, Ghaziabad, India
- **Appasami Govindasamy**
- **Arash Habibi Lashkari**
  University Technology Malaysia(UTM)
- **Aree Mohammed**
  Directorate of IT/ University of Sulaimani
- **ARINDAM SARKAR**
  University of Kalyani, DST INSPIRE Fellow
- **Aris Skander**
  Constantine 1 University
- **Ashok Matani**
  Government College of Engg, Amravati
- **Ashraf Owis**
  Cairo University
- **Asoke Nath**

CONTENTS

Novel Altered Region for Biomarker Discovery in Hepatocellular Carcinoma (HCC) Using Whole Genome SNP Array

Novel cytogenetic aberration for hepatocellular carcinoma

Esraa M. Hashem[1]
[1]Biomedical Engineering Department
Misr University for Science and Technology (MUST University)
Cairo, Egypt

Mai S. Mabrouk[2*]
[2]Biomedical Engineering Department
Misr University for Science and Technology (MUST University)
Cairo, Egypt

Ayman M. Eldeib[3]
[3]Systems and Biomedical Engineering
Faculty of engineering, Cairo University, Cairo, Egypt

***Abstract*—cancer represents one of the greatest medical causes of mortality. The majority of Hepatocellular carcinoma arises from the accumulation of genetic abnormalities, and possibly induced by exterior etiological factors especially HCV and HBV infections. There is a need for new tools to analysis the large sum of data to present relevant genetic changes that may be critical for both understanding how cancers develop and determining how they could ultimately be treated. Gene expression profiling may lead to new biomarkers that may help develop diagnostic accuracy for detecting Hepatocellular carcinoma. In this work, statistical technique (discrete stationary wavelet transform) for detection of copy number alternations to analysis high-density single-nucleotide polymorphism array of 30 cell lines on specific chromosomes, which are frequently detected in Hepatocellular carcinoma have been proposed. The results demonstrate the feasibility of whole-genome fine mapping of copy number alternations via high-density single-nucleotide polymorphism genotyping, Results revealed that a novel altered chromosomal region is discovered; region amplification (4q22.1) have been detected in 22 out of 30-Hepatocellular carcinoma cell lines (73%). This region strike, *AFF1 and DSPP*, tumor suppressor genes. This finding has not previously reported to be involved in liver carcinogenesis; it can be used to discover a new HCC biomarker, which helps in a better understanding of hepatocellular carcinoma.**

Keywords*—Hepatocellular carcinoma; copy number alternation; biomarkers; single-nucleotide polymorphism

I. INTRODUCTION

Cancer begins when cells in a part of the body begin to grow out of control. There are many types of cancer, but they all starts duo to out-of-control growth of abnormal cells. In many cases, the cancer cells form a tumor. Liver cirrhosis is the end stage of chronic liver diseases. Cirrhosis is a disease in which hepatic cells become damaged. People with cirrhosis have an increased risk of liver cancer. The main causes of cirrhosis are alcoholic liver disease (ALD), hepatitis B (HBV), hepatitis C (HCV), non-alcoholic steatohepatitis (NASH), 10 haemochromatosis, auto-immune hepatitis (AIH), primary biliary cirrhosis (PBC) and primary sclerosing cholangitis (PSC) [1], as shown in Fig. 1. Chronic HCV infection is the main reasons for cirrhosis and HCC [2]. Rates for liver cancers in men are usually 2 to 4 times higher than in women [3]. The most common primary malignancy of the liver is Hepatocellular carcinoma (HCC), it accounts for as many as one million deaths annually around the world [4].

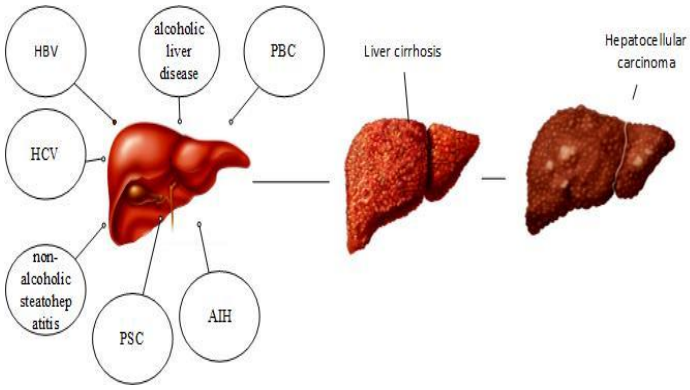


Fig. 1. The main causes of liver cirrhosis

The liver is a unique organ in that it can regenerate up to 70 to 80 % after resection. As liver transplantation is the only available treatment for HCC. The genomic signatures of HCC can help to characterize the molecular changes responsible for its development. In cancer, the genomes of affected cells are rearranged in complex or even unpredictable ways, single-nucleotide polymorphism arrays (SNP) considered as new physical detection technologies employed to identify chromosomal gain, loss, deletion and amplification to detect known point mutations. These detection methods measure several hundred thousand sites throughout the genome at the same time, and when use these detection methods in high-throughput to measure thousands of samples, it's generate terabytes of data per experiment. This technology helps to understand the underlying genetic alterations in HCC, aid in molecular classification of HCC and patient prognosis. Gene expression profiling has provided important insights into the biology of HCC. A better understanding of gene expression in HCC may help to understand the pathophysiology of HCC and improved prognostication.

One of the characteristics in cancer is chromosomal instability it results in the structural alterations of DNA copy number alterations (CNAs) and it is expressed as loss, gain, amplifications and deletions, which are frequently contributing to tumorigenesis. That alterations change the level of gene expression, which modify normal growth control and survival pathway [5]; identification of cancer specific CNAs will not only provide new insight into understanding the molecular basis of tumorigenesis but also facilitates the discovery of new cancer genes. Among such alternations, frequently detected of DNA copy number gains at 1q, 8q, 11q, 17q, and 20q [6-8] and losses at 4q, 8p, 9p, 13q, 16q, and 17p have been identified in HCC [9-11] using traditional methodology.

Early efforts focus on discover and categories CNAs depend on a technique called comparative genomic hybridization (CGH) which used to provide a genome-wide investigation of CNAs in cancer. The disadvantage of CGH method is the resolution of mapping copy number alteration limited by 2 Mbp (high- copy number amplifications) to 10 Mbp (low-copy-number amplification or deletion) [12].

Through the uses of DNA clones spotted in array format on slides as targets for hybridization of normal and test DNA array comparative genomic hybridization (a-CGH) opened the way for a marked increase in resolution up to 30 kb [13-14] however, the main drawback of a-CGH its difficult to detect aberrations that do not result in copy number changes and is limited in detect mosaicism [15].

During the last decade, the whole genome single nucleotide polymorphism (SNP) array has been the common element in a highly productive synergistic relationship between advances in biological understanding and computational methodologies. SNP is a single base change in a DNA sequence, with a usual alternative of two possible nucleotides at a specific position. The Illumina Bead Array has gradually increased in capacity over the years from 100 000 SNPs (Human-1) to the current (HumanHap1M) one million, with intermediary steps 240- 000, 317 000, 550 000 and 650- 000 [16].

The use of SNP arrays in copy number detection has a number of advantages. Instead of the two applications for the data that are SNP genotyping and copy number analysis, other aspects promote their use more than other techniques. SNP arrays work with fewer samples per experiment compared to a-CGH. The SNP array is a cost-effective technique, which allows the user to increase the number of samples examined on a limited budget. Although the progress in high-throughput sequencing technology has made copy number discovery much easier, the application of known CNA information means that we can target structural variation in a sample using less expensive techniques such as the SNP array without a large reduction in genome- wide coverage [17].

There are many automated algorithms used to determine the characteristic of genomic profiles, Justin K. et al. use Hidden Markov Models (HMM) algorithm to demonstrate that highly accurate SNP genotypes can infer from very low coverage shotgun [18]. Cooper et al. use SNP conditional mixture modelling (SCIMM) by applied a mixture-likelihood clustering method within the R statistical package to identify deletions of copy number changes [19]. Franke et al. present a combined approach focus on single SNP interpretation; Tri-typer uses maximum likelihood estimation to detect deletions in Illumina SNP data in unrelated samples [20].

Signal noise often causes false positive predictions and it is a strong limitation of many automated algorithms to detect biomarkers in CNAs [21]. To fill this draw back, segmentation method is implemented based on wavelet decomposition and threshold in which detects significant breakpoints in the data called: Discrete Stationary Wavelet transforms (DSWT), that helps to understanding of the genomic basis of the disease process and develop a method to find biomarkers for early diagnosis of HCC by, identification of genome-wide alternations in copy number from whole genome SNP genotyping. The normalized R-value is uses as a representation of intensity on individual SNP plots. The log R ratio value can calculate from the expected normalized intensity of a sample and observed normalized intensity, log R ratio used by a number of copy number event detection algorithm [17].

In the work at hands, one-dimensional DSWT model is proposed to analyze 30- HCC cell line SNP array data to identify gain, amplification, deletion and loss of chromosomes and try to predict new biomarkers that help to pre-diagnose of HCC.

II. Material and Method

SNP array plays a key role in genome-wide association and population genetic studies, which are the most common genetic variants in the human genome [21]. Wavelet analysis unlike traditional Fourier transform (FT), it is able to decay time series into time-frequency space and gets more attention as a potential tool to study cancer genomic data.

To remove noise from the signal by using DSWT algorithm First, load the data 30 cell lines, then decompose the signal at single-level, then estimate the approximation and detail sub bands to generate the signal using inverse stationary wavelet transform, finally remove the noise and display denoised signal with ideogram of a specific chromosome. The steps of the proposed algorithm summarized in Fig.2

A. SNP Genotyping Array Data:

The data used in this study is SNP Genotyping Array, performed on the Human Omni1-Quad Bead Chip by Illumina Fast Track Services (Illumina, San Diego, CA); Thirty HCC cell lines, processed according to the manufacturers' instructions. Raw data processed using an in-house pipeline analysis tool to obtain copy number segments [22]. Using HCC cell- lines of specific human chromosomes 1, 4, 8,9,11, 13, 16, 17 and 20 to compare the normal versus tumor DNA (hepatocellular carcinoma). It provides a whole-genome genotyping microarray (WGGT) screening of DNA-copy number changes. For each SNP array, its two alleles defined as the A and B alleles using a set of specific naming rules. The raw signal intensity values measured for the A and B alleles are then submitted to a five-step normalization procedure using the signal intensity of all SNPs. This step produces the X and Y values for each SNP array, representing the

experiment overall normalized signal intensity on the A and B alleles, respectively. As normalization measure of the total signal intensity, the log R Ratio (LRR) value for each SNP is calculated as LRR = log2 (R observed/R expected), where (R expected) is computed from linear interpolation of accepted genotype clusters [23]. LRR of Zero means copy number neutral, positive values mean copy number gains and negative values mean copy number losses. The data of normalized log R based intensity ratios was stored in Excel files and are available at: [http://www.ncbi.nlm.nih.gov/geo].

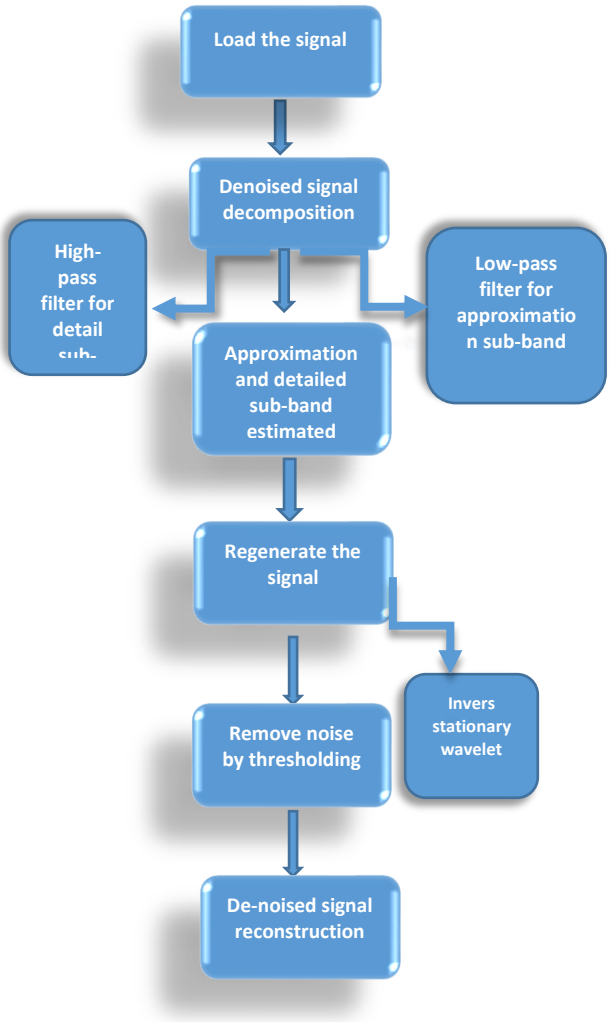


Fig. 2. The steps of discrete stationary wavelet algorithm

B. DISCRETE stationary wavelet TRANSFORM (DSWT):

Hepatocarcinogenesis is a complex multi-state process occurring usually after many years of chronic hepatitis providing mutagenic environments to precipitate random genetic alterations.

With the rapid development of high-density single-nucleotide polymorphism array and array-based comparative genomic hybridization, it has become feasible to characterize CNAs involved in tumor development and progression across the entire genome. Advances in SNP genotyping technologies have played a key role in the proliferation of large-scale genomic studies, leading to the discovery of hundreds of genes associated with complex human diseases (HCC). Wavelet bases are generated by dilation and translation of a single wavelet function (x) called the mother wavelet (*Haar basis).* It is simplest wavelet base, which generated by the *Haar* function in equation 1, it identifies statistically significant breakpoints in the data, using the maxima of the Haar wavelet transform, and segments accordingly [5]. The advantage of wavelets is their excitability.

$$\psi(x) = \begin{cases} -1/\sqrt{2} & 1 < x \leq 0 \\ 1/\sqrt{2} & 0 < x \leq 1 \\ 0 & otherwise \end{cases} \quad (1)$$

The family of dyadic dilations and translations $\{\psi_{jk}\}_{j,k\in z}$ where ψ jk(x) = $2^{j\backslash 2}$ ψ (2^{j}x-k) forms an orthonormal(two vectors in an inner product space are orthonormal if they are orthogonal and all of unit length) basis for the space of L^2 functions, where j and k are indices for the scale and location, respectively [24]. After using Haar function, there is a need to remove noise from the signal by decomposing, the input signal into an approximation sub-bands and a set of detail sub-bands at different resolution scales using a set of high pass and low pass filters [25]. The number of samples contained by each sub-band at level N, so the number of input samples divided by 2^N [26]. Sub-band coding found to yield a fast computation of wavelet transform, in the work at hand, only one level of approximation and detail sub-bands need to implement, then regenerate the signal using inverse stationary wavelet transform, after that, remove the noise by threshold selection rule to display denoised signals.

Wavelet thresholding is very simple non-linear technique, which operates on one wavelet coefficient at same time. In its simplest form, each coefficient is threshold by compare against threshold, if the coefficient is smaller than threshold, set to 0, otherwise it is modified or kept. There are two thresholds frequently used, hard threshold, and soft threshold. The soft-thresholding function has a somewhat different rule from the hard-thresholding function. It shrinks the wavelet coefficients by λ towards zero in equation 2, the threshold λ chosen according to the signal energy and the noise variance [27].

$$\begin{aligned} f(x) = x - \lambda \quad & \text{if } x \geq \lambda \\ =0 \quad & \text{if } x < \lambda \\ = x + \lambda \quad & \text{if } x \leq -\lambda \end{aligned} \quad (2)$$

Finally, to display the denoised signal with specific chromosome the ideogram added to the denoised signal. A threshold level assigned to each dyadic resolution level according to the principle of minimizing the Stein Unbiased Estimate of Risk (SURE) for estimate the threshold. SURE provides a means for unbiased estimation of the true mean-squared error (MSE). Without the need knowledge of the noise-free signal, this unbiased evaluate solely depends on the given data and on some description of the first-order dependence of the denoising operator with respect to the data [5].

III. RESULTS

The most common adult genetic liver disease in which a particular genetic defect leads to iron accumulation in the

liver, leading to liver cirrhosis and liver cancer in some cases. Liver cancer is one of the few cancers that is increasing in incidence and in mortality. It is difficult to find HCC early because signs and symptoms often do not appear until it is in its later stages [28].

Comprehensive identification of CNVs is required to provide a complete view of human genetic variation, which frequently detected during the development of HCC. Mapping of chromosome gains, amplifications, deletions and losses have frequently resulted in the identification of oncogenes and tumor suppressors gens. Recently, there has been increased interest in inferring copy number from SNP arrays due to their widespread use in genome-wide association studies and significantly greater probe density.

In this study, the genome-wide CNAs have been systematically analyzed on specific chromosomes(1q,4q,4p,8q,8p,9q,11q,13q,16p,17q,17p,20q) by applying discrete stationary wavelet transform technique on HCC cell lines data using high-density SNP-array. This allowed us to identify regions in the HCC genome that have undergone recurrent high-level amplifications or deletions. The 30 HCC cell lines data can download from GEO with accession number [GSE38207].

Figure.3 shows the coefficients of approximation and detail sub-band at level (1) in chromosome 4, sample (GSM936756-39376). Figure 4 displays the original and denoised signal using the soft threshold rule for chromosome 4. By performing DSWT algorithm in HCC cell lines using SNP array, it has been noticed that the most commonly gained region of chromosome 1q22.1–23.1 in 17 out of 30-hepatocellular carcinoma (57%), Fig.5 gains of 1q are the most frequent aberrations and occur early during tumorigenesis [29]. This region contains the tumor suppressors gens JTB, SHC1, CCT3and COPA [30].

A region of 4q25-26 losses in 30 out of 30-hepatocellular carcinoma (100%) Fig.6. The allelic loss of chromosome 4q was significantly associated with hepatocellular carcinoma having elevated serum AFP [31]. Gain of 8q observed in 6 out of 30-hepatocellular carcinoma (20%), Fig. 7 a loss of 8p was detected in 16 out of 30-HCC (53%) Fig.8 loss for 9p was observed in 30 out of 30-HCC (100%), Figu.9 gain of 11q was observed in 15 out of 30-HCC (50%), Fig.10 loss of 13q was observed in 14 out of 30 HCC (47%), Fig. 11 loss of 16q was observed in 18 out of 30-HCC (60%), Fig. 12 gain of 17q was detected in 17 out of 30-HCC (57%), deletion of 17p was observed in 29 out of 30-HCC (97%), Fig. 13 loss of 17p is one of the most frequent chromosomal aberration in primary hepatocellular carcinoma [32]. It has may affect the tumor suppressor gene TP53. TP53 is responsible for regulation of the cell cycle at G1/S and G2/M interfaces, as well as induction of apoptosis in response to severe damage to cellular DNA [33]. Gain of 20q was observed in 10 out of 30-HCC (33%), Fig. 14. All the results summarized in table 1.

Mai.S.Mabrouk et al [5], applied discrete stationary wavelet transform technique to a number of human chromosomes for analyzing array-CGH data to evaluate the prognosis of 67 HCC patients. The results show that a gain of 1q detected in 55% and a gain of 20q detected in 67% of HCC cases. In addition, a loss of 4q detected in 67%, a loss and gain of 8q, 8p detected in 87%, a loss of 13q detected in 72%, loss in 16q detected in 75%, and loss of 17q detected in 33% of HCC. Comparing the results obtained in this work with previous study it has been found that, high-density SNP array have better resolution and detect copy number variation regions better than array-CGH. That can help for mapping the whole genome of CNAs.

Performing DSWT analysis shows that there is an important observation for a novel altered region that detected an amplification region of *4q22.1* in 22 out of 30 HCC cell lines (73%) of AFF1 and DSPP genes, as shown in Figure 13, this finding have not previously reported to be involved in liver carcinogenesis. This observation may lead to discover new biomarker for HCC patients, or may be associated with other types of cancer [34]. It should be considered for future work on detecting copy number alternation regions.

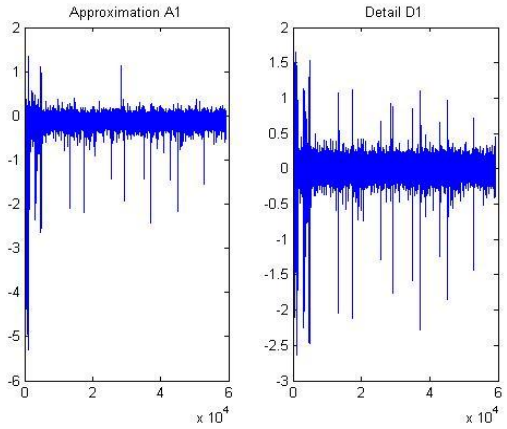


Fig. 3. The coefficients of approximation and detail sub band

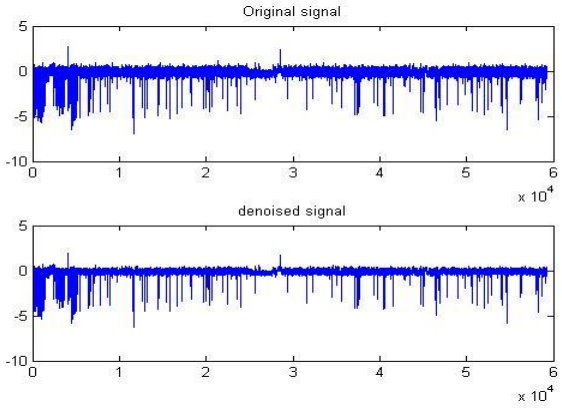


Fig. 4. The original and denoised signals of chromosome 4

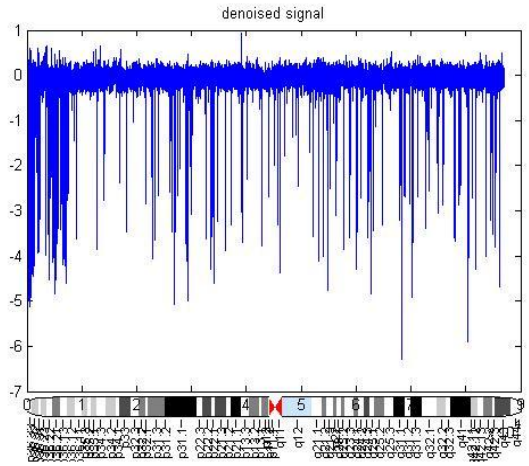


Fig. 5. Gain of chromosome 1q21-23

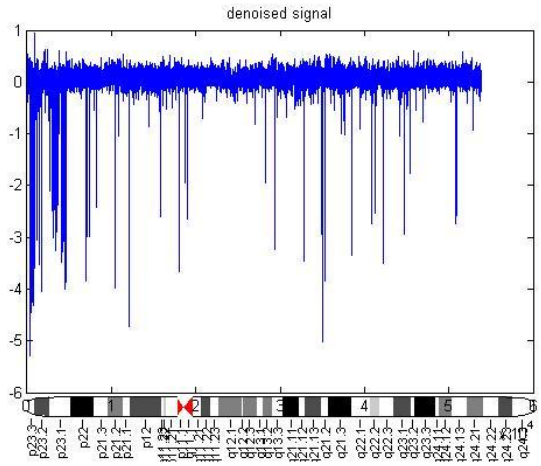


Fig. 8. loss of chromosome 8p21.3

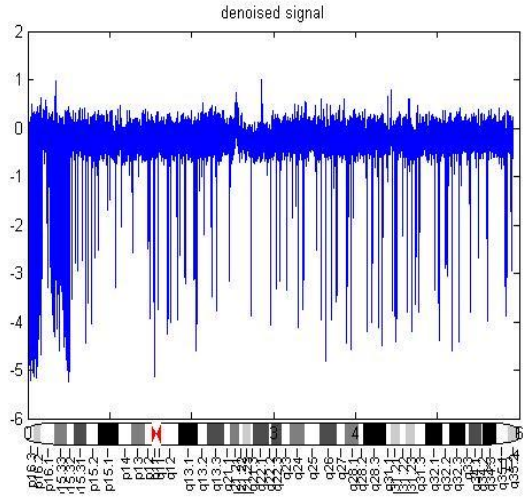


Fig. 6. Loss of chromosome 4q25-26

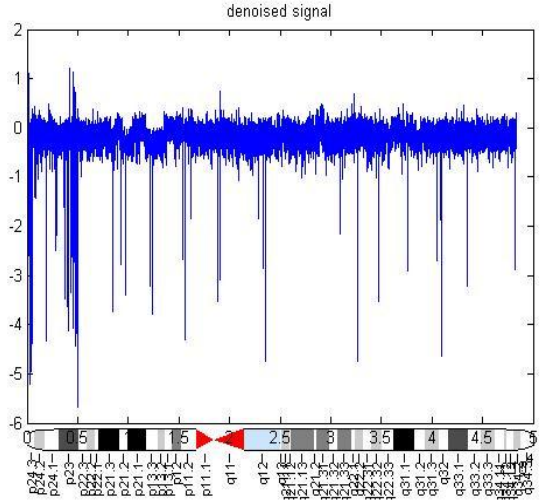


Fig. 9. loss of chromosome 9q21.3-23

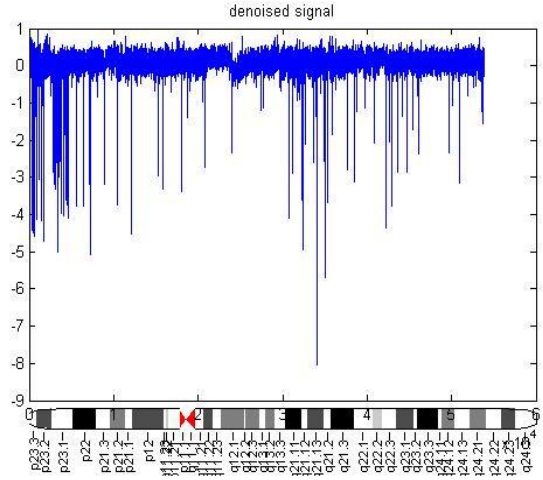


Fig. 7. Gain of chromosome 8q24.12

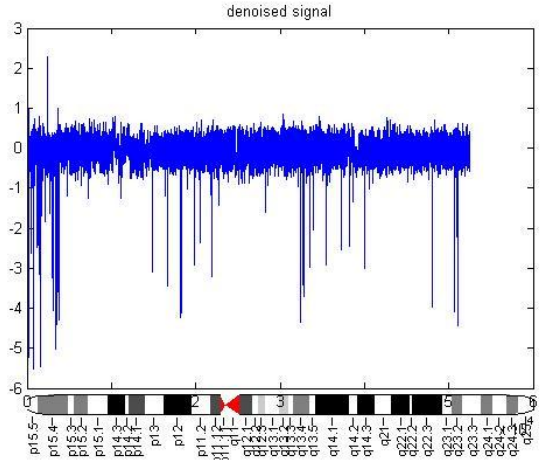


Fig. 10. Gain of chromosome 11q13.2-13.3

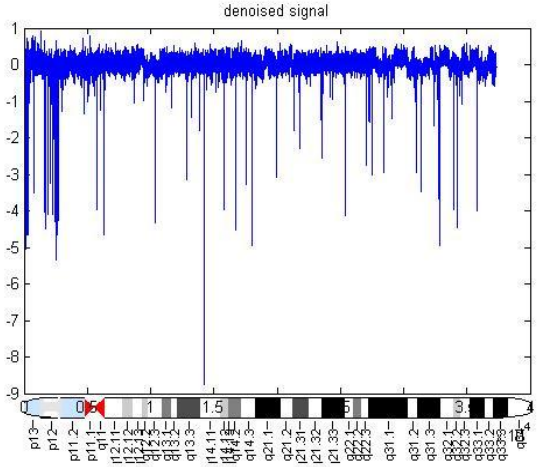


Fig. 11. loss of chromosome 13q12.11

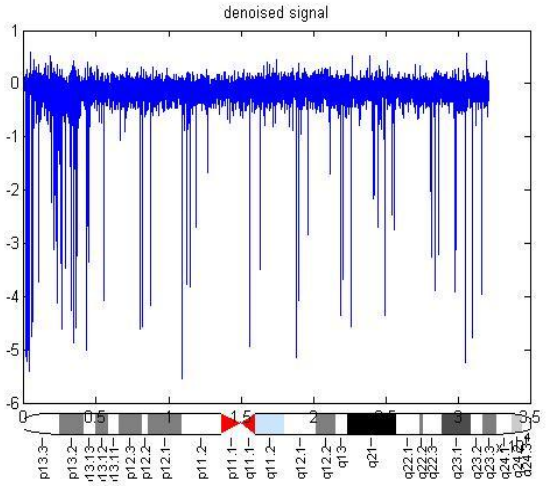


Fig. 12. loss of chromosome 16q11.2

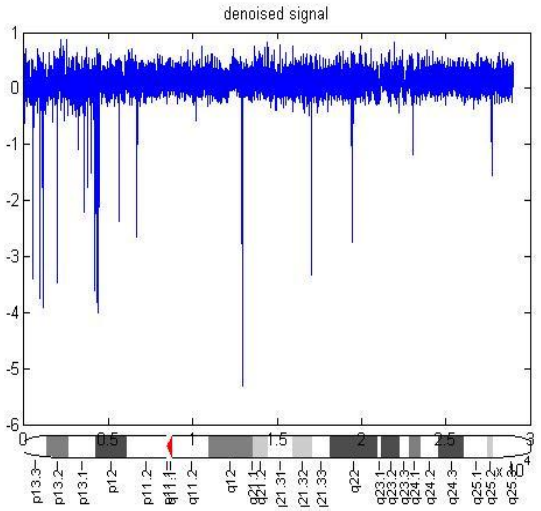


Fig. 13. Gain and loss of chromosome 17q12,17p13.1

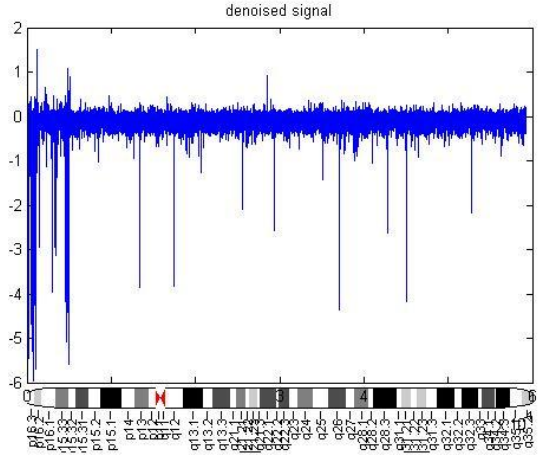


Fig. 14. Gain of chromosome 4q22.1

TABLE I. RESULTS OF EACH ABNORMAL CHROMOSOME

CHROMOSOME CNAS TYPE	CYTOBAND	PERCENTAGE
GAIN	1q21-23	57%
GAIN	4q22.1	73%
GAIN	8q24.12	20%
GAIN	11q13.2-13.3	50%
GAIN	17q12-21	57%
GAIN	20q12	33%
LOSS	4q25-26	100%
LOSS	8p21.3	53%
LOSS	9p21.3	100%
LOSS	13q12.11	47%
LOSS	16q11.2	60%
LOSS	17p13.1	97%

IV. CONCLUSIONS

The prognosis for patients with hepatocellular carcinoma is poor due to, the low chance of therapeutic treatment. The large majority of HCC cases happened in those with chronic liver disease, particularly HBV and HCV. To date, the exact mechanism of hepatocarcinogens is still remaining unclear. Many disease states, different types of cancer, can be better understood by discovering tumor biomarkers. These markers not only help in prediction of prognosis or recurrence but can also help in deciding appropriate modality of therapy and may represent novel potential targets for therapeutic interventions

Genetic association studies have been a popular process for evaluating the association between common SNP and complex diseases. However, the recent development of technologies for gene expression profiling and genome-wide copy number analysis has allowed the comprehensive characterization of entire cancer genomes. This information helps to improve our understanding of the genetic and epigenetic alterations driving the development of liver cancer, instead of ultimately provide guidance for personalized treatment of patients. Clearly, there is a persistent need for novel clinical HCC biomarkers that have a sufficient specificity and sensitivity to aid in the early diagnosis of HCC. A common drawback of previous techniques is the long running time required to segment real high-density arrays and the resolution, it is clearly identifying when comparing our

result with a-CGH data. DNA-SNP array studies aim to identify novel serological markers, by targeting genes that characteristically highly expressed in HCC tissues.

This study demonstrates that genomic high-density SNP-array test could be used to detect and characterize genome-wide copy number alterations on more than one thousand clones arising from several chromosomes. By applying one-dimensional discrete stationary wavelet transform technique on 30 HCC cell lines, it has been found that there is a certain number of chromosome aberration including gains of 1q,8q,11q,17q,20q chromosomes and losses of 4q,8p,13q,16q,17p. A new observation region has been altered for amplification of chromosome 4 region *4q22.1*. Genes expression profiling DSPP and SPP1 can lead to discover new biomarkers that may help improve diagnostic for detect HCC, represent novel targets for therapeutic agents, and allow us to identify regions in the HCC genome that have undergone frequent high-level focal amplifications or deletions.

REFERENCES

[1] N.I. Guha, J.P. Iredale, ''Clinical and diagnostic aspects of cirrhosis. In: Textbook of hepatology from basic science to clinical practice. 3rd edition. Edited by: Rodés J, Benhamou J-P, Blei A, Reichen J, Rizzetto M. Oxford'', Blackwell Publishing, 2007, pp.604-622.

[2] E.S.Hashim, M.S.Mabrouk, ''A study of support vector machine algorithm for liver disease diagnosis'', American Journal of Intelligent Systems, 2014, vol.4(1),pp. 9-14

[3] F.X. Bosch, J. Ribes, M. Díaz, R. Cléries, ''Primary liver cancer: Worldwide incidence and trends'' Gastroenterology, 2004, vol.127, pp.5:16.

[4] A. S. Befeler and A. M. Bisceglie, '' Hepatocellular Carcinoma: Diagnosis and Treatment'', GASTROENTEROLOGY, 2002, vol.122, pp.1609-1619.

[5] M.S.Mabrouk, E.S.Hashim, A.Shrawy, '' Discrete Stationary Wavelet Transform of Array CGH Data for Biomarkers Identification of Hepatocellular Carcinoma'', 2012, vol. 1 (2), pp.148-154.

[6] R. Beroukhim, CH. Mermel, D. Porter, G. Wei, S. Raychaudhuri, J. Donovan, et al. ''The landscape of somatic copy-number alteration across human cancers'',Nature,2010, vol.463, pp.899-905.

[7] S.A. Lee, C .Ho, R. Roy, C .Kosinski, M.A Patil, A.D. Tward, et al. ''Integration of genomic analysis and in vivo transfection to identify sprouty 2 as a candidate tumor suppressor in liver cancer.'', HEPATOLOGY 2008, vol.47, pp.1200-1210.

[8] Y. Tanaka, F .Kanai , M .Tada, R .Tateishi M. Sanada M, Y. Nannya ,''Gain of GRHL2 is associated with early recurance of hepatocellular carcinoma''J Hepatol. 2008, vol.49 (5), pp.746-57.

[9] Y. Midorikawa, W. Tang, Y .Sugiyama, ''High-resolution mapping of copy number aberrations and identification of target genes in hepatocellular carcinoma'', Biosci Trends 2007; 1:26-32.

[10] M.A Patil, I.Gutgemann, J. Zhang , C. Ho, S.T. Cheung, D.Ginzinger, et al. ''Array-based comparative genomic hybridization reveals recurrent chromosomal aberrations and Jab1 as a potential target for 8q gain in hepatocellular carcinoma.'', Carcinogenesis, 2005,vol.26,pp.2050-2057.

[11] T.C.Chen ,L.L. Hsieh , T.T. Kuo , K.F. Ng , W.C. YH, et.al, ''p16INK4 gene mutation and allelic loss of chromosome 9p21-22 in Taiwanese hepatocellular carcinoma'' Anticancer Res. 2000, vol.20(3A),pp.1621-6.

[12] A.Kallioniemi, O.P Kallioniemi, D .Sudar, D. Rutovitz, J.W Gray, F. Waldman, et al. ''Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors.'', Science 1992, vol.258, pp.818-821.

[13] B.P Coe,B. Ylstra , B.Carvalho,G.A. Meijer, et al. ''Resolving the resolution of array CGH.'' Genomics 2007, vol.89, pp.647–653.

[14]Mai.S.M, Esraa. M. H, Amr. Sh. ''Statistical Approaches for Hepatocellular Carcinoma (HCC) Biomarker Discovery.'' American Journal of Bioinformatics Research, 2012, vol.2 (6), pp.102-109.

[15] A.E. Ostlander, G.A. Meijer, B. Ylstra, ''Microarray-based comparative genomic hybridization and its applications in human genetics''. Clin Gene, 2004, vol.66, pp.488-495.

[16] T.L.Framboise, '' Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances'', Nucleic Acids Research, 2009, Vol. 37(13), pp. 4181–4193.

[17] L.Winchester, C. Yau and Jiannis Ragoussis, '' Comparing CNVdetection methods for SNP arrays'', BRIEFINGS IN FUNCTIONAL GENOMICS AND PROTEOMICS, 2009, VOL 8(5), pp.353:366.

[18] J.Kennedy, '' Efficient Algorithms for SNP Genotype Data Analysis using Hidden Markov Models of Haplotype Diversity'', M.Sc., Rensselaer, Hartford, CT, USA, 2002.

[19] G.M. Cooper, T. Zerr, J.M. Kidd, et al. ''Systematic assessment of copy number variant detection via genome-wide SNP genotyping.'' Nat Genet 2008, vol.40 (10), pp.1199–203.

[20] L.Franke, C.G. Kovel, Y.S Aulchenko, et al. ''Detection, imputation, and association analysis of small deletions and null alleles on oligonucleotide arrays.'' AmJ Hum Genet, 2008, vol.82 (6), pp.1316–33.

[21] L. Kruglyak, D.A. Nickerson,''Variation is the spice of life'',Nat. Genet., 27 (2001), pp. 234–236.

[22] https://icom.illumina.com/ icom/software.ilmn, 2015.

[23] K. Wang,M. Li, D. Hadley, R. Liu, J. Glessner, S. F. Grant, H. Hakonarson, and M. Bucan, 'PennCNV: An integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data'', Cold Spring Harbor Laboratory Press,2007, vol.17, pp. :1665–1674.

[24] R.S. Stankovic, B.J.Falkowski, '' The Haar wavelet transform: its status and achievements'', Computers and Electrical Engineering, 2003, vol. 29, pp.25:44.

[25] E.S.Hashim, M.S.Mabrouk, ''Impact of parallel computing on identifying biomarkers of hepatocellular carcinoma'', J. Med. Imaging Health Inf, 2014, vol.4, pp. 1–5.

[26] Nha Nguyen, S. Oraintara, H.Huang,and Y. Wang, '' Denoising of Array-Based DNA Copy Number Data Using The Dual-tree Complex Wavelet Transform'' International Conference on Bioinformatics and Bioengineering,2007, pp. 137 – 144.

[27] P. Hedaoo1 and S.S Godbo, '' wavelet thresholding approach for image denoising'', International Journal of Network Security & Its Applications (IJNSA), Vol.3, No.4, July 2011.

[28] E. M.Hashim, M. S. Mabrouk, A.Sharawy, ''Circular Binary Segmentation Modeling of Array CGH Data on Hepatocellular carcinoma.'', Radio Science Conference (NRSC), 29th National, 2012, pp. 667 – 674.

[29] A. Subramanian, P. Tamayo, V.K. Mootha, et al. ''Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.'', Proc Natl Acad Sci USA, 2005, vol.102, pp.15545–15550.

[30] Y .Wang, M.C. Wu, J.S. Sham, W .Zhang, W.Q .Wu, and X.Y. Guan. "Prognostic significance of c-myc and AIB1 amplification in hepatocellular carcinoma. A broad survey using high-throughput tissue microarray.'' Cancer, 2002, vol.95, pp.2346–52.

[31] Q. Deng, S. Huang. "PRDM5 is silenced in human cancers and has growth suppressive activities.'', Oncogene, 2004, vol.23, pp.4903–10.

[32] RS. Cornelis, M.v. Vliet, CB. Vos, AM. Cleton-Jansen, MJ .van de Vijver, JL .Peterse, et al. "Evidence for a gene on 17p13.3, distal to TP53, as a target for allele loss in breast tumors without p53 mutations.'' Cancer Res1994, vol.54, pp. 4200–6.

[33] A. Singhal , M.Jayaraman , D.N. Dhanasekaran , V. Kohli, '' Molecular and serum markers in hepatocellular carcinoma: Predictive tools for prognosis and recurrence'', Critical Reviews in Oncology/Hematology,2012, vol. 82 ,pp.116–140.

[34] R.Hill , RK .Kalathur , L .Colaço , R. Brandão , S .Ugurel , M. Futschik , ''TRIB2 as a biomarker for diagnosis and progression of melanoma.'' Carcinogenesis. 2015, vol.36 (4), pp. 469-77.

A Framework for Satellite Image Enhancement Using Quantum Genetic and Weighted IHS+Wavelet Fusion Method

Amal A. HAMED
National Authority for Remote Sensing and Space Science
Aswan, Egypt

Osama A. OMER
Electrical Engineering Department
Aswan University, 81542
Aswan, Egypt

Usama S. MOHAMED
Electrical Engineering Department
Assiut University
Assiut, Egypt

***Abstract*—this paper examined the applicability of quantum genetic algorithms to solve optimization problems posed by satellite image enhancement techniques, particularly super-resolution, and fusion. We introduce a framework starting from reconstructing the higher-resolution panchromatic image by using the subpixel-shifts between a set of lower-resolution images (registration), then interpolation, restoration, till using the higher-resolution image in pan-sharpening a multispectral image by weighted IHS+Wavelet fusion technique. For successful super-resolution, accurate image registration should be achieved by optimal estimation of subpixel-shifts. Optimal-parameters blind restoration and interpolation should be performed for the optimal quality higher-resolution image. There is a trade-off between spatial and spectral enhancement in image fusion; it is difficult for the existing methods to do the best in both aspects. The objective here is to achieve all combined requirements with optimal fusion weights, and use the parameters constraints to direct the optimization process. QGA is used to estimate the optimal parameters needed for each mathematic model in this framework "Super-resolution and fusion." The simulation results show that the QGA-based method can be used successfully to estimate automatically the approaching parameters which need the maximal accuracy, and achieve higher quality and efficient convergence rate more than the corresponding conventional GA-based and the classic computational methods.**

***Keywords*—Quantum genetic algorithm (QGA); HIS; fusion; wavelet; registration; super-resolution**

I. INTRODUCTION

Image fusion is the process of merging two or more images obtained from two or more sensors for the same scene. The objective is to extract more information from the fused image than information in individual images. In satellite images, the low resolution (LR) multispectral (MS) image are merged with the high resolution (HR) panchromatic (pan) image to obtain the MS HR image by using fusion technique [1]. In the recent years, many fusion methods such as Intensity Hue Saturation (IHS) transform, High Pass Filtering (HPF) method [2], Laplacian pyramid [3] and wavelet transform have been proposed. The stand-alone IHS method is the most commonly used fusion technique because it can convert a standard RGB (Red, Green, Blue) image into (I), (H) and (S) components, the transform, and inverse equations mentioned in [4]. This color space has the advantage of the human beings visual system in which I, H, and S components considered as roughly orthogonal perceptual axes. Therefore, it can add the spectral and spatial information smoothly for satellite images with overlapping spectral sensitivity between pan image and MS image. But this method has a disadvantage that the color quality of the fused image strongly depends on the similarity between the HR pan image and the intensity image (I) of the MS LR image [5]. The gray value distribution of the intensity of the IHS image should be close enough to that of the pan image to preserve the spectral information [6]. However, the difference between the intensity image and the pan image causes a major spectral color distortion. Among the existing fusion methods, wavelet transform based method. It has the advantage of qualified localization in both space and frequency domains [7]. The stand-alone wavelet fusion outperforms other conventional (conv.) fusion techniques, such as IHS, Principal Component Analysis (PCA) in preserving spectral information [8] because it usually injects the high spatial details from the HR image into all three low spatial resolution MS bands. However, these high spatial information in the HR image have gray values different from that of an MS band. This difference may cause some spectral distortion in the wavelet-fused image, the combination between color and spatial details appear unnatural [9]. To better employ the advantages of IHS and wavelet fusion methods and to get over the shortcomings of the two stand-alone methods, researchers has introduced an IHS and wavelet integration fusion in previous work; explained in [10]. In general, it uses the IHS transform to integrate the spectral information of LR MS with the spatial detail information of HR pan to achieve a smooth combination of spectral and spatial information, while wavelet transform is utilized to generate a new image "New Intensity" that has a similar gray values distribution to "I" component of MS and contains the high spatial details of the pan. As illustrated in Fig. 1, the process steps of this method is explained (before image fusion, the MS image is resampled to have the same pixel size as the HR pan by using the cubic interpolation). The integrated IHS with wavelet transforms produces efficient results than either standard methods or stand-alone wavelet-based methods [11]. However, the trade-off is higher complexity and cost [12]. On the other hand, the HR pan image can be reconstructed from multi LR images by applying super-resolution techniques utilizing the subpixel shifts between them. The quality of the reconstructed image depends on the accuracy of subpixel shifts estimation process. Image restoration and interpolation techniques are implemented to

obtain the estimated HR pan image [13]. Several techniques have been introduced in many studies such as robust super-resolution (RS) based on bilateral total variations by Farsi [14], iterative back projection (IBP) by Irani [15], projection onto convex sets (POCS) by Stark and Oskoui [16], and structure-adaptive normalized convolution (SANC) by Pham [17]. These techniques use a priori information about the degradation and the imaging system. The key to apply these techniques is to give appropriate values for the parameters utilized in an image processing system designed to achieve some criteria implied by the constraints. Several methods to find optimal or quasi-optimal solutions (parameters optimization) for problems in image processing based on evolutionary computation introduced in the last decades. Evolving solutions rather than computing them is considered a favorable programming approach. Among those techniques, genetic algorithm (GA) techniques try to find the solution over the natural selection of the possible solutions (individuals) among the iterations of the algorithm (generations) [18].

However, another alternative of evolutionary algorithms was introduced: QGA, it is a combination of GA and quantum computing. Quantum computation has attracted researchers concern. There were some efforts to use QGA for exploring search spaces. In this work, we applied the QGA using Spot-4 & Spot-5 data sets; the framework starts with sub-pixel shift registration, then image restoration and finally IHS+Wavelet fusion. All models based on QGA that used for parameters estimation to extract some computational abilities of QGA to perform processing in an effective and an efficient manner. Results compared with those obtained by corresponding GA-based registration, restoration, and IHS+Wavelet fusion methods and also with those got by corresponding classic Computational methods. The objectives of this work as follows:

- Extracting more spatial information from 2 subpixel-shifted images of the same scene by super-resolution.
- Introducing relevant information from multiple images from two sensors in a single image by fusion.
- Improving the image registration by accurate estimation of the sub-pixel transformation matrix.
- Improving blind image restoration by image-dependent estimation of blur kernel instead of assuming.
- Improving fusion by automatic adaptive-weights estimation according to the application.
- Comparing the proposed methods with classic ones by visual inspection, measuring metrics and plotting the Line Spread Function (LSF) curves for estimating the spatial resolution enhancement.

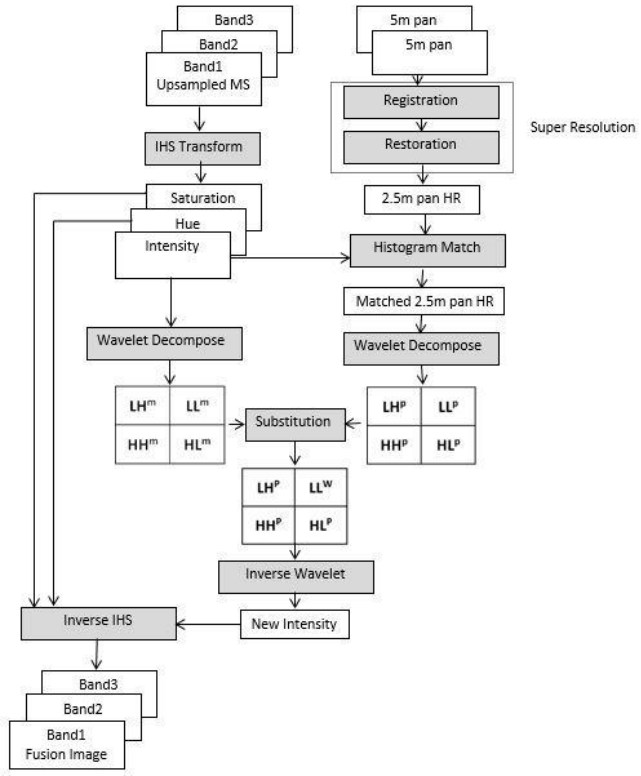


Fig. 1. The IHS+Wavelet fusion steps

This paper is organized as follows. Section 2 explains the quantum genetic algorithm. Section 3 describes the proposed QGA-Based satellite image enhancement framework. Section 4 presents the experiments with different restoration methods. Section 5 discusses the results and the comparison with other classic methods. Finally, conclusion is explained in Section 6.

II. Quantum Genetic Algorithm

Unlike classic computing in which the smallest unit (the bit) can be 0 or 1 to represent the data, the quantum computing uses qubit which can be in the "1" state, in "0" state or in any superposition of them [19]. A state of a qubit described as formula:

$$|\Psi> = \propto |0> + \beta|1> \quad (1)$$

Where |0> and |1> represent the bit classical values 0 and 1 respectively, α and β are complex numbers satisfy the condition:

$$|\propto|^2 + |\beta|^2 = 1 \quad (2)$$

$|\alpha|^2$ is the probability to have the value 0 and $|\beta|^2$ is the probability of having the value 1. However, when the 'measure' or 'to observe' is taken, the qubit will converge into a single

state. If there is a system of m-qubits, the resulting state space has 2^m dimensions. It is an exponential growth of the state space. This exponential parallelism could lead to exponentially faster convergence than the classical systems. The QGA chromosome is string of N qubits can be represented by:

$$q_j^t = \begin{bmatrix} \propto_{11}^t \propto_{12}^t .. \propto_{1k}^t ..... \propto_{m1}^t \propto_{m2}^t .. \propto_{mk}^t \\ \beta_{11}^t \beta_{12}^t .. \beta_{1k}^t ..... \beta_{m1}^t \beta_{m2}^t .. \beta_{mk}^t \end{bmatrix} \quad (3)$$

Where q_j^t represents the j^{th} chromosome of the t^{th} generation, k represents the No. of the qubit in each gene; m represents No. of genes in each chromosome. Comparing with the traditional GA that uses crossover and mutation to achieve population diversity, in QGA the chromosome values updated by a Q-gate as the following:

$$\begin{bmatrix} \propto_i' \\ \beta_i' \end{bmatrix} = U(\Delta\theta_i) \begin{bmatrix} \propto_i \\ \beta_i \end{bmatrix} = \begin{bmatrix} \cos(\Delta\theta_i) & -\sin(\Delta\theta_i) \\ \sin(\Delta\theta_i) & \cos(\Delta\theta_i) \end{bmatrix} \begin{bmatrix} \propto_i \\ \beta_i \end{bmatrix} \quad (4)$$

Where α_i & β_i represent the qubit before update, α_i' & β_i' represent the qubit after update and $\Delta\theta_i$ represent the rotation angle. The lookup table of rotation angle in QGA explained in [20]. Therefore, QGA has better search ability as well as convergence speed and the performance of the algorithm will not be affected with a small population size [21].

III. Quantum Genetic-Based Satellite Image Enhancement

The main goal of this framework is to improve and stabilize the performance of the image enhancement methods by choosing optimal or quasi-optimal parameters through QGA-based techniques. The framework consists of registration, then restoration and at last image fusion. The overall structure of the QGA in general:

1) Initialize the population Q(t0); the initial (α, β) of each are equal 1/sqrt(2) [19], that to begin presenting all states with equal probability.

2) Produce P(t) by observing states of Q(t) (extract a classic gene from a quantum gene, it is selected randomly based on the α and β values of the qubit).

3) Evaluate the fitness of every solution of P(t) by applying objective fitness function.

4) Use the best-evaluated solution in next generation as the evolutional goal.

5) Update population Q(t) by quantum rotating gate, that obtain Q(t+1).

6) Store the best solution and its fitness value.

7) Repeat above steps till convergence to an optimum value or till no improving can get.

A. Quantum Genetic-based Image Registeration

QGA is proposed to estimate the registration parameters (transformation matrix). The fitness function is to minimize the error (difference image) between the reference and input (estimated transformed) images. It defined as following:

$$\text{Minimize: Fit}(C_i): \sqrt[2]{\text{mean}\left(\text{abs}(Y - X)\right)} \quad (5)$$

Where, C_i is genotype, Y and X are the input image and the reference image respectively.

In this work, the initial population is selected to be 10 genotype, the registration parameters are estimated with regards to optimized matching between images. Therefore, these values can be used in warping the images successfully. For comparison, the corresponding conv. GA procedures are applied on the same image and with the same fitness function. Also a classic computational conv. registration method; forwards additive algorithm (Lucas-Kanade) as in [22] is implemented. Bilinear pixel interpolation is selected to calculate the intensity of a transformed pixel with better accuracy. So by interpolating their intensities, the intensities of neighbor pixels are taken into account. This is to improve overall minimization.

B. Quantum Genetic-based Image Restoration

Blur kernel is used for restoring the degraded image and reconstructing an HR image from multi LR images by applying super-resolution techniques. In many studies, the values of blur kernel have been determined by the trial and error for simplicity as work in [23]. In our approach; QGA is applied to estimate three unique values of a symmetric 5×5 blur kernel simultaneously. To evaluate the fitness solution, the multi LR images and the estimated registration parameters from previous step are needed to get the HR in each iteration. The fitness function is defined as the following equation:

$$\text{Minimize: Fit}(C_i): \frac{\text{norm}(Y - Y_{\text{prev}})}{\text{norm}(Y)} \quad (6)$$

Where, C_i is genotype, Y and Y_{prev} are the reconstructed HR image at an iteration and at the previous one. The initial population is selected to be 15 genotype. The blur kernel is being estimated with regards to optimized image quality. Then it is used in reconstructing the HR image. For comparison, another GA-based restoration method is applied to the same image with the same fitness function. Also, the same technique is implemented, but instead of estimating the blur kernel according to metrics, it is assumed by try and error.

C. Quantum Genetic-based (IHS+ Wavelet) Image Fusion

Weighting parameters for fusion step are estimated during this framework by using an optimization approach based on QGA searching technique. The "IHS + Wavelet" fusion is implemented in this research by fusing the HR Image reconstructed in the previous step to each band of the up-sampled MS image, it is performed by using an automatic standard deviation-based injection model in order to standardize the method. Increasing the weight value means that the high-spatial details of reconstructed HR image is more intensely incorporated into the resulted fused MS image. In this framework, fusion weights of each MS band are automatically estimated separately by QGA according to the its pixel values and the required standard deviation while preserving the visual spectral information of the colored MS fused image. Weights are estimated according to the application, here in spatial enhancement application, high-spatial details are most required while preserving the spectral information for visual comparison as possible. For implementing QGA, the wavelet component of reconstructed HR image from previous step and wavelet of intensity of IHS of up-sampled MS bands of original image are needed. The fitness function is defined as following equation:

$$\text{Maximize: Fit}(C_i): \sqrt[2]{\frac{\sum\sum(F-API)^2}{p\times q}} \quad (7)$$

Where, C_i is genotype, F is the resulted fused image, API is the Average Pixel Intensity (Mean) of resulted fused image, p*q is the image size. In this work, the initial population is selected to be 5 genotype. The weights are being estimated with regards to optimized image quality. Therefore, these values can be used in IHS+Wavelet fusion step. For comparing, another GA-based IHS+Wavelet fusion method is applied to the same image with the same fitness function. The same fusion method is implemented by using try and error experimentation for estimating the weights. For visual comparison, linear histogram match is implemented to adapt standard deviation (SD) and mean of the fused image bands to those of the original MS.

IV. Experiments

We conducted experiments to demonstrate the proposed QGA-based framework. Two types of datasets "Spot-5 and Spot-4" for the same scenes are used. Spot-5 datasets are two pan 5m-spatial resolution images, Spot-5 satellite features a new dual linear detector array configured as being offset in the focal plane in such way as to provide coincident imagery of the same instantaneous field view with offset by 2.5m on both lines and columns; that produces two pan images definitely shifted by (0.5, 0.5) [24]. Spot-4 dataset is 20 m spatial resolution MS colored image for the same scene with different viewing angle as shown in Fig. 2. The main framework steps:

1) Estimate sub-pixel shifts between two pan 5m Spot-5 images by the QGA-registration.

2) Use the estimated sub-pixel shifts (step1) in estimating an HR pan 2.5m image by applying the proposed QGA-based restoration method (super-resolution technique).

3) Co-registering the MS Spot-4 image with the pan 5m Spot-5 image by applying geometric correction (standard Erdas program) to overcome the problem of viewing angle difference.

4) Upsample the MS image (from step3) by using cubic interpolation to produce an MS has the same pixel size as the HR pan 2.5m image (from step2).

5) Transform the upsampled MS image into IHS components (forward IHS transform).

6) Apply histogram matching to the HR pan 2.5m to that of the intensity (I) of the MS.

7) Decompose both the matched HR pan 2.5m image and the intensity component (I) of MS (from step5) into wavelet planes respectively (a one-level decomposition is applied).

8) Replace the approximation image of the wavelet-transformed matched HR pan 2.5m image (LL_p) by that of the intensity decomposition of MS (LL_m) to inject gray value information of the intensity image of MS into the HR pan image. To avoid an over injection of the intensity information, the LL_p is not completely, but partially, replaced by the LL_m; namely a new approximation image (LL_w) is first generated through a weighted combination of LL_p and LL_m, and then replaces the LL_p of the matched HR pan 2.5m decomposition. Weights are estimated by the proposed QGA-fusion weights estimation. The detail components (LH_p, HH_p and HL_p) of the matched HR pan 2.5m wavelet decomposition remain unchanged. The method to generate the new approximation image LL_w expressed as:

$$c = w_1.a + w_2.b \quad (8)$$

Where a and b are the approximation images LL_p and LL_m, respectively, and w_1 and w_2 are the corresponding QGA-based estimated weights coefficients.

9) Perform an inverse wavelet transform to obtain a new intensity has similar gray distribution to that of the intensity image from the IHS transformed MS and contains the same spatial detail of the HR pan 2.5m image.

10)Transform the new intensity (step 9) together with the H and S components back into RGB space (inverse IHS transform) to obtain the spatially-enhanced fused MS colored image.

The whole framework is examined on different restoration methods; such as IBP, RS, POCS and SANC (step 2), the corresponding fusion results are shown in Figs. 5& 6& 7 and 8. The restoration parameters used are such as; step size "α" = 0.05 & regularization factor "λ" = 0.2. Beside visual evaluation, spectral and spatial quality metrics are used in this work to evaluate the performance of proposed work.

V. Results and Discussion

The proposed framework of image enhancement based on QGA optimization is fine-tuned by means of three parameters; Transformation matrix of registration, blur kernel for restoration process and injection IHS+Wavelet fusion weights of the added reconstructed pan HR image to the up-sampled MS bands. To evaluate the accuracy of registration parameters, Root Mean Square Error (RMSE) is calculated as shown in table 1 are compared to the well-known displacement values of 5m pan Spot-5 images (0.5, 0.5) in horizontal and vertical directions. Results show QGA is more accurate than conv. GA-based method and more accurate than the gradient-based (Grad.) registration method implemented as in [22]. Investigating the proposed QGA-based weighted IHS+Wavelet fusion method (estimated weights are presented in table 2), by inspecting the quality of enhanced MS fused images; it is noticed that the proposed method preserves the original spectral properties of the added upsampled MS images to a high degree, although images are subject to spectral distortions during fusion operations. The spectral quality of fused MS images is determined according to the changes in colors as compared to the original MS images those before fusion process. The objective is to obtain the fused image with the optimal combination of spectral characteristics preservation and spatial improvement. In this study, three metrics: peak signal-to-noise ratio (PSNR) [25], ERGAS [26] and The Mean Structure Similarity (MSSIM) index [27] have been used in order to determine the spatial and spectral quality of the MS fused images. Tables 3, 4 and 5 show a significantly higher spatial fidelity of QGA approach with regard to conv. GA-based framework and also to traditional conv. method such as work in [23] in which parameters such as blur kernel and fusion weights are assumed or estimated by visually try and

error. The spectral metrics also show better spectral quality in case of QGA approach with regard to GA approach and conv. methods. Evaluating the whole image quality; PSNR values of spatial and spectral metrics are summed, and then their average is calculated, the same for ERGAS and MSSIM spatial and spectral metrics. In case of QGA-based method; results show better averaged values more than GA-based and conv. methods. A visual analysis indicates an increase in spatial quality with respect to the original image (Figs. 5& 6& 7 and 8) while maintaining the spectral quality. To evaluate spatial resolution enhancement, line spread function (LSF) is calculated by edge-knife method. Fig. 3 shows a comparison of the measured LSFs between analogues bands in case of QGA, GA and the conv. Approach (band 1 as an example). We can notice that there is more enhancement in case of QGA. By measuring full width half maximum (FWHM) from LSF curves and comparing with those of classic cubic interpolation method ("cubic" curve); it is obvious that enhancement by factor more than two. Moreover, a comparison between the proposed QGA-based restoration, GA-based method and the conv. method, in sense of convergence is shown in Fig. 4. It shows that the QGA-based method exhibits faster convergence compared to GA-based and classic conv. restoration methods

TABLE I. REGISTRATION PARAMETERS AND THEIR RMSE

	Parameters		**RMSE**	
	dx	***dy***	***dx***	***dy***
Grad.	0.43	0.39	9.1	11.02
GA	0.463	0.44	7.9	8.2
QGA	0.51	0.48	4.6	6.04

TABLE II. IHS+WAVELET FUSION WEIGHTS FOR THE THREE BANDS

Pan weight percentage	**Band1**	**Band2**	**Band3**
GA	0.787	0.80	0.83
QGA	0.807	0.828	0.85

TABLE III. PSNR OF FUSED MS IMAGES USING SEVERAL RESTORATION

PSNR (DB.)		**IBP**	**RS**	**POCS**	**SANC**
Classic	**Spectral**	15.36	15.36	15.64	15.31
	Spatial	4.48	4.38	4.51	4.49
	Average	**9.92**	**9.87**	**10.07**	**9.90**
GA.	**Spectral**	15.54	15.43	15.78	15.54
	Spatial	4.49	4.47	4.53	4.51
	Average	**10.02**	**9.95**	**10.16**	**10.03**
QGA.	**Spectral**	15.70	15.66	15.85	15.62
	Spatial	4.51	4.49	4.57	4.56
	Average	**10.10**	**10.07**	**10.21**	**10.09**

TABLE IV. ERGAS OF FUSED MS IMAGES USING SEVERAL ESTORATION

ERGAS		**IBP**	**RS**	**POCS**	**SANC**
Classic	**Spectral**	23.15	23.15	22.47	23.25
	Spatial	97.98	98.78	97.51	97.69
	Average	**60.56**	**60.96**	**59.99**	**60.47**
GA.	**Spectral**	22.70	22.96	22.16	22.71
	Spatial	97.78	98.05	97.20	97.56
	Average	**60.24**	**60.50**	**59.68**	**60.13**
QGA.	**Spectral**	22.34	22.43	22.00	22.50
	Spatial	97.43	97.60	97.08	97.50
	Average	**59.88**	**60.01**	**59.54**	**60.00**

TABLE V. MSSIM OF FUSED MS IMAGES USING SEVERAL RESTORATION

MSSIM		**IBP**	**RS**	**POCS**	**SANC**
Classic	**Spectral**	0.84	0.85	0.80	0.84
	Spatial	0.90	0.92	0.91	0.93
	Average	**0.87**	**0.88**	**0.86**	**0.89**
GA.	**Spectral**	0.87	0.89	0.83	0.86
	Spatial	0.95	0.98	0.93	0.94
	Average	**0.91**	**0.93**	**0.88**	**0.90**
QGA.	**Spectral**	0.90	0.91	0.86	0.88
	Spatial	0.96	0.99	0.96	0.97
	Average	**0.93**	**0.95**	**0.91**	**0.92**

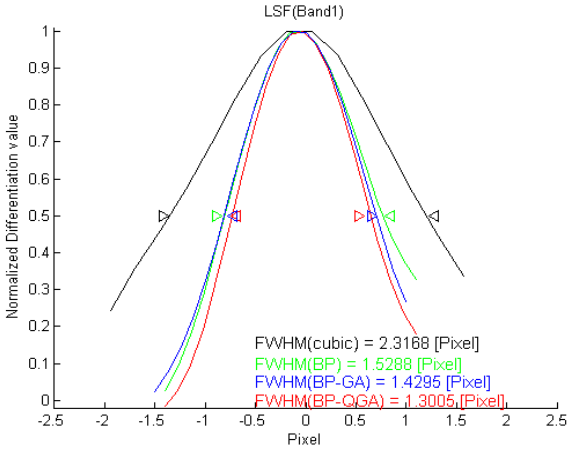
VI. CONCLUSION

The obtained results show that the proposed QGA-based satellite image enhancement framework is much more powerful and efficient compared to the classic GA-Based one. There are two main reasons for this; the first reason is that the quantum encoding of solutions reduces the needed number of chromosomes that achieves reasonable search variance. So, all possible solutions can be represented by only one chromosome at the same time. Therefore, the size of the population does not to be great. It is possible theoretically to use only one chromosome, but in practice, this usually leads to trapping into local optima. Thus, we need little more chromosomes to increase the search space. The second reason is that the advantage of QGA-operations such as rotation gates, that provide in someway a guide for the population individuals, thus the number of necessary iterations to have an acceptable solution is significantly smaller (can be about 60 iterations), and therefore that increase convergence rate. On the other side, benefits of QGA and GA methods in comparison to traditional computational methods are accuracy, the stability of estimation "convergence," automated solution, and the low computational cost. According to the obtained results, the proposed QGA-based method assures accuracy, convergence and better visual image enhancement, it offers very effective solutions for the studied problem. In proposed work, our focus has been to introduce a low-complexity estimation algorithm for using in enhancement MS satellite image. We have shown that the proposed QGA-based registration algorithm rivals many of the more complex state-of-the-art gradient-based motion estimation algorithms. Also we demonstrated that QGA optimization technique can be applied to estimate blur kernel dependent of the image itself instead of assuming or try and error technique in various restoration methods (IBP, RS, POCS, and SANC). Also for implementing the weighted IHS+Wavelet fusion, QGA can be used successfully in the automatic estimation of adaptive injection weights. Simulations and results show that this framework also works in practice.

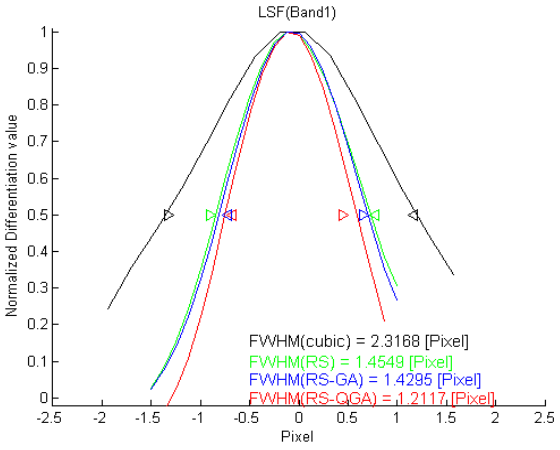
a)

b)

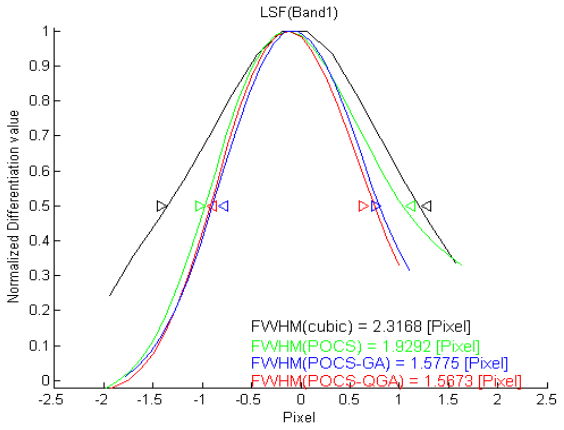
Fig. 2. a) SPOT-5 5m pan image. b)SPOT-4 20m MS image



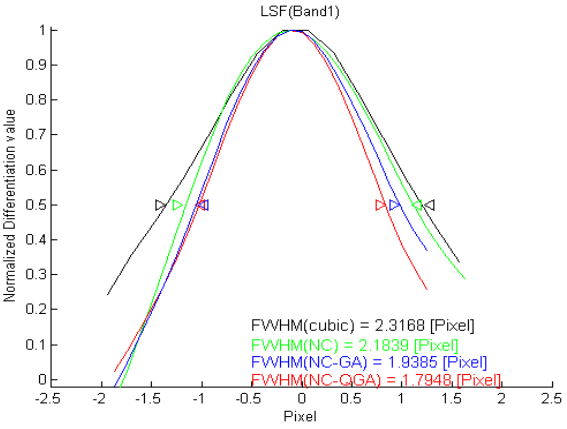
a)



b)



c)



d)

Fig. 3. LSF curves of fused image"band1" in case of these restoration methods a) IBP b) RS c) POCS d) SANC

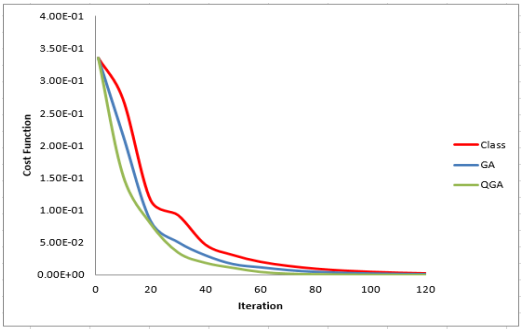


Fig. 4. Convergence rates of Classic, GA, and QGA-Based POCS restoration method

REFERENCES

[1] Andreja, S., Kristof, O. High-resolution Image Fusion: Methods to Preserve Spectral and Spatial Resolution. Photogrammetric Engineering & Remote Sensing 2006, vol. 72, no. 5, p. 565–572.

[2] Hana, S. S., Li, H. T., Gu, H. Y. The Study on Image Fusion for High Spatial Resolution Remote Sensing Images. International Society for Photogrammetry and Remote Sensing Beijing 2008, vol. XXXVII, Part B7, p. 1159-1164.

[3] Aiazzi, B., Alparone, L., Baronti, S., Pippi, I., Selva, M. Generalized Laplacian pyramid-based fusion of MS + P image data with spectral distortion minimization. ISPRS Internat. Archives Photogram, Remote Sensing 2002, vol. 34, no. 3B-W3, p. 3–6.

[4] Qiu, Z. C. The study on the remote sensing data fusion. Acta Geodaetica et Cartographica Sinica 1990, vol. 19, no. 4, p. 290–296.

[5] Dheepa, G., SukumaranSatellite, S. Image Fusion Technique using Integration of IHS Transform and Contrast based Wavelet Packets. International Journal of Computer Applications 2014, vol. 107, no. 9, p. 37-43.

[6] Tu, T. M., Su, S. C., Shyu, H. C., Huang, P. S. A new look at IHS-like image fusion methods. Inf. Fusion 2001, vol. 2, no. 3, p.177–186.

[7] Kusum, Rani, Reecha, Sharma. Study of Image Fusion using discrete wavelet and Multiwavelet Transform. International Journal of Innovative Research in Computer and Communication Engineering 2013, vol. 1, no. 4, p. 795-799.

[8] Nikolakopoulos, G. K. Comparison of nine fusion techniques for very high resolution data. Photogrammetric Engineering &Remote Sensing 2008, vol. 74, no. 5, p. 647–659.

[9] Yocky, D. A. Multiresolution wavelet decomposition image merger of Landsat Thematic Mapper and SPOT panchromatic data. Photogrammetric Engineering & Remote Sensing 1996, vol. 62, no. 3, p. 295–303.

[10] Hong, G., Zhang, Y., Mercer, B. A Wavelet and IHS Integration Method to Fuse High Resolution SAR with Moderate Resolution Multispectral Images. Photogrammetric Engineering & Remote Sensing 2009, vol. 75, no. 10, p. 1213–1223.

[11] Harpreet, Kaur, Rachna, Rajput. A Combined Approach using DWT & PCA on Image Fusion. International Journal of Advanced Research in Computer and Communication Engineering 2015, vol. 4, no. 9, p. 294-296.

[12] Gagandeep, kaur, Anand, Kumar, Mittal. A New Hybrid Wavelet Based Approach for Image Fusion. International Journal of Innovative Research in Science, Engineering and Technology 2015, vol. 4, no. 1, p. 19034- 19043.

[13] Metwalli, M. R., Nasr, A. H., Faragallah, O. S., El-Rabaie, S., Abd El-Samie, F. E. Combining Super-resolution and Fusion Methods for Sharpening Misrsat-1 Data. Geoscience and Remote Sensing, IEEE Transactions 2013, vol. 51, No. 4, p. 2292 – 2301.

[14] Farsiu, S., Robinson, D., Elad, M., Milanfar, P. Fast and Robust multi frame super resolution. IEEE transactions on Image Processing 2004, vol. 13, p. 1327 – 1344.

[15] Irani, M., Peleg, S. Improving resolution by image registration. In CVGIP: Graphical Models and Image Processing 1991, vol. 53, p. 231-239.

[16] Stark, H., Oskoui, P. High resolution image recovery from image plane arrays using convex projections. J. Opt. Soc. Am. A 1989, vol. 6, p. 1715-1726.

[17] Pham, T. Q., Van Vliet, L. J., Schutte, K. Robust fusion of irregularly sampled data using adaptive normalized convolution. EURASIP J. Appl. Signal Process. 2006, vol. 2006, p. 1-12.

[18] Rahman, M. F., Karim, S. M. M. Performance analysis of estimation of distribution algorithm and genetic algorithm in zone routing protocol. International Journal of Computer Science and Information Security (IJCSIS) 2010, 8, p. 203-207.

[19] Han, K. Genetic Quantum Algorithm and Its Application to Combinatorial Optimization Problem. In Proceedings of IEEE Congress on Evolutionary Computation 2000, pp. 1354-1360.

[20] Jiang, Sh., Zhou, Q., Zhang, Y. Analysis on parameters in an improved quantum genetic algorithm. International Journal of Digital Content Technology and its Applications (JDCTA) 2012, vol. 6, no. 18, p. 176-184.

[21] Zhang, G., Li, N., Jin, W., Hu, L. A novel quantum genetic algorithm and its application. Acta Electronica Sinica 2004, vol. 32, no. 3, p. 476-479.

[22] Baker, S., Matthews, I. Lucas-Kanade 20 years on: A unifying framework. International Journal of Computer Vision 2004, vol. 56, no. 3, p. 221–255.

[23] He, H., Kondi, L. P. A regularization framework for joint blur estimation and super-resolution of video sequences. In proceeding of: Image Processing, ICIP, IEEE Int. Conf. 2005, 3, p. 329-332.

[24] Han, Y., Lee, S. Parameter Estimation-based Single Image Super Resolution. 1st IEEE Global Conference on Consumer Electronics GCCE, February 2012, p. 565-569.

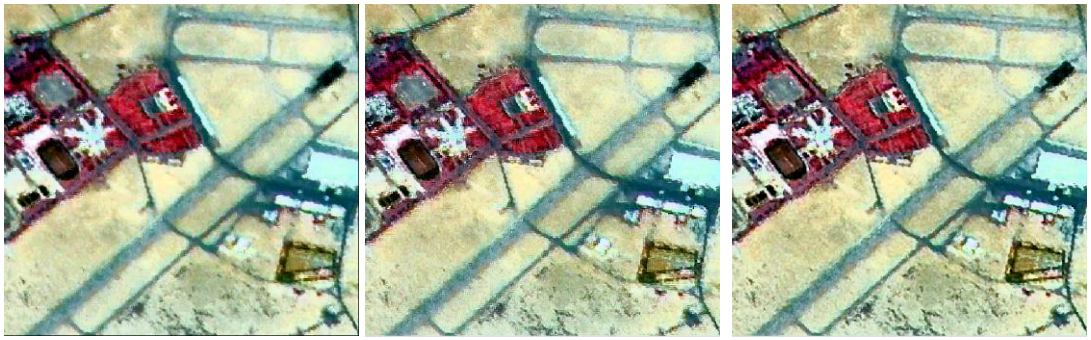
[25] Liyakathunisa, Ravi Kumar, C. N. A novel super resolution reconstruction of low resolution images progressively using DCT and zonal filter based denoising. International Journal of Computer Science & Information Technology (IJCSIT) 2011, vol. 3, no. 1.

[26] Gonzalo-Martín, C., Lillo-Saavedra, M. Balancing the Spatial and Spectral Quality of Satellite Fused Images through a Search Algorithm. Search Algorithms and Applications, Prof. Nashat Mansour (Ed.), InTech. 2011, vol. 156, no. 307, p. 953-978.

[27] Wang, Z., Bovik, C. A, Simoncelli, P. E. Image Quality Assessment: From Error Visibility to Structural Similarity. IEEE transactions on image processing 2004, vol. 13, no. 4, p. 600-612.

(a) (b) (c)

Fig. 5. Fused MS images in case of IBP restoration. (a) Classic (b) GA. (c) QGA



(a) (b) (c)

Fig. 6. Fused MS images in case of RS restoration. (a) Classic (b) GA. (c) QGA

(a) (b) (c)

Fig. 7. Fused MS images in case of POCS restoration. (a) Classic (b) GA. (c) QGA

(a) (b) (c)

Fig. 8. Fused MS images in case of SANC restoration. (a) Classic (b) GA. (c) QGA

Improved Appliance Coordination Scheme with Waiting Time in Smart Grids

Firas A. Al Balas, Wail Mardini,Yaser Khamayseh, Dua'a Ah.K.Bani-Salameh
Department of Computer Science,Jordan University of Science and Technology, Jordan
Irbid, Jordan 22110

***Abstract*—Smart grids aim to merge the advances in communications and information technologies with traditional power grids. In smart grids, users can generate energy and sell it to the local utility supplier. The users can reduce energy consumption by shifting appliances' start time to off-peak hours. Many researchers have proposed techniques to reduce the previous issue for home appliances, such as the Appliances Coordination (ACORD) scheme and Appliances Coordination with Feed In (ACORD-FI) scheme.**

The goal of this work is to introduce an efficient scheme to reduce the total cost of energy bills by utilizing the ACORD-FI scheme to obtain an effective solution. In this work three scheduling schemes are proposed: the Appliances Coordination by Giving Waiting Time (ACORD-WT), the Appliances Coordination by Giving Priority (ACORD-P), and using photovoltaic (PV) with priority and waiting time scheduling algorithms.

A simulator written in C++ is used to test the performance of the proposed schemes using. The performance metric used is the total savings in the cost of the energy bill in dollars. The first comparison for the proposed schemes with the ACORD-FI, and the results show that the efficiency of the proposed ACORD-WT is better than the ACORD-FI, regardless of the number of appliances. Moreover, the proposed ACORD-P, is also better than the standard ACORD-FI.

Keyword—smart grids; energy bill; off-peak

I. INTRODUCTION

Energy sources are classified into renewable and nonrenewable sources. Renewable sources are those that can be accessible to humans in a timely scaled manner that comes from natural resources on a regular or irregular basis. For example, sunlight is available on a daily basis in the summertime in many areas around the world. However, the wind would be available in some areas on a regular basis and would be slightly useful in others around the world. A renewable source contains many sources that can be listed as follows: tidal power, wave power, solar power, wind power, hydroelectricity, radiant energy, geothermal power, biomass, compressed natural gas, and nuclear power.

On the other side, nonrenewable sources are those that do not renew in enough amounts. For example, coal needs thousands of years to build naturally and cannot be available at a relevant rate of consumption. Examples of such sources are petroleum, coal, natural gas, and nuclear power.

Electricity is an energy form called electricity energy, and it is not similar to the other sources of energy, such as coal, petroleum, and solar energy. It is defined as the set of physical phenomena associated with the flow of an electric charge. The traditional sources of generating electricity were nonrenewable sources but can be generated from renewable sources [1].

Electrical devices that are common in homes are ovens, washers, dishwashers, televisions, microwaves, and others. Each one can consume an already measured amount of energy, and with the demand for electricity increasing, energy consumption has increased. Therefore, the cost of energy consumption also has increased.

The increasing demand for electricity in the future must proceed through updating the electric grid and creating smart ones. The term "grid" refers to the electrical distribution system, which transmits electricity from power plants located near fuel sources to the consumption locations, where the previous electric grid or the traditional grid has worked well for many years. Fig. 1 shows the traditional power infrastructure.

The smart grid was founded to solve the increasing demands on electricity. The smart grid establishes and distributes electricity more efficiently, economically, and securely, and it combines different technologies, products, services, from generation, transmission, and distribution to and from consumer appliances by using advanced sensing, communications, and control technologies [2]. Fig. 2 shows the modern power infrastructure (smart grid).

Smart-grid technologies can control and monitor the power consumption in both homes and buildings, where each device has different operations that can be used and scheduled, resumed, suspended, and stopped by a smart meter. The smart meter enables scheduling of these operations, which ensures savings by 1) reducing the energy during peak demand time, 2) reducing cost, 3) increasing reliability, 4) and reducing power-interruption periods.

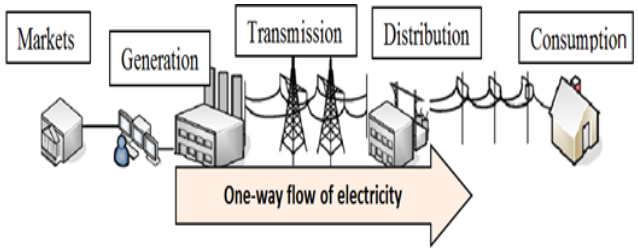


Fig. 1. Traditional power infrastructure

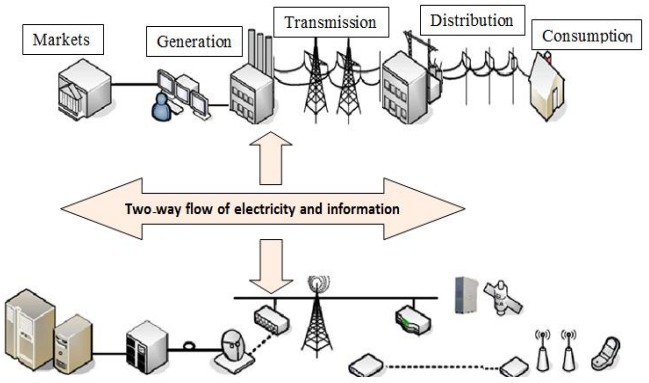


Fig. 2. Modern power infrastructure (smart grid)

The smart grid consists of four main parts [3]:

1) Advanced Transmission Operations (ATS) aim to achieve intelligent transmission and decrease the risk of failure.

2) Advanced Metering Infrastructure (AMI) is considered a key component of the smart grid because it deploys communication networks to connect each customer with utility companies and also interacts with smart meters for scheduling the energy management.

3) Advanced Distribution Operations (ADO) are full automation of all control devices, and their primary focus is on the self-healing capabilities of the smart grid.

4) Advanced Asset Management (AAM) deals with the management of the industrial equipment at the user.

Smart grids have many objectives; they allow for a two-way flow of information between consumers and utility, improve energy storage, are self-healing, environmentally friendly and able to implement consumer energy management policies [18]. However, smart grids face many challenges such as regulation, standardization, and security.

The drivers that are involved in developing the smart grids can is into three groups: the government, customer behavior, and industry and technology changes. Fig. 3 shows the smart-grid layout.

A. Wireless Sensor Networks (WSNs)

Wireless sensor networks (WSNs) are an essential part of smart grids and provide low-cost, low-power solutions. Some applications of WSNs in smart grids are the home appliance and environment monitoring that include consumption and fault detection. The main challenges faced by WSNs' applications in smart grids are hard environmental conditions (e.g., high humidity levels, vibrations, dirt, and dust), security issues, resource constraints, packet errors, and system reliability. There are three resources (energy, memory, and processing) responsible for the design and implementation of WSNs [6].

B. Electrical devices

Electrical devices can consume major amounts of energy (for example, washers, dishwashers, dryers, coffee makers, plug-in hybrid electric vehicle (PHEV), and air conditioners). Each device consumes a specific amount of electrical energy that can be measured in kWh and can be scheduled to reduce energy cost and consumption.

Each device has many different cycles with different energy-consumption level for each. The duration time in minutes for each cycle are 10, 30, 60, 60, 60, and 90 for a coffee maker, washer, dryer, PHEV, air conditioner, and dishwasher, respectively, where the energy consumption for each device is 0.4, 0.89, 2.46, 9.9, 1.5, and 1.19, respectively [4].

PHEV cars are hybrid electric vehicles equipped with researchable batteries in addition to the traditional liquid fuel tank. The battery fully charged by linking the plug with an external electric power source. Most PHEVs are recharged during off-peak hours. The cost of electricity to operate the hybrid car has been estimated at less than one quarter the cost of gasoline. Also, they reduce air pollution. Typical recharging of a PHEV battery takes several hours. The quick-mode charge to around 80% capacity may take as little as 30 minutes. Most PHEVs need 0.2–0.3 kWh charging power for 1 mile of driving. The energy consumption for this device is 9.9 kWh [5].

A washing machine is a typical appliance that exists in almost every home. Electrical energy is used for driving the drum motor and heating up the water. The energy consumption for this device is typically around 0.89 kWh.

A dryer uses huge amount of energy to remove the humidity from the clothes. It was invented in England and France in the early 1800s. It is used commonly nowadays in North America. The energy consumption for this device is around 2.46 kWh.

A dishwasher is a mechanical device that is used in cleaning dishes and may be found in many restaurants and homes. The amount of energy used depends on whether it is connected to hot or cold water. The power consumption for this device is 1.19 kWh.

An air conditioner (AC) is used in many places to control the temperature of the limited area around it. Its cooling capacity is measured regarding BTU and considered as the amount of power used to lower the temperature of the air. There are different types of AC systems: window, split unit, and central AC. The capacity of the window and split-unit type is around 6000–24000 BTU, while the capacity of the central type is around 9000–60000 BTU. The energy consumption for this device is typically 1.5 kWh [6]. A coffee maker is used commonly to make coffee in western countries, and there are many different types. The power consumption for this device is 0.4 kWh.

TABLE I. ENERGY CONSUMPTION AND CYCLE DURATION OF APPLIANCES

Appliance	Energy consumption (kWh)	Duration (min)
Washer	0.89	30
Dishwasher	1.19	90
Dryer	2.46	60
Coffee Maker	0.4	10
PHEV	9.9	60
AC	1.5	60

C. Real Time Pricing

There are different pricing schemes proposed in typical power grids; some are specific for smart grids, for example, Real Time Pricing (RTP), Time of Use (TOU), Critical Peak Pricing (CPP), Day Ahead Pricing (DAP), and Inclining Block Rate (IBR).

The previous schemes are used widely. For example, the TOU pricing scheme is used in the Appliances Coordination with Feed In (ACORD-FI), Optimization-Based Residential Energy Management (OREM), and in-Home Energy Management (iHEM). Also, RTP is used in the Residential Energy Load Consumption (RLC) scheme, where TOU and CPP pricing are used in a decision support tool scheme [13].

In RTP, the price is changed hourly and is fixed during the period. RTP reflects the wholesale prices, weather conditions and generator failures.

D. Job Scheduling

Job scheduling is the process of deciding how to assign resources to different tasks to optimize one or more objectives, such as minimum waiting time and maximum response time. The job is scheduled by using priorities, delay, and custom scheduling conditions [7].

Scheduling algorithms are required for most modern systems to perform the multitasking (e.g., operate more than one process at the same time) and multiplexing (send multiple flows at one time). The scheduling can be classified as preemptive or non-preemptive scheduling. Following is a detailed description of those scheduling types.

In this paper, we used the preemptive scheduling used in real-time systems. It implements the highest priority task of all those tasks that are currently ready to implement.

Preemptive scheduling includes Priority and Round Robin (RR) algorithms. In Priority Preemptive scheduling, each process at the ready list is in descending order by its priority, so the process in the beginning of the list has the highest priority and is picked first by the scheduling algorithm. However, in RR scheduling, each process has a small unit of time, and the jobs move to the next process and continue until all processes are completed.

Non-preemptive scheduling is defined as when a process enters the state of operation; the state of that process is not removed from the scheduler until it is completed. Non-preemptive scheduling includes the algorithms of First Come First Served (FCFS) and Shortest Job First (SJF). FCFS, which is also known as First in First Out (FIFO), is the simplest scheduling algorithm. The jobs are completed in the same order they arrive, but this algorithm has a disadvantage in that it has long waiting times.

Fair scheduling is a method of assigning resources to requests in which the requests are distributed equally such that all requests get an average share of resources over time. The available Center Processing Unit (CPU) is divided initially among the groups, then among the users within each group. The requests into each pool (group) are scheduled using either fair scheduling or FIFO scheduling. The fair scheduler can limit the number of running requests per user and per pool. The key goal of the fair scheduler is to run small requests quickly in case the large requests are running [23].

The objectives of the fair-share scheduler are to ensure fairness, fast response time, and load spreading without making any request wait for too long.

The FCFS is implemented if all requests have the same weight, which means all processes in the requests queue are given time in the form of a time slice that increases when the weight increases. The average wait time for Weighted Round Robin (WRR) is better than for RR [24].

E. Problem Definition

Electricity is defined as a secondary source of energy that uses other primary sources, like coal and wind, that are increased during that use. But the use of electricity increases every year by consumers due to the ease of which it moves from the producer's position (power plant) among long distances to the consumption position.

Therefore, the electricity grid traditionally was proposed and built to give and distribute the energy service. Now, though, consumers consume large amounts of energy to operate several appliances, such as microwaves, washers, lights, Coffee Maker, and more, at the same time, and the cost of electricity is dynamic at peak and off-peak pricing.

For this reason, smart grids are being developed to manage energy consumption by reducing energy consumption and its cost and, as a result, the energy bill.

Many approaches have been implemented to manage energy consumption. Most of these approaches helped to reduce the total energy cost, but according to our knowledge, much saving can be accomplished with an improved approach.

F. Paper Objectives

In this paper, a set of goals were achieved, as are listed below:

1) To allow for monitoring and controlling in to reduce the amount of energy consumption of home appliances.

2) To reduce the cost of energy consumption and power interruption periods, and after that, reduce the energy bill for the customer.

3) To reduce peak demand, which will also help lower electricity rates.

G. Paper Organization

This paper is composed of five sections. After the introduction section, we list the related works for this paper in Section II. Section III presents the methodology and the system model of this paper. In Section IV, the experiments and results are discussed. Section V presents the conclusion of the suggested work and future work.

II. Related Works

In the literature, the increasing demand for electricity in the grid and the need to manage the energy consumption has been studied in several works from 2009 until now.

Many approaches that have been proposed and implemented to manage energy consumption to decrease the rate of total energy consumption and reduce the total energy cost in energy bills. We present in the following the most important papers discussing this issue, and then, we present a summary and discussion of the most important papers related directly to our work.

The work in [24] proposed an optimal and automatic residential energy consumption scheduling framework with the goal of decreasing the cost of the energy bill and minimizing the waiting time for the operation of each device. The authors studied the consumption during the period between September 1, 2009, and December 31, 2009 (122 days; four months). The number of devices used each day varied from 10 to 25. The devices were divided into two parts: fixed-consumption devices, such as electric stove, lighting, heating, refrigerator–freezer, and devices with a varying consumption energy rate, such as dishwasher, clothes washer, clothes dryer, and PHEV. The results showed that their technique reduced the user's cost along with the peak-to-average ratio in the load demand.

In [14], the authors employed TOU-aware energy management in a smart home with a wireless sensor home area network that affected the peak load to reduce the energy bills. The rate of electricity was different in each on-peak, moderate-peak, and off-peak hours. The smart grids were divided into three parts, smart meters, home gateway, and user devices, and the devices could collaborate to reduce the consumer demand to decrease the energy bills and the load on the grid.

This application uses wireless communication between user devices and sensor network; the devices also communicate with an energy management unit (EMU), which manages the user requests by scheduling the duration of devices to off-peak hours or provides the use of local energy, if available. The residential energy management application is an important component of the smart grid that combines the Information and Communication Technologies (ICT) to the traditional power grid, and the communication among devices and energy uses Zigbee with IEEE 802.15.4 standard with short-range wireless links.

The simulator used was implemented in C++, and the user request was modeled as a Poisson process. The interarrival rate for off-peak hours had a negative exponential distribution with a mean of 12 hours, while in the peak hours, the mean was 1 hour. The devices used were a washer, dryer, dishwasher, and coffee maker, which deployed in the peak winter period from 7 to 11 am and from 5 to 9 pm and the mid-peak hours from 11 am to 5 pm.

The simulation ran between 10 and 210 days, and the maximum delay chosen was 24 hours. The performance metrics were taken (e.g., the peak load ratio, total payment of energy consumption, and delay by the users). The results showed a reduction in the user's sharing on the peak load by 30%.

The authors in [18] presented a home energy management application that is used in a WSN. The sensor nodes are communicated by IEEE 802.15.4 and specify the performance metrics. The metrics are 1) delivery ratio (the ratio of the number of metrics received to the number sent) and 2) delay and packet delay variance for two forms (differential interval times and different network sizes).

The periods of time are divided into two interval times, depending on the load on the network. The first period is called the on-peak period, in which there is high loading on the network, and the other period is called the off-peak period, in which there is low loading on the network. When the load becomes a high load on the network, it is called an on-peak period, but when the load is low, the network it is off-peak.

In [13] they presented an application called in-Home Energy Management (iHEM) that uses WSNs, which employ smart devices. The message flow for iHEM is given in Fig. 3. When the consumer turns on the devices, it generates the start request (START-REQ) packet and sends it to the EMU, which communicates with the smart meter, which in turn gives the updated price information.

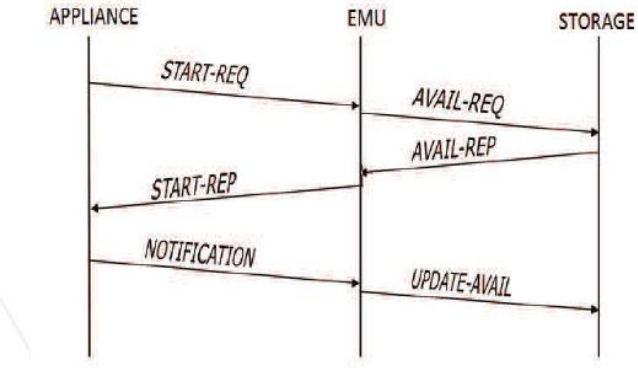


Fig. 3. Message flow for iHEM [13]

The EMU's role is to send the available request (AVAIL-REQ) packet to the energy units, which then retrieve the available amount of energy in the unit and return the available reply (AVAIL-REP) packet back to the requested EMU. The EMU determines the starting time of the devices and computes the waiting time, which is defined as the difference between the suggested and requested start time. Finally, the user decides whether to operate the device immediately or wait.

The simulator used was Qualnet (network simulator). The devices can be defined by two types: 1) a full-function device, such as a smart meter, and 2) a reduced-function device. The WSN is organized in this simulation using a cluster-tree topology, and the interval time is specified between periods (100s–300s). The numbers of the selected nodes are between 20 and 30 nodes. The results from this simulation showed that the delivery ratio increased for small size networks, which reached 85% , while the end-to-end delay decreased. Therefore, the performance of the WSN was better for smaller networks.

The authors in [16] presented an evaluation of the performance of iHEM, in which they showed the effectiveness of iHEM and the Optimization-Based Residential Energy

Management (OREM) scheme. OREM reduces users' energy consumption, while iHEM is able to achieve more objectives, such as reducing the share of the devices in energy bills, reducing the contribution to the peak load, and reducing carbon emission.

Performance was measured by a number of scenarios: local energy generation, iHEM with priority, and iHEM with RTP. There are three components to iHEM: EMU that communicates with devices, a smart meter, and a storage unit. It also employs WSNs for communication through a ZigBee protocol. The simulation period was extended for seven months (from 20 to 210 days).

The researchers applied the iHEM simulation through Microsoft Visual studio C++, where the interarrival time between two requests is a negative exponential distribution with a mean of 12 hours while the mean is 2 hours for both the morning and evening periods. Four devices were used in this simulation (washer, dishwasher, dryer, and Cofee Maker), and the delay accepted in the OREM was 12 hours. The simulation time was between 20 to 210 days for each OREM and iHEM scheme.

The results reduced the expenses of the consumers, compared to the case without energy management, and in all scenarios, the total contribution of the devices to the energy bill decreased when compared without priority, feed-in (local generation), and to the TOU pricing scheme.

In [13], the authors developed the Energy Management and Monitoring system (EMM), which manages the power in buildings with a Building Energy Management System (BEMS). EMM contains an EMM client placed in the building with two forms of interfaces (wired and wireless) to gather the energy metering and sensor data through the Internet by many sensors and then sends it to the EMM server installed in the EMM center and linked with smart meters to compute the energy consumption.

The EMM system is proposed to provide Energy Management Service (EMS) functions, select energy resources with a low price, reduce unnecessary loads, and control the battery of PHEVs that are interworking with a smart grid.

The work in [8] proposed an Optimum Load Management (OLM) technique for RTP that utilizes the communication infrastructure of the future smart grid that will enable the consumer to balance between energy bills and their economic situation. The aim of this scheme is to reduce energy consumption cost, and the results showed high potential by reducing the energy bill by 8–22%.

The authors in [8] presented details of the various Home Energy Management schemes (HEMs), which aim to reduce the peak demand, an average ratio (PAR) and increase savings. That makes the grid smarter and faster in making decisions.

The demand curve and flat pricing rates scenario in the traditional power grid shows that the load demand during peak periods is very high vs. off-peak periods. HEMs enable Demand Side Management systems (DSM) and Demand Response (DR).

There are different techniques for energy management in the smart grid:

1) OREM aims to manage the energy consumption by scheduling home devices, and it specifies the maximum delay for each device as equal to the length of two-time slots.

2) The iHEM system uses smart devices, a central EMU, and WSNs for communication purposes through ZigBee protocol, and there are two types of devices: Full-Function and Reduced-Function devices. Full-function devices are interconnected in a mesh topology, and Reduced-Function devices are interconnected in a star topology. That aims to manage the home energy by shifting the load to off-peak periods.

The application works when the consumer presses the start button of the device, and the device generates a data packet that is sent to the EMU. The EMU communicates with the smart meter, and local generation units provide the price information. The EMU schedules the time of the device from this information.

The results showed that the sharing of the device to the total load was reduced during peak hours, and the peak load was reduced up to 5%.

The authors in [26] studied the development of the smart residential load simulator with a user-friendly graphical interface that aims to achieve easy study of energy management systems in smart grids by simulating the on–off decisions of residential devices. For this study, they used a specific tool based on Matlab Simulink-GUID toolbox available at www.power.uwaterloo.ca. Appliances used in this study were thermostats, air conditioners, furnaces, water heaters, refrigerators, stoves, dishwashers, clothes washers, dryers, lights, and pool pumps, as well as wind, solar, and battery.

The authors in [21] presented a new scheduling method to smooth the demand situation of each house to reduce the energy prices by using a genetic algorithm, which controls the occurrence time of devices and coordinates the groups to set optimization of each group at the same time.

The objectives of the proposed method were to shift the peak demand, control a wide range simultaneously, and reduce the utility bill. The results showed that the proposed method can reduce electric costs by 4.71%.

The authors in [10] studied the problem of the increased level of demand response management in the smart grid called offline scheduling. The objective was to be able to schedule all requests with a minimum total electricity cost. They proposed a polynomial time offline algorithm to achieve the optimal solution, and it was able to optimize the time complexity to O (n T log n), where the time complexity before the optimization is O (n2T).

A simulator was implemented by using Python for a six-hour timeframe and divided into a sequence number of time slots, and it is available online at [23].

The work in [17] proposed the Appliance Coordination (ACORD) scheme for smart grids that allows flexibility in the start time for home devices. The main goal of the ACORD scheme is to shift devices' start time to off-peak hours when the consumer's desired start time falls between peak hours. The scheme uses the in-home WSN to relay the data between the coordinator and the different devices in the home. The architecture for the ACORD scheme is given in Fig. 4. When the user operates the device by the start button, it generates a START-REQ packet that contains the desired on duration cycle of the device (e.g., washing cycle of the washer) and the packet is sent to the EMU by the WSN.

Once the EMU receives the START-REQ packet, it schedules the available start time, if no hard start time is requested, after communicating with the smart meter to check the TOU rate and peak hour information. In a large house environment, the EMU may be physically far away from the appliances and not reachable on one hop by all devices. Thus, multi-hopping is required for message delivery.

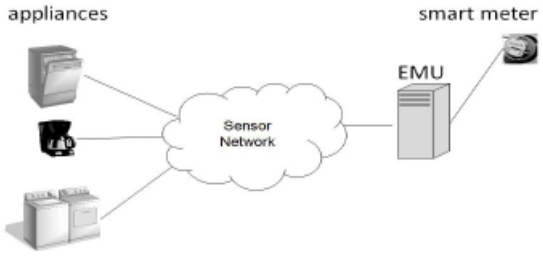


Fig. 4. The architecture for the ACORD scheme [17]

The waiting time or the scheduled start time is sent back to the consumer by the START-REP packet and set to zero if the consumer requests a hard start time or the desired start time is in off-peak hours and no other requests are scheduled on that time. The consumer's final decision is sent back to the EMU in notification packets, using the decision in the reserve time slot for the device

The researchers tested the performance of the scheme using simulation. They used two different load scenarios, high and low consumer activity cases. In the high- and low-activity cases, the interarrival times between the two requests were assumed to be a negative exponential distribution with means of 6 and 48 hours. The devices used included a washer, dryer, dishwasher, and coffee maker. The peak hours chosen were between 7 and 11am and 5 and 9pm in the winter on weekdays, and the maximum acceptable delay was 10 hours. The performance metrics used were the total cost saving in dollars and the number of lost requests in the sensor network. The results showed that the total contribution of the devices to the energy bill was $47 of the consumer requests at the period.

The authors in [15] proposed the ACORD-FI scheme as an improvement of the ACORD scheme proposed in [17]. The main assumption here is that the device can manage the user requests and the energy is generated locally to reduce the sharing of the devices and provide savings on the energy bill. The scheme uses the in-home WSN to relay the data between the coordinator and the different devices in the home.

They tested the performance of the scheme using simulation. They used two different load scenarios, high- and low-activity cases. The interarrival times between two requests were assumed to be negative exponential distribution with means of 6 and 48 hours. The devices used included a washer, dryer, dishwasher, and coffee maker. The peak hours chosen were between 7 and 11 am and 5 and 9 pm in the winter weekdays, and the maximum acceptable delay was 10 hours. The performance metrics used were the total cost savings in dollars and the number of lost requests in the sensor network. The results showed an improvement over ACORD by a rate of $37 as a total contribution of devices to the energy.

In [23], the authors proposed a new energy scheduling algorithm to minimize the expenses of a customer energy bill. The proposed algorithm takes into account the uncertainty in household appliances and the irregular renewable energy generation. It also takes into consideration the variable power generation from renewal resources (e.g., solar, air, etc.) and the capacity-limited energy storage in attached batteries.

The proposed scheme was claimed to achieve up to a 45% cost reduction compared to traditional scheduling algorithms. Also, this proposed scheme was claimed to be able to generate a scheduling solution in 10 seconds, which is fast enough for home appliances applications.

In [24], the authors proposed a dynamic scheduling scheme that depends on the idea of optimal portfolio selection to generate a user's energy consumption history, called a weighted graph. By using this weighted graph, the scheme can detect the user's need of energy close to the optimal need. The performance of the proposed scheme is evaluated with different performance metrics-peak-demand, demand variation, energy cost, and the utility of the customers. Simulation results showed that the proposed dynamic scheduling scheme, D2S, yielded improved performance compared to the existing ones of no scheduling and static scheduling.

In [25], the authors proposed a power scheduling scheme formulated as an optimization problem that includes integer and continuous variables. An optimal scheduling strategy is obtained by solving the optimization problem. The proposed work assumed that consumers have two types of appliances. The first type of appliances has a flexible starting time and works continuously with fixed power.

The second type of appliances works with flexible power in a predefined working time. At the same time, the consumers can adjust the starting time of the first type of appliances or reduce the power consumption of the second type of appliances to reduce the payments. However, this also will incur discomfort to the consumers. As claimed, the simulation results achieved the desired trade-off between the payments and the discomfort by solving the optimization problem.

Table 2 presents a summary of works related directly to ours and shows the goal of the reference, techniques used, main results, simulator used, period used, and the number of appliances used.

TABLE II. A COMPARISON BETWEEN THE DEPLOYMENT TECHNIQUES THAT HAVE BEEN USED IN RELATED WORKS

Reference number and name	Goal of the reference	Technique used	Results
[8] Home Energy Management Systems in Future Smart Grids	Enable the consumer to balance between energy bills and economic situation.	Optimum Load Management (OLM) technique.	Reduced the energy bill by 8–22%.
[8] Home Energy Management Systems in Future Smart Grids	Reduce the peak demand, average ratio (PAR), and increase savings.	Home Energy Management schemes (HEMs).	Reduced the peak load up to 5%.
[9] Scheduling for Electricity Cost in Smart Grid	Schedule all requests with minimum total electricity cost.	Polynomial time offline algorithm.	Achieved the optimal solution and was able to optimize the time complexity to O (n T log n).
[10] Proactive energy management system architecture interworking with smart grid	Manage the power in buildings with Building Energy Management Systems (BEMS).	Energy Management and Monitoring (EMM) system.	Reduced unnecessary loads and controlled the battery of PHEVs that were interworking with the smart grid.
[11] TOU-Aware Energy Management and Wireless Sensor Networks for Reducing Peak Load in Smart Grids	Achieve efficient use of green energy, increase automation in distribution, and enable residential energy management.	Time of Use (TOU)-aware energy management.	Reduction in the consumer's the contribution on the peak load by 30%.
[12] Using Wireless Sensor Networks for Energy-Aware Homes in Smart Grids	Reduce the total cost of the energy bill and provide more savings on the energy bill.	Appliance Coordination with Feed In (ACORD-FI) scheme.	Improved total contribution of devices to the energy bill to $37.
[13] Wireless Sensor Networks for Cost-Efficient Residential Energy Management in the Smart Grid	Evaluate the performance of In-Home Energy Management application (iHEM).	iHEM and Optimization-Based Residential Energy Management (OREM) scheme.	Reduced the consumer's expenses, compared with the case without energy management.
[14] Wireless Sensor Networks for Smart Grid Applications	Achieve the performance of the wireless sensor networks (WSN) under varying interarrival times and varying network sizes.	A home energy management application.	Increased delivery ratio for small-size networks that reached 85% while the end-to-end delay decreased.
[16] Distributed Demand Scheduling Method to Reduce Energy Cost in Smart Grid	Shift peak demand, control a wide range simultaneously and reduce utility bill.	Distributed Demand Scheduling method.	Reduced electricity costs by 4.71%.
[18] Optimal Residential Load Control with Price Prediction in Real-Time Electricity Pricing Environments	Differentiate between minimizing the payment and minimizing the waiting time for the operation of each device.	Optimal and automatic residential energy consumption scheduling framework.	Reduction in user's payments and in peak-to-average ratio in load demand.
[19] Development of a Smart Residential Load Simulator for Energy Management in Smart Grids	Achieve easy study of energy management systems in smart grids.	Development of the smart residential load simulator.	Reduced the peak load in dynamic pricing (delay the demand to the periods of the low electricity price).
[25] Wireless Sensor Networks for Domestic Energy Management in Smart Grids	Shift devices start time to off-peak hours.	Appliance Coordination (ACORD) scheme.	Improved total contribution of devices to the energy bill to $47.

III. METHODOLOGY AND SYSTEM MODEL

In this section, we propose two novel scheduling schemes to enhance the ACORD-FI scheme [15]. The proposed techniques are based on the non-preemptive and preemptive scheduling schemes, both implemented on the Giving Waiting Time and the Priority of Devices approaches

A. System Models

1) Non-preemptive scheduling scheme (ACORD-FI): Before discussing our system models, we will discuss the scheme's steps of using on- and off-peak hours.

The ACORD-FI steps:

- Step 1: Define all parameters that will be used in the simulation.
- Step 2: Create the current queue as a linked list, and add the events or appliances into the current queue with two parameters. The first parameter is the type of event (start or stop), and the second is the timestamp of the event or appliance, which is computed by the Poisson process model [19], where the interarrival time between two requests is a negative exponential distribution with random numbers of the timestamp given the event or appliance using the following equations:

Where the function returns a random number between 0 and 1,

$$Generate\_RandomNumbers = rand()/(RAND\_MAX + 1)$$
$$Exponentia\,l\_Distrbution = -time * \log(Generate\_RandomNumbers) \quad (1)$$

where the function is the negative exponential distribution and the time is the interarrival time between requests.

- Step 3: Check if the timer is less than the simulation, time where the value of the timer in the beginning is zero.
- Step 4: Return the event with the smallest timestamp.
- Step 5: Select the smallest timestamp of the event in the current queue. If the case is a start event, schedule the stop event for this event in the current queue. If the case is a stop event, calculate the energy consumption and the cost of energy consumption of the appliances through the period using the following equations, and schedule the start event for this event in the current queue.

$$Energy\_Consumption = (Power * Dtime)/100 \quad (2)$$

where power is the energy consumption of the appliance in watts and Dtime is the time the appliance takes to finish the work. Divide the result by 1000 to convert it to kilowatts.

$$Cost\_Energyconsumption() = Energy\_Consumption * \cos t\_on - or - off - Peakhours()/1000 \quad (3)$$

The energy consumption in (3) is computed kilowatts, and cost_on–or–off–Peakhours is the price of the period in the on- or off-peak hours in cents for Ontario and fills for Jordan. The total cost is calculated in Canadian dollars.

2) Non-preemptive scheduling scheme (ACORD-FI): In this scheme, without using on- and off-peak hours, we change the equation for calculating the cost of energy consumption used with on- and off-peak hours (4).

The ACORD-FI steps are the same as with using the on- and off-peak hour scheme's steps, except it differs when calculating the cost of the appliances' energy consumption through the period using the following equation:

$$Cost\_Energyconsumption() = Energy\_Consumption * \cos t\_on - and - off - Peakhours()/1000 \quad (4)$$

where the energy consumption is computed by (3) in kilowatts, and the cost_on–and–off-Peakhours is the average of prices of both periods where the prices are in Ontario or in Jordan. The total cost is calculated in Canadian dollars.

We are going to enhance the effectiveness (reducing energy consumption and total cost of energy consumption) of the non-preemptive scheduling scheme using two schemes: a preemptive scheduling scheme by giving priority (weight) for several appliances and a preemptive scheduling scheme by giving waiting time (delay).

B. Proposed Schemes

1) Appliances Coordination by Giving Waiting Time (ACORD-WT) scheme: The preemptive scheduling scheme is considered one of the most effective scheduling techniques to reduce energy consumption and the total cost of the energy consumption.

This scheme differs from the ACORD-FI scheme [15] by developing a new technique for choosing the best event. The ACORD-FI scheme [15] depends on choosing the event with the smallest timestamp, unlike the ACORD-WT scheme for some appliances, which depends on choosing the event with the smallest timestamp and amount of energy required.

The ACORD-WT scheme is an energy management scheme, and it is an enhancement of the ACORD-FI by including scheduling algorithms that give waiting time to several devices.

In the ACORD-WT scheme, the user may operate a device at any time regardless of the peak hours (on-peak hours or off-peak hours). When the user operates a device, the device communicates with the EMU to check for the smallest timestamp for devices. The interval between the start time and the requested start time is computed by the EMU is called the waiting time, and it is sent back to the device. The message flow for the ACORD-WT scheme is given in Fig. 5. The device generates a START-REQ packet and sends it to the EMU. The START-REQ packet contains the type of request and the device cycle.

When the EMU receives the START-REQ packet, the EMU links with the storage unit by generating an AVAIL-REQ packet and retrieves the amount of the available energy. The storage unit replies with the amount of available energy to the EMU with an AVAIL-REP packet.

After receiving the AVAIL-REP packet, the EMU specifies the starting time of the device by using Algorithm 1, as shown in Fig. 6. The user decides whether to start the device or wait, depending on the waiting period for each device. The user's decision is sent back to the EMU with a notification (NOTIFICATION) packet.

The EMU sends an update available (UPDATE-AVAIL) packet to the storage unit to update the amount of available energy on the unit after receiving the user's decision. The algorithm of the ACORD-WT scheme (Algorithm 1, Fig. 6) works as follows. The EMU first checks if the stored energy is available, and the devices will be operated immediately; otherwise, the devices will be operated depending on the waiting period given for each device.

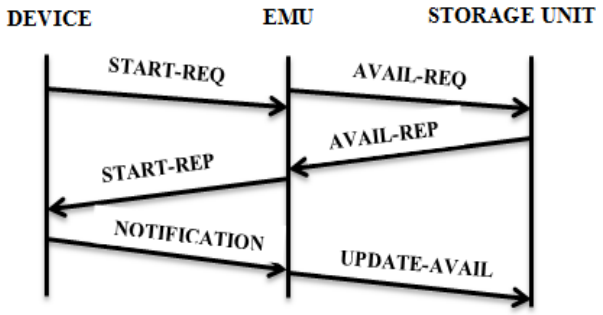


Fig. 5. Message flow for ACORD-WT scheme

In the below the specification of the scheme steps:

The first five steps are the same as the ACORD-FI scheme [15], but it differs in phase seven after we select the smallest timestamp of the event in the current queue. If the case is a start event, we calculate the energy consumption, which is computed by (3), and the energy required by adding the value of the current energy with the value of the event's energy consumption using the following equations:

$$\text{Re}\,quired\_Energy = Current\_Energy + Energy\_Consumption() \quad (5)$$

After we check if the amount of energy required is less than or equal to the amount of the energy threshold, where the energy threshold is equal to 2.47 kWh when using four devices and 8.16 kWh when using six devices, and if the condition is true, we schedule the stop event for this event in the current queue. We then add the value of energy consumption to the current energy, where the value of the current energy, in the beginning, is zero, using the following equations:

$$Current\_Energy = Current\_Energy + Energy\_Consumption() \quad (6)$$

But if the condition is not true, we remove the event from the current queue and schedule the stop event for this event in the current queue by changing the timestamp parameter, which equals the negative exponential distribution [19] of the timestamp computed by (2) with the value of the current timer.

In the stop case, we calculate the energy consumption and the cost of appliances' energy consumption through the period by using (3) and (4), and then we subtract the value of energy consumption from the value of the current energy using the following equation, (7), and schedule the start event for this event in the current queue. The total cost is calculated in Canadian dollars for Ontario and Jordan. The ACORD-WT algorithm is given in Fig. 6.

$$Current\_Energy = Current\_Energy - Energy\_Consumption() \quad (7)$$

Algorithm 1 | Preemptive Scheduling Scheme bygiving Waiting Time.

```
1:  {Wti : waiting time of appliance i}
2:  {Sti : requested start time of appliance i}
3:  {T: Timer}
4:  {S: Simulation time}
5:  While T < S
6:     PickSmallest_TimeStamp()
7:     If (stored energy available = TRUE) Then
8:        StartImmediately()
9:     Else
10:       If (Sti is in peak) Then
8:           Wti ← shift to another time()
9:        Else
10:          StartImmediately()
11:    End If
12:  End If
```

Fig. 6. Appliances Coordination by Giving Waiting Time scheme (ACORD-WT) algorithm

2) Appliances Coordination by Giving Priority (ACORD-P) scheme: The implementation of this scheme differs from the previous schemes in scheduling the event and the technique for selecting the best event.

The ACORD-P scheme for some appliances depends on choosing the event with the smallest timestamp, checking if the current queue is empty or not, and the amount of energy required.

The ACORD-P scheme is an energy management scheme and is proposed to enhance the ACORD-FI, which includes the locally generated energy with energy management decisions by giving priority to some devices.

In the ACORD-P scheme, the user may operate a device at any time regardless of the peak hours (on-peak hours or off-peak hours). When the user operates a device, the device communicates with the EMU to check for the smallest timestamp for devices.

The interval between the start time and the requested start time, called the waiting time, is computed by the EMU and sent back to the device. The device generates a START-REQ packet and sends it to the EMU. The START-REQ contains the type of request, the device cycle, and the weight of the device.

When the EMU receives the START-REQ, the EMU communicates with the storage unit by generating an AVAIL-REQ and retrieves the amount of the available energy. The storage unit replies with the amount of available energy to the EMU with an AVAIL-REP.

After receiving the AVAIL-REP, the EMU determines the starting time of the device by using Algorithm 2, as shown in Fig. 7.

The user decides which device to start or wait, depending on the rank of priority (the device that has maximum weight is the first in the operating list). The user's decision is sent back to the EMU with a NOTIFICATION.

The EMU sends an UPDATE-AVAIL to the storage unit to update the amount of available energy on the unit after receiving the user's decision.

The algorithm of ACORD-P scheme (Algorithm 2, Fig. 7) works as follows. The EMU, in the beginning, checks if the stored energy is available, and the devices will be operated immediately, and the amount of energy in the storage unit will be updated; otherwise, the devices will be operated depending on the rank of priority (weight) for each device.

The following are the specifications of the scheme steps:

- Step 1: Define all the parameters used in the simulation.
- Step 2: Create the current queue as a linked list and array called "running appliances array," and then add the events or appliances into the current queue with two parameters. The first parameter is a type of event (start or stop), and the second is a timestamp of an event or appliance that is computed by the Poisson process model [19], where the interarrival time between two requests is a negative exponential distribution with random numbers of the timestamp given of the event or appliance, which is computed by (2).
- Step 3: Check if the timer is less than simulation time.
- Step 4: Check if the current queue is empty or not.
- Step 5: Return the event with the smallest timestamp.
- Step 6: Select the smallest timestamp of the event in the current queue.
- Step 7: In the current queue, we check if the case is a start event, and we calculate the energy consumption, which is computed by (3), and energy required, which is computed by (6).

After that, we check if the amount of energy required is less than or equal to the amount of the energy threshold, where the energy threshold is equal to 2.47 kWh when using four devices and 8.16 kWh when using six devices. If the condition is true, we schedule the stop event for this event in the current queue, and then we add the value of energy consumption to the current energy, where the value of the current energy in the beginning is zero, which is computed by (7).

- Step 8: Return the maximum weight for all events in the running appliances array.
- Step 9: Select the maximum weight of the event in the running appliances array.
- Step 10: In the array, if the case is a start event, we calculate the energy consumption, which is computed by (3), and the energy required, which is by (6).

After that, we check if the amount of energy required is less than or equal to the amount of the energy threshold, where the energy threshold is equal to 2.47 kWh when using four devices and 8.16 kWh when using six devices. If the condition is true, we schedule the stop event for this event in the current queue, and then we add the value of energy consumption to the current energy, where the value of the current energy in the beginning is zero, which is computed by (7). But if the condition is not true, we remove the event from the current queue and schedule the start event in the running appliances array with three parameters. The first parameter is the type of event (start or stop). The second parameter is a timestamp of the event or appliance computed by the Poisson process model [19], where the interarrival time between two requests is a negative exponential distribution with random numbers of the timestamp given of the event or appliance using (2). The third parameter is the weight, where the weights are specified by the questioner for the appliances washer, dishwasher, dryer, Cofee Maker, PHEV, and AC, and the weight for each device is 9, 4, 5, 8, 6, and 7, respectively.

In the stop case, we calculate the appliances' energy consumption and cost of energy consumption through the period by using (3) and (4) and then subtract the value of energy consumption from the value of the current energy in (8) and schedule the start in the running appliances array. The total cost is calculated in Canadian dollars for Ontario and Jordan. The ACORD-P algorithm is given in Fig. 7.

Algorithm 2 | Preemptive Scheduling Scheme by giving Priority.

```
1:  {Cq : current queue}
2:  {Wi : weight of appliance i}
3:  {Sti : requested start time of appliance i}
4:  {T: Timer}
5:  {S: Simulation time}
6:  While T < S
7:     PickSmallest_TimeStamp()
8:  If (Cq is full = = TRUE) Then
9:     If (stored energy available = TRUE) Then
10:       StartImmediately_SmallestTimeStamp()
11:    Else
12:       Wi  ← move to running appliances array()
13:    End If
14:  Else
15:    StartImmediately_MaximumWeight()
16:  End If
```

Fig. 7. Appliances Coordination by Giving Priority (ACORD-P) scheme algorithm

Until the current queue becomes empty, we go from working on the running appliances array, to select the best event, depending on the maximum weight from all events. After that, we select the maximum weight of the event in the running appliances array. If the case is a start event, we update with stop event for this event in the running appliances array. If the case is a stop event, we calculate the energy consumption and the cost of the energy consumption of the appliances through the period in (3) and (4), and we update the start event with the stop for this event in the running appliances array. The total cost is calculated in Canadian dollar for Ontario and Jordan.

3) Appliances coordination by adding photovoltaic (PV) power source: In this paper, we suggest the use of a solar photovoltaic (PV) power supply. One solar panel with two hours of effective energy generation in winter will generate 350 w, and the feed-in tariff rates is 80.2 cents/kWh [15]. For our experiments with simple calculations, it is 28.07 cents/kWh for the Jordan tariff rate. During experiments, we used the stored power from the PV for the appliances' request in on-peak hours, and they were not used in off-peak hours. Therefore, we could guarantee full utilization benefits of the generated power. The on-peak time, as suggested, was between 7 and 11 am, which is the best time for the PV to renew the power in its batteries for future use.
The scenario used during experiments is as follows:

When the appliance (event) request arrives at the scheduling system, it checks if this request is in on-peak time. If yes:

- The system checks the energy amount of power needed and the available stored power from the PV. If it is enough, it immediately will start the event.
- If the energy amount stored is less than what is required, the system checks the end time of the event request, depending on appliance duration time. If this request ends at off-peak, the event starts immediately, also.
- If the energy amount is less than what is required and the end time is in on-peak, it moves this event to other scheduling algorithms used in the system.

If the appliance (event) request arrives at the scheduling system at off-peak time, the system will not do anything regarding the PV power stored and will run the event from the ordinary power using the other scheduling algorithms discussed previously.

IV. Experiments and Results

This section discusses the different experiments and results conducted to assess the performance of the proposed approaches discussed in Section III. We evaluate the performance by measuring the total cost of energy consumption for home appliances. The efficiency of the ACORD-FI scheme and the different proposed scheduling schemes have been tested based on the pricing in both Ontario and Jordan.

A. Simulation Parameters and Assumptions

Table 3 presents the main parameters of the simulation that applies to all schemes. All of these parameters and assumptions used are similar to those used in [15], except what applies to Jordan is proposed by us and based on the information gathered from the Ministry of Utilities in Jordan. The user requests are modeled as a Poisson process, during peak periods the interarrival times between two requests are negative exponentially distributed with a mean a 12 hours while during off-peak period, the interarrival times between two requests are negative exponentially distributed with a mean of 1 hour.

TABLE III. Parameters Used for Each Scheme

	Parameter	Value
1	Simulation time	210 days (approximately seven months).
2	Total number of devices	six devices: washer, dishwasher, dryer, Cofee Maker, PHEV, AC
3	Interarrival time	Poisson process, with a negative exponential distribution, during peak periods with a mean of 12 hours and during off-peak periods with a mean of 1 hour.

The details of the simulated devices used have been discussed in detail in Section I, including the cycle duration and energy consumption. The peak hours are selected for both Ontario [27] and Jordan as determined in Table 4.

TABLE IV. TOU Rates in Ontario as of 2011 and Jordan as of 2015

	Period	Time	Rate
Ontario	On-peak On-peak Off-peak	7:00am to 11:00am 5:00pm to 9:00pm 9:00pm to 7:00am	9.3 cent/kWh 9.3 cent/kWh 4.4 cent/kWh
Jordan	On-peak Off-peak	7:00am to 11:00pm 11:00pm to 7:00am	62.71 /kWh 52.66 /kWh

We simulate user requests between 10 days to 210 days (approximately seven months). The first 5 days from the period are spared for warm-up, and we present results as the average of 10 simulation runs.

Our simulator has been coded using in C++ under Ubuntu version 12.04.4. The main performance measure used to compare between approaches is the total savings for cost of energy consumption in dollars.

B. Results and Discussions

In the following set of figures, we will show the different sets of comparisons between four and six devices, with and without delay, with and without priorities, and using PV with and without using proposed scheduling algorithms.

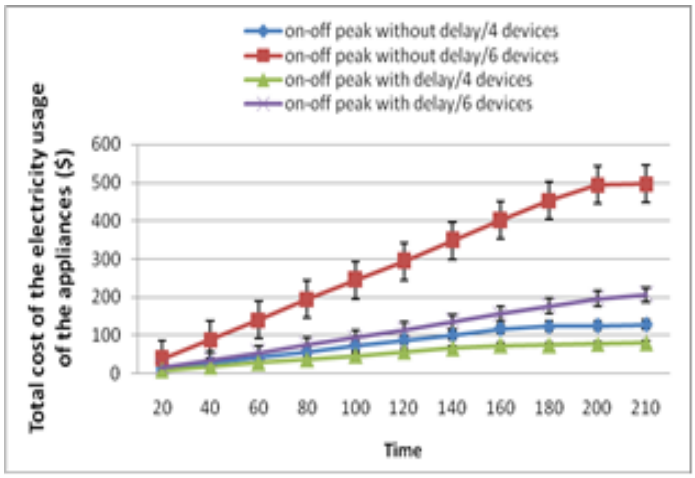


Fig. 8. The total cost of the electricity consumed by four and six devices with and without delay in Ontario

Fig. 8 shows the relation between the time and total cost of energy consumption in dollars for different cases when on–off peak is used. In this set of experiments, we used the four and six devices as discussed in Section I. The two cases considered for each set of devices are with and without delay for the requests that came during the day. The same electricity prices are used for the province of Ontario. In either case of four or six devices, using the delay reduced the total cost since the requests were delayed for the off-peak periods whenever possible. For example, in the case of four devices without delay, the total cost for 210 days was $128.1; however, the cost when a delay was used was reduced to $79.4, resulting in savings of approximately $49 for this period.

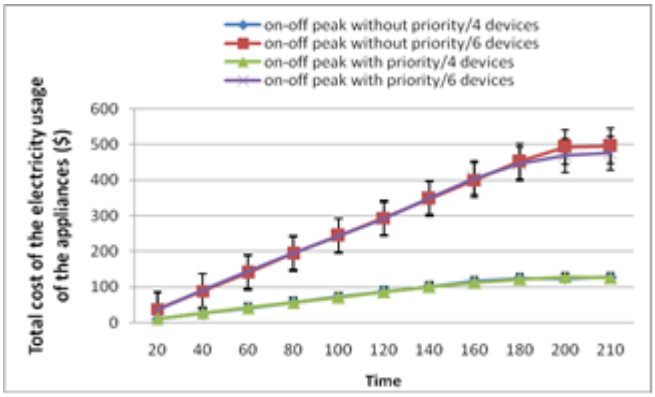


Fig. 9. The total cost of the electricity consumed by four and six devices with and without priority in Ontario

Fig. 9 shows the relation between the time and total cost of energy consumption in dollars for different cases when on–off peak is used. In this set of experiments, we used the four and six devices, as discussed in Section I. The two cases considered for each set of devices are with and without priority for the requests that came during the day. The same electricity prices were used for the province of Ontario. In either case of four or six devices, using the priority reduced the total cost since the requests were given priority to some appliances whenever possible. For example, in the case of four devices without priority, the total cost for 210 days was $128.1; however, the cost when giving priority was reduced to $125.6, resulting in a savings of approximately $3 for this period.

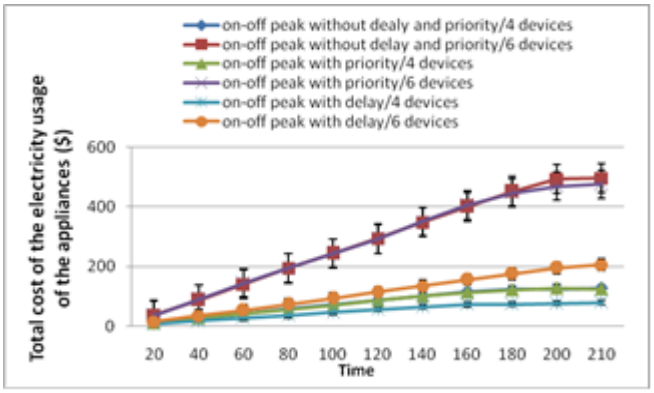


Fig. 10. The total cost of the electricity consumed by four and six devices with and without delay and priority in Ontario

Fig. 10 shows the relation between the time and total cost of energy consumption in dollars for different cases when on–off peak is used. In this set of experiments, we used the four and six devices as discussed in Section I. The three cases considered for each set of devices are with and without delay and priority for the requests that came during the day. The same electricity prices are used for the province of Ontario. In either case of four or six devices, using the delay reduced the total cost since the requests were delayed for the off-peak periods whenever possible. For example, in the case of four devices without delay and priority, the total cost for 210 days was $128.1; however, the cost when giving priority was reduced to $125.6, resulting in savings of approximately $3 for this period, while the cost when a delay was used was reduced to $79.4, resulting in savings of approximately $49 for this period.

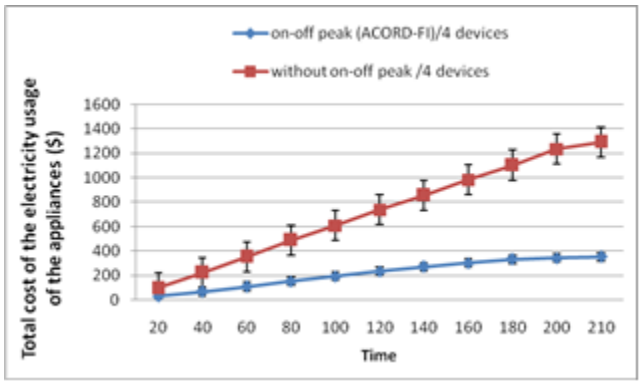


Fig. 11. The total cost of the electricity consumed by four basic devices for the cases of on–off peak and without on–off peak in Jordan

Fig. 11 shows the relation between the time and total cost of energy consumption in dollars for the cases when on–off peak is used (similar to ACORD-Fi) and the case when on–off peak is not used. In this set of experiments, we used the four basic devices discussed in Section I. The electricity prices that were applied in Jordan 2015 for the first case of on–off peak were the same prices in Table 3 for the province of Jordan. The electricity price used for the second case without on–off peak were the average price of the prices in the first case. The simulation period ran from 20 to 210 days. We can see from the figure that using the average price rather than different prices for the on- and off-peak periods costs more.

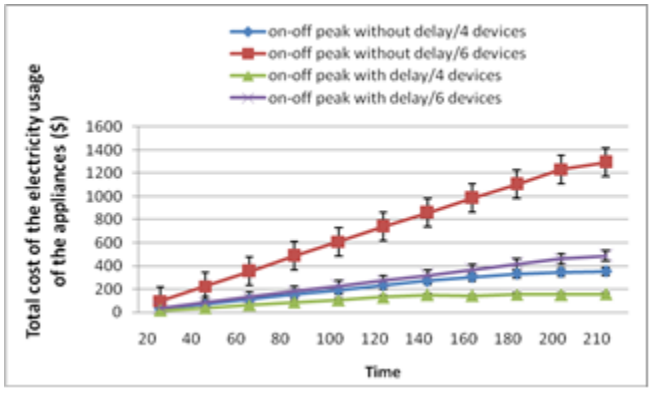


Fig. 12. The total cost of the electricity consumed by four and six devices with and without delay in Jordan

Fig. 12 shows the relation between the time and total cost of energy consumption in dollars for different cases when on–off peak is used. In this set of experiments, we used the four and six devices as discussed in Section I. The two cases considered for each set of devices are with and without delay for the requests that came during the day. The prices are the same as in Table 3 for the province of Jordan. In either case of four or six devices, using the delay reduced the total cost since the requests were delayed until the off-peak periods whenever possible. For example, in the case of four devices without delay, the total cost for 210 days was $353.3; however, the cost when delay was used was reduced to $156.6, resulting in a savings of approximately $197 for this period.

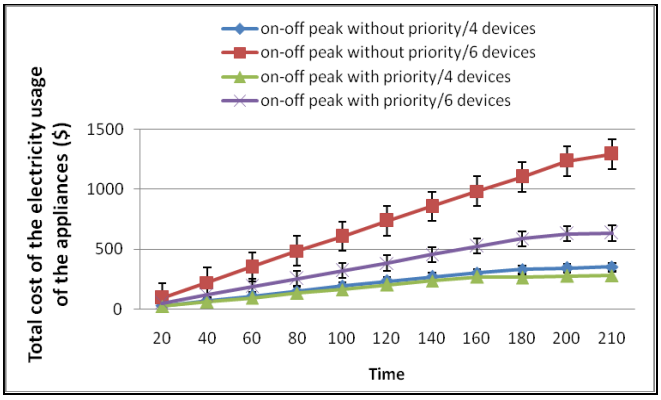


Fig. 13. The total cost of the electricity consumed by four and six devices with and without priority in Jordan

Fig. 13 shows the relation between the time and total cost of energy consumption in dollars for different cases when on–off peak is used. In this set of experiments, we used the four and six devices as discussed in Section I. The two cases considered for each set of devices are with and without priority for the requests that came during the day. The same electricity prices are used for the province of Jordan. In either case of four or six devices, using the priority reduced the total cost since the requests were given priority to some appliances whenever possible. For example, in the case of four devices without priority, the total cost for 210 days was $353.3; however, the cost when giving priority was reduced to $281.8, resulting in savings of approximately $72 for this period.

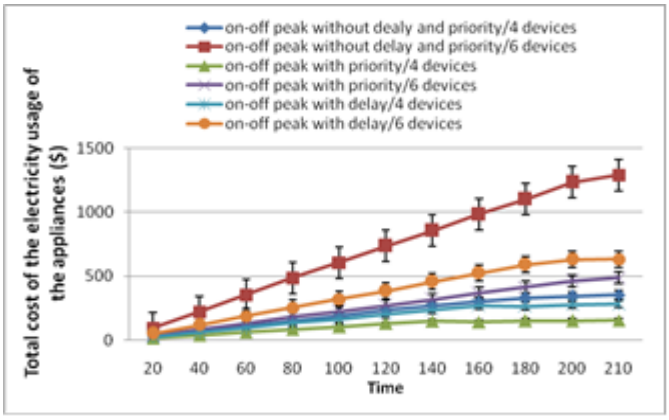


Fig. 14. The total cost of the electricity consumed by four and six devices with and without delay and priority in Jordan

Fig. 14 shows the relation between the time and total cost of energy consumption in dollars for different cases when on–off peak is used. In this set of experiments, we used the four and six devices as discussed in Section I. The three cases considered for each set of devices are with and without delay and priority for the requests that came during the day. The same electricity prices are used for the province of Jordan. In either case of four or six devices, using the delay reduced the total cost, since the requests were delayed until the off-peak periods whenever possible. For example, in the case of four devices without delay and priority, the total cost for 210 days was $353.3; however, the cost when giving priority was reduced to $281.8, resulting in a savings of approximately $72 for this period, while the cost when delay was used was reduced to $156.6, resulting in a savings of approximately $197 for this period.

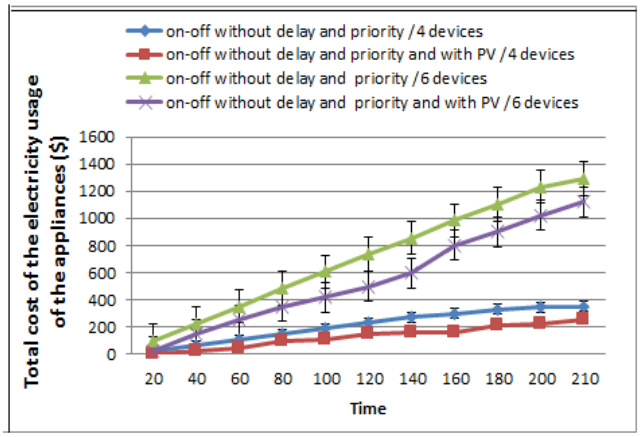


Fig. 15. The Total cost of the electricity consumed by four and six devices using and not using PV and without delay and priority in Jordan

Fig. 15 shows the total cost of energy consumption in dollars during the period of study in the case of with and without using PV and without using any type of proposed scheduling algorithms. The results show that there is a factor of saving when using PV, but still the cost is high because, in our experiments, we used one panel of PV, but in real life, more than one panel can be used, which will give more savings.

Fig. 16 shows the total cost of energy consumption in dollars during the period of study in the case of using and without using PV and with using one of proposed scheduling algorithms: delay. The results show that there is a factor of saving when using PV, and this saving comes from the idea of using the PV energy and then using the delay algorithm, which makes the cost savings in two layers.

Fig. 17 shows the total cost of energy consumption in dollars during the period of study in the case of using and without using PV and with using one of the proposed scheduling algorithms: priority. The results show that there are good savings when using PV, and it can be noticed that using 6 devices with PV is close to using 4 devices, both with priority. As a result, using PV as layer 1 in using appliances and moving to the priority scheduling algorithm gave interested results.

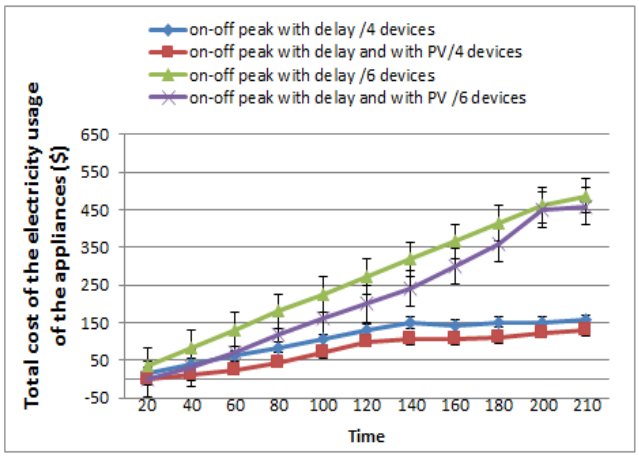


Fig. 16. The total cost of the electricity consumed by four and six devices using and not using PV and with delay in Jordan

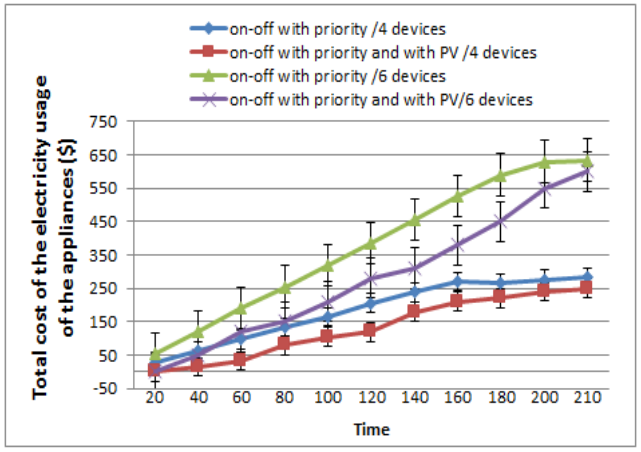


Fig. 17. The total cost of the electricity consumed by four and six devices using and not using PV and with priority in Jordan

V. Conclusions and Future Works

A. Conclusions

In this work, many experiments were conducted to study the behavior of the proposed schemes, including the preemptive scheduling scheme by giving waiting time and the preemptive scheduling scheme by giving priority. Another purpose of these experiments was to study the efficiency of the proposed schemes when we compare them against the ACORD-FI scheme when using on- and off-peak hours and without using on- and off-peak hours.

Our proposed preemptive scheduling schemes (ACORD-WT and ACORD-P) are an enhancement of the non-preemptive scheduling scheme (ACORD-FI), and they aim to reduce the total energy consumption of home appliances and reduce the total cost of energy bills.

When comparing the ACORD-FI scheme with the two proposed schemes (by giving waiting time and by giving priority) using different number of devices, the results show that the ACORD-FI scheme has the worst savings in cost for the solutions, and the ACORD-WT scheme has the highest savings of cost for the solutions, regardless of the number of devices and peak hour periods.

B. Future Work

Our proposed work improved the energy cost regarding the ACORD-FI scheme by adding new scheduling algorithms. In our future work, we will try to investigate a new scheduling algorithm to be compared with our work and add new energy sources and new devices.

References

[1] T. Anderson, T. Bon, L. Backer"Basic Electrical Terms and Definitions" North Dakota State University of Agriculture and Applied Science , 1994, pp. 1–4.

[2] M. Shabanzadeh, M. P. Moghaddam, “What is the smart grid? Definitions, perspectives, and ultimate goals,” 28th Power System Conference, Tehran, Iran, pp. 1–5, 2013.

[3] W. Wang, Z. Lu “Cyber security in the smart grid: Survey and challenges,” vol. 57, pp. 1344–71, April 2013.

[4] R. Stamminger, “Synergy potential of smart appliances,” De-liverable 2.3 of Work package 2 from the Smart-A project, University of Bonn, March 2009.

[5] I. Ipakchi and F. Albuyeh, “Grid of the future,” IEEE Power Energy Magazine, vol. 7,pp. 52–62, March-April 2009.

[6] V. Gungor, S. Bin Lu, G. Hancke, “Opportunities and challenges of wireless sensor networks in smart grid,” IEEE Transactions on Industrial Electronics, vol. 57, pp. 3557–3564, October 2010

[7] T. Li, D. Baumberger, S. Hahn, “Efficient and scalable multiprocessor fair scheduling using distributed weighted round-robin,” Proceedings of the 14th ACM Sigplan Symposium on Principles and Practice of Parallel Programming, vol. 44, pp.65–74, April 2009.

[8] I. Khan, A. Mahmood, N. Javaid, S. Razzaq, R. Khan, M. Ilahi, “Home energy management systems in future smart grids,” Nadeem Javaid, COMSATS Institute of IT, Islamabad, www.javaid.com, pp. 459–464, 2013.

[9] M. Burcea, W. K. Hon, H. H. Liu, P. W. H. Wong, D. K. Y. Yau, “Scheduling for electricity cost in smart grid,” Springer International Publishing, vol. 8287, pp. 306–317, 2013.

[10] C. S. Choi, J. Han, W. K. Park, Y. K. Jeong, I. W. Lee, “Proactive energy management system architecture interworking with smart grid,” IEEE 15th International Symposium on Consumer Electronics(ISCE) pp. 621–624, June 2011.

[11] M. Erol-Kantarci, H. T. Mouftah. “TOU-Aware Energy Management and Wireless Sensor Networks for Reducing Peak Load in Smart Grids,” IEEE Vehicular Technology Conference , pp. 1–5, May 2010.

[12] M. Erol-Kantarci, H. T. Mouftah, “Using wireless sensor networks for energy-aware homes in smart grids,” IEEE Symposium on Computers and Communications, 2010, pp. 456–458.

[13] M. Erol-Kantarci, H. T. Mouftah, “Wireless sensor networks for cost-efficient residential energy management in the smart grid,” IEEE Transactions on Smart Grid, vol. 2, pp. 314–325, June 2011.

[14] M. Erol-Kantarci, H. T. Mouftah, “Wireless sensor networks for smart grid applications,” IEEE. Saudi International Electronics, Communications And Photonics Conference (SIECPC) 2011, pp. 1–6.

[15] “Poisson process model (Negative exponential distributed),” http://www.pamvotis.org/vassis/RandGen.htm.

[16] A. Imamura, S. Yamamoto, T. Tazoe, H. Onda, H. Takeshita, S. Okamoto, et al., “Distributed demand scheduling method to reduce energy cost in smart grid,” IEEE Humanitarian Technology Conference (R10-HTC) 2013, pp. 148–153, Aug 2013.

[17] P. LI, S GUO, Z. CHENG, “Joint optimization of electricity and communication cost for meter data collection in smart grid,” IEEE Transactions on Emerging Topics in Computing, vol. 1, pp. 297–306, December 2013.

[18] A. H. Mohsenian-Rad, A. Leon-Garcia, “Optimal residential load control with price prediction in real-time electricity pricing environments,” IEEE Transactions on Smart Grid, vol. 1, pp. 120–133, September 2010.

[19] E. Pouresmaeil, J. M. Gonzalez, C. A. Canizares, K. Bhattacharya, "Development of a smart residential load simulator for energy management in smart grids," IEEE Transactions on Power System, vol 4, pp. 1-8, Feb 2013.

[20] M. Erol-Kantarci, H. T. Mouftah. "Demand Management and Wireless Sensor Networks in the Smart Grid," Energy Management system book, Aug 2012.

[21] "Hydro Ottawa TOU rates," https://hydroottawa.com/accounts-and-billing/residential/time-of-use/rate-periods .

[22] X. Chen, T. Wei, S. Hu., "Uncertainty-aware household appliance scheduling considering dynamic electricity pricing in smart home," IEEE Transactions on Smart Grid, vol. 4, no. 2 pp. 932–941, March 2013.

[23] S. Bera, P. Gupta, S. Misra, "D2S: Dynamic demand scheduling in smart grid using optimal portfolio selection strategy," IEEE Transactions on Smart Grid, vol.6, no. 3, pp.1434–1442, March 2015.

[24] K. Ma, T. Yao, J. Yang, X. Guan, "Residential power scheduling for demand response in smart grid," International Journal Of Electrical Power & Energy Systems vol. 78, pp: 320–325, 2015.

[25] M. Erol-Kantarci, H. T. Mouftah, "Wireless sensor networks for domestic energy management in smart grids," IEEE Biennial Symposium on Communications (QBSC) 2010, pp. 63–66, May 2010

Network Attack Classification and Recognition Using HMM and Improved Evidence Theory

Gang Luo
College of Computer Science and Electronic Engineering
Hunan University
Changsha, China

Ya Wen
College of Computer Science and Electronic Engineering
Hunan University
Changsha, China

Lingyun Xiang
Hunan Provincial Key Laboratory of Intelligent Processing of Big Data on Transportation
Changsha University of Science and Technology
Changsha, China

***Abstract*—In this paper, a decision model of fusion classification based on HMM-DS is proposed, and the training and recognition methods of the model are given. As the pure HMM classifier can't have an ideal balance between each model with a strong ability to identify its target and the maximum difference between models. So in this paper, the results of HMM are integrated into the DS framework, and HMM provides state probabilities for DS. The output of each hidden Markov model is used as a body of evidence. The improved evidence theory method is proposed to fuse the results and encounter drawbacks of the pure HMM for improving classification accuracy of the system. We compare our approach with the traditional evidence theory method, other representative improved DS methods, pure HMM method and common classification methods. The experimental results show that our proposed method has a significant practical effect in improving the training process of network attack classification with high accuracy.**

***Keywords*—Hidden Markov Model; Evidence theory; Network attack; KDD CUP99; Classification**

I. INTRODUCTION

With the development and popularity of Internet, the network environment in today's society is more and more complex. Security of network has become a very important problem in the network. Intrusion detection system which attempts to use data mining and machine learning methods to detect and classify intrusion activities plays an important role in detecting and preventing network attacks[1]. However, intrusion detection systems can be split into two groups: 1) anomaly-based detection system and 2) misuse-based detection system[2]. Each of them has a different way in detecting and protecting data security and has both advantages and disadvantages. The misuse-based detection system, especially the reasoning system based on model matching, can achieve high classification accuracy for known attacks. Scholars proposed various classifier models to solve classification problem in network intrusion detection, including Bayesian network, fuzzy logic, k-nearest neighbor, decision tree, neural networks, support vector machine, the hidden Markov model.

Cheng Xiang [3] proposed a multiple-level hybrid classifier, a novel intrusion detection system, which combined supervised tree classifiers and unsupervised Bayesian clustering to detect intrusions. The performance of this approach was shown to have high detection and low false alarm rates. In [4], a multiple classifier intrusion detection model was presented, which was based on a new data mining method called hidden Naive Bayesian. This method was better than other models, but it only had a high detection rate for the DoS (the denial of service) attack while the other attack detection accuracy was not high. Yuk [5] applied intelligent dynamic swarm based rough set for feature selection and simplified swarm optimization for intrusion data classification. The performance of the hybrid intrusion detection system on KDD Cup 99 dataset is better than others with high classification accuracy.

Some researchers use machine learning methods to design intrusion detection systems. Most of them are based on SVM technology which has a solid theoretical basis and can classify data records into multiple classes. Horng et al. [6] proposed an SVM-based intrusion detection system based on a hierarchical clustering algorithm to preprocess dataset before training. The simple feature selection procedure was applied to eliminate unimportant features from the training set so that the obtained SVM model could classify the network traffic data more accurately. However, this system showed better performance in the detection of DoS and Probe attacks but not very good in U2R and R2L attacks. In [7], a pipeline of the data preprocess and data mining was put forward in IDS to choose critical features. With the combination of clustering method and support vector machine, an efficient and reliable classifier was developed to judge a network. The performance of SVM was good in data classification, but not suitable for large scale dataset. Training complexity is deeply dependent on the data volume of training dataset, and the greater amount of data will lead to higher training complexity. However, many data mining applications involve millions or even billions of pieces of data records. The system failure caused by the lack of memory makes the SVM can't run such a large dataset.

The Markov model and hidden Markov model which are initially used for speech recognition (Rabiner,1989), handwriting recognition (Gunter and Bunke, 2003), biological sequence analysis (Durbin et al., 2006) have been applied to computer security model in recent years. In the field of computer security, HMM is mainly used for anomaly-based

intrusion detection systems. Warrender et al. (1999) made a pioneering work in this area and they use HMM for system modeling. HMM can also be used in network security. Ariu [8] proposed a novel solution where the HTTP payload is analyzed using hidden Markov model. The proposed system, named HMMPayl, had high classification accuracy and was very effective on most common attacks of the Web application. The core idea of attack classification is pattern recognition. Hidden Markov model can effectively describe the hidden Markov process containing unknown parameters. HMM can get hidden parameters of the process from observable feature parameters, and use these parameters to make further analysis for attack classification [9]. However, when the dimension of feature parameters space is high, the training structure is complex, the training time is very long, and the classification recognition accuracy is quite low. In intrusion detection, several attacks may show some similar features. That is, under certain features, the attacks are likely to have a certain probability of occurrence. Fusion all kinds of feature information to obtain the occurrence probability of each attack and the maximum probability of occurrence can be judged as the main attack. Therefore, the use of evidence theory [10-11] is particularly suitable for classification and recognition of information fusion.

The study of neural biology showed that the information process of biological sensing system can be divided into two relatively independent processing procedures: information unit decomposition and fusion. Such way of early decomposition and late fusion with high ability in information processing and intelligent decision [12]. According to this, the idea of multi-features fusion and decision making can be used in the classification of attacks to achieve the purpose of improving classification accuracy. Assume that there is a classification problem of N kinds of attacks. The whole attack feature parameters space is divided into K subspaces according to certain rules. Then decision model of each feature subspace is constructed to achieve the mapping of feature subspace. If use hidden Markov model, K feature subspace will lead to K hidden Markov models process, and K diagnostic results will be obtained. This process is equivalent to the decomposition of information unit. K diagnostic results of K sub hidden Markov models are then used as K bodies of evidence. By using the evidence theory to combine K bodies of evidence, the fusion of information units can be realized, and the final decision can also be made.

Therefore, in this paper, the hidden Markov model and the evidence fusion theory are applied to network security. A new information fusion system based on HMM and DS evidence theory is proposed which can effectively achieve the target of network attack classification and recognition. At the same time, a new method of evidence fusion based on entropy weight is proposed. By calculating the information entropy of source data to obtain weight of evidence, and modify the basic probability assign (BPA) of original evidence. Finally, the rule of combination is used to combine the modified BPA.

The rest of the paper is organized as follows: Section II briefly describes the principal theory of hidden Markov model and DS evidence theory. Section III presents the improved DS evidence theory and explains the details of the theoretical concept of the proposed HMM-DS system. The analysis of experimental results for KDD CUP99 using HMM-DS has been compared with C4.5, LibSVM, Naïve Bayes, which are presented in section IV. In Section V, the study concludes with a summary of the research undertaken.

II. Related Work

A. Hidden Markov Model

A hidden Markov model is a statistical model which is used to describe a Markov process containing unknown parameters. It is mainly used to determine the hidden parameters of the process from the observable parameters, and then use these parameters to make a further analysis. Figure 1 shows the general architecture of an instantiated HMM. The random variable x_i is the hidden state, y_i is possible observation, a_{ij} is transition probability matrix, and b_{ij} is emission probability matrix.

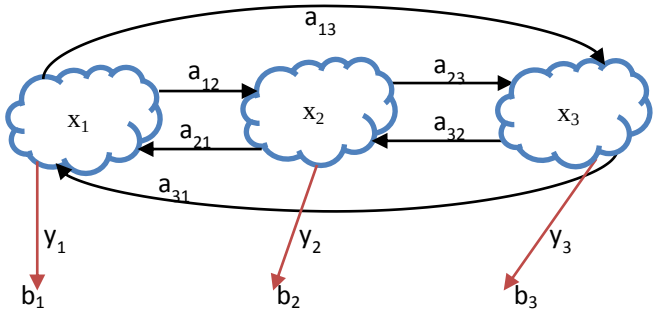


Fig. 1. The architecture of an instantiated HMM

An HMM can be described as five elements, which is $\{Q, O, \pi, A, B\}$.Q is the number of hidden states which is accurately known or guessed. O represents the number of observable states which can be achieved by training datasets. A is the matrix of state transition probabilities, B is probability distribution in each of the states which is also called the mixture matrix and π is the initial state of probability distribution. In state transition matrix and mixture matrix, each probability is independent of time. Namely, when the system is in an evolution, these matrices do not change over time. Therefore, we can use the compact notation $\lambda = \{\pi, A, B\}$ to denote an HMM with discrete probability. HMM can solve three problems [13]:

1) Evaluation problem. For a large number of sequences of HMMs (λ_1, λ_2, $\lambda_3 \ldots \lambda_k$) and observation sequence $O = \{o_1, o_2, \ldots, o_T\}$, Forward algorithm is used to calculate the probability of a given observation sequence, and then an HMM is chosen that best matches the observations.

2) Decoding problem. For a given model λ and observation sequence O, Viterbi algorithm is used to calculate the most likely sequence of hidden state.

3) Learning problem. For a given observation sequence and the related set of hidden states, Baum-Welch or Forward-Backward algorithm is applied for parameter estimation.

B. DS evidence theory

The evidence theory was first put forward by Dempster and developed by Shafer. In evidence theory, elements in the frame of discernment Θ are exclusive and exhaustive. Define m: $2^{\Theta}\rightarrow$ [0, 1] as basic probability assignment (BPA, also called mass function) satisfying: $\forall A\subset\Theta$, $m(\Phi)=0$, $\sum_{A\subset\theta} m(A)=1$

where A is called the focal element [14].

The core of DS evidence theory is the rule of combination. Two mass function m_1 and m_2, based on the evidence of two independent and reliable sources, can be combined into a new mass function by the use of conjunctive combination.

$$m_1 \oplus m_2(A)=\frac{1}{1-k}\sum_{B\cap C=A} m_1(B)m_2(C) \quad (1)$$

Where $k=\sum_{B\cap C=\varnothing} m_1(B)m_2(C), k\neq 1$ measures the conflict between m_1 and m_2. K is called the conflict coefficient. Dempster's rule of combination satisfies the associative law and the commutative law. There are n mass function (m_1, m_2,...m_n) in the frame of discernment, the conjunctive combination is calculated as

$$(m_1 \oplus \cdots m_n)(A)=\frac{1}{1-k}\sum_{\bigcap_{i=1}^{n} A_i=A} m_1(A_1)\cdots m_n(A_n) \quad (2)$$

Where $k=\sum_{\bigcap_{i=1}^{n} A_i=\varnothing} m_1(A_1)m_2(A_2)\cdots m_n(A_n)$.

As a kind of uncertain reasoning method, DS has attracted more and more attention. DS evidence theory can not only solve the problem of unknown and uncertainty, but also provides a very useful rule of combination which can help us fuse the evidence provided by multiple sources of evidence. Two key problems need to be solved in the process of DS evidence theory applied in data fusion classification. On the one hand, how to construct the basic probability assignment function of DS evidence theory, which is an important issue that must be solved in the process of combination and is not easy to determine. On the other hand, when the bodies of evidence to be combined are highly conflicting, counter-intuitive results may be obtained based on Dempster's rule of combination.

Scholars have proposed a variety of solutions to solve the issue. Some of them think that counter-intuitive results are caused by Dempster's rule, so they modified Dempster's rule to build a new combination rule. Yager [16] proposed an algorithm to distribute conflict belief to unknown proposition completely. This algorithm is more reasonable than that of D-S evidence theory in dealing with the combination paradox. However, the combination results are undesirable in combining multiple sources of evidence. In [17], Yager proposed a very interesting approach which made use of a weighted aggregation of the belief structures where the weights were related to the degree of dependence. It is too theoretical to be used in real applications. However, how to define the degree of dependence is not given. Sun [18] allocated part of the basic probability assignment of the conflict to the set of propositions supported by the evidences by a certain proportion. The difference between Yager and Sun is the proportion of the conflicts allocated. Thierry [19] presented a modified combination rule with mass function of dependent information sources. This rule used a special description of the body of evidence to ensure the combination, but the results given are very strange, and it does not consider the degree of confidence in the source of evidence. Destercke and co-workers [20] generalized the minimum rule of possibility theory, but did not respect the fundamental equivalence between belief functions and their empty set.

Some researchers deal with conflict evidences based on the method of modifying evidence source while keeping the combination rule unchanged. Haenni [21] thought it may not be the problem of combination rule when results were not matched with the real situation. However, the evidence of conflict should be modified. The rule of combination proposed by Murphy [22] was just to average all the BPAs of relevant hypothesis to get new belief assignments. This method can get good convergence effect, but the weight of each sensor in practical system is not the same. Yong Deng [23] put forward a novel sequence weighted evidence combination approach by using the variances of BOE sequences to generate the weights. In [24], the proposed method used training data to build a normal distribution model for each attribute of the data. Then, a nested structure BPA function was constructed by using the relationship between the test data and normal distribution model. To deal with the outer dependence, Su [25] proposed a model based on the intersection of influencing factors identified during the information propagating and evaluating process. The relative weights of BPAs for a specific element in the outer dependence phase and the relative weights of elements in the inner dependence phase were used as the discount coefficient in the discounting operator.

III. Decision Model of Information Fusion Classification Based on HMM-DS

A. Improved DS evidence theory

In this paper, the feature space is divided into K subspaces according to the character of the feature space, and the function of each subspace is different. Some features of KDD CUP99 may be irrelevant and some others may be redundant. The importance of each feature subspace is also different. The result is more accurate by obtaining the importance coefficient of features to get new basic probability assignment. In this paper, based on the information entropy of the specific features of data source, the entropy weight to determine the importance coefficient as the weight of the fusion feature is obtained.

The basic idea of entropy method is depended on the variability of indices to determine the objective weight. The smaller the information entropy of the index is, the greater the degree of variation of the index value, and the more information it will provide. In the comprehensive evaluation, the index can play a bigger role with high weight. On the contrary, the greater the information entropy of the index is, the smaller the degree of variation of the index value, and the less information will be provided. Then the index just plays a smaller role with low weight [26].

The following procedure has been used in building the weight vector.

1) Assume that all the features of the source data are: $(X_1,X_2,X_3,\ldots,X_n)$, and $X_j=\{x_{1j}, x_{2j},\ldots x_{mj}\}$,which represents m record of evaluating data with n features. Equation (9) is used for data standardization and a data matrix $R=(r_{ij})_{m\times n}$ will be get after standardization of all indexes.

$R=\begin{Bmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ r_{21} & r_{22} & \cdots & r_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ r_{m1} & r_{m1} & \cdots & r_{mn} \end{Bmatrix}$, r_{ij} represents the data value of j *-th* index of the i *-th* record.

2) Calculate the proportion of the index value of the j *-th* index of the i *-th* item $p_{ij}=r_{ij}/\sum_{i=1}^{m} r_{ij}$. If $p_{ij}=0$, define $\lim_{p_{ij}=0} p_{ij}\ln p_{ij}=0$.

3) The information entropy of the j *-th* feature index is:

$$E_j=-\ln(m)^{-1}\sum_{i=1}^{m} p_{ij}\ln p_{ij} \tag{3}$$

4) According to the calculation formula of information entropy, the information entropy of each index is calculated like $S_i=(E_1,E_2,\ldots Ep), i=1,2,\ldots k$. The sum of the indexes of feature subspace is calculated by $H_i=\sum_{j=1}^{p} E_j$, and weight coefficient of H_i obtained as follows:

$$w_i=\frac{1-H_i}{k-\sum_{i=1}^{k} H_i} \quad i=1,2,\cdots,k \tag{4}$$

5) Based on weight coefficients of each evidence, weight vector can be obtained. $W=(w_1,w_2,\cdots,w_k)$. The basic probability assignment $m_i(A_j)$ of each element in the frame of discernment was modified by the weight vector.

$$m_i^*\left(A_j\right)=w_i * m_i\left(A_j\right) \tag{5}$$

6) In the equation(5), $j=1,2,\ldots,n$, n is the number of focal elements in the frame of discernment except for Φ. But the sum of basic probability assignment $m_i^*(A_j)$ value after adjustment is not 1 which does not meet the requirements of basic probability assignment function definition. In order to satisfy the definition of basic probability assignment, a definition is in need.

$$m_i^*\left(\Theta\right)=1-\sum_{j=1}^{n} m_i^*\left(A_j\right) \tag{6}$$

The basic probability distribution function is defined by (5) and (6). Finally, the combination in (2) is used to combine the modified evidences.

B. HMM-DS System Design

The whole feature parameter space is divided into several sub parameter spaces, and then a HMM model is designed for each feature parameter subspace. Parameters of each HMM model will be constructed with training data. After that, the models have the ability to learn. Meanwhile, these sub-models can form the preliminary judgment layer. Outcome probabilities based on every sub-HMM model can be obtained and these probabilities will act as basic probability assign of evidences in the frame of discernment. In the meantime, considering the problem of consistency variation among evidences, the improved evidence theory is used to fuse them to get the result of the cooperation of each sub model, and to improve the recognition accuracy of attack classification.

The combination of HMM and DS evidence theory can have complementary advantages, and it is beneficial to improve the speed and accuracy of classification identification. In this paper, the HMM-DS attack fusion classification decision model is shown in Figure 2. In this model, the whole classification process is divided into two layers: a preliminary identification layer based on HMM; a fusion decision layer based on HMM-DS evidence theory.

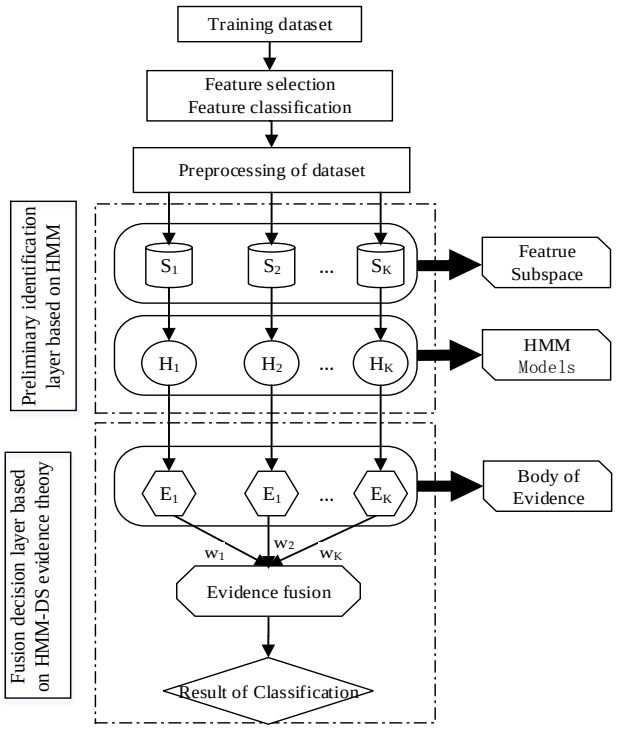


Fig. 2. HMM-DS system design

1) Preliminary identification layer based on HMM

Firstly, the feature parameter space S is divided into K sub-parameter space S_i, According to the definition of the parameter space, the corresponding learning dataset of sub HMM is obtained. Secondly, each independent sub hidden Markov model H_i is constructed and trained with learning dataset, and is capable of learning. Finally, the corresponding test samples are used to test the trained hidden Markov model, and the results obtained are the basis for the fusion decision layer in the next step.

2) Fusion decision layer based on HMM-DS theory

In the fusion decision layer, the output of each sub model in the preliminary recognition layer is used as a body of evidences E_i. The improved evidence combination method is used to fuse the evidence and obtain the final decision results

to achieve the attack classification. The creative process consists of the following five steps:

a) Establish a framework of discernment. According to expert experience and previous history records, establish an identification framework of discernment, $\Theta = \{A_1, A_2, \ldots, A_M\}$. In this paper, the proposition in the framework of recognition is corresponding to the attack type: Normal, Probe, Dos, U2R, R2L.

b) Construction of evidence. The output of each sub-HMM model is used as a body of evidence.

c) Calculate BPA of every element in the frame of discernment. DS evidence theory does not give the general calculation method of basic probability assignment, and methods used in relative papers were also different. The evaluation problem of HMM can obtain probability according to the observation sequence. Therefore, the BPA can be directly obtained by the probability calculated by the forward algorithm. As DS requires the sum of BPA of all elements must be 1, the probability gotten from HMM need to be normalized. The evidence H_i assigns to the BPA of proposition A_j can be expressed as follows:

$$m_i(A_j) = \frac{H_i(A_j)}{\sum_{j=1}^{M} A_j} \quad M \text{ is the number of attack type} \qquad (7)$$

Original evidences are modified by the improved evidence method in this paper, and new $m_i^*(A_j)$, $m_i^*(\Theta)$ is acquired

d) Evidence combination. DS evidence fusion method can be used to calculate the fused BPA $m(A_j)$.

e) Decision making. Decision methods used in evidence theory includes: decision making based on belief function, decision making based on minimum risk and decision making based on basic probability assignment. In this paper, the third method was used. That is, if $A_1, A_2 \subset U$, satisfy $m(A_1) = max\{m(A_i), A_i \subset U\}$; $m(A_2) = max\{m(A_i), A_i \subset U \text{ and } A_i \neq A_1\}$.

$$\text{If} \quad \begin{Bmatrix} m(A_1) - m(A_2) > \varepsilon_1 \\ m(\Theta) < \varepsilon_2 \\ m(A_1) > m(\Theta) \end{Bmatrix} \qquad (8)$$

Where A_1 is the result of the decision. Among them, ε_1 and ε_2 are the predefined thresholds. Θ is the uncertainty set.

IV. Experimental results and analysis

The experiments run in an Intel Pentium 2.7 GHz computer with 2.0G memory running Windows7.The code for data processing and data mining is written in MatlabR2014a.

A. KDD Cup99 Dataset description

The experiment data used in this paper is a benchmark database downloaded from KDD Cup99 [27]. This dataset contained a wide variety of intrusion simulated in a military network environment. It consists of two dataset, the training dataset and test dataset. Each network connection record is marked as Normal or Attack. The classification of attack behavior is a 5- class problem, and each network connection belongs to one of the following behavior: normal, denial of service (DOS), unauthorized access from a remote machine (U2R), unauthorized access to local supervisor privileges (R2L), probing. The test dataset includes some specific attacks that do not appear in the training dataset to make the task more difficult and realistic, which contains 24 training attack types, with additional 14 types in the test dataset only. KDD CUP99 is mainly used for binary classification (normal and attack) and multiple classification (normal and four kinds of attack).

The following data shows the connection record data format, and each feature records separated by a comma. Each record in the KDD Cup99 data set contains 41 various quantitative and qualitative features which can be divided into three groups: basic features of the network connection, features based on the content of the network connection and features based on time flow in 2 second. The last feature is the label.

2, tcp, smtp, SF, 1684, 363, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0.00, 0.00, 0.00, 0.00, 1.00, 0.00, 0.00, 104, 66, 0.63, 0.03, 0.01, 0.00, 0.00, 0.00, 0.00, 0.00, normal.

0,icmp,ecr_i,SF,1032,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,511,511,0.00,0.00,0.00,0.00,1.00,0.00,0.00,255,255,1.00,0.00,1.00,0.00,0.00,0.00,0.00,0.00,smurf.

0,udp,private,SF,28,0,0,3,0,0,0,0,0,0,0,0,0,0,0,0,0,2,2,0.00,0.00,0.00,0.00,1.00,0.00,0.00,255,2,0.01,0.02,0.01,0.00,0.00,0.00,0.77,0.00,teardrop.

The data has been preprocessed before using for training and testing of the classification model. The preprocessing of dataset has been explained in section 4.B.

B. Data preprocessing

The standard KDD Cup99 dataset is in text format. Some of the 41 features are irrelevant, and some others may be redundant, which can reduce efficiency and lead to wrong results. In this paper, after use of general feature selection techniques for feature simplification, features with the same value and less value are deleted. Finally, features that can improve the classification accuracy and running efficiency of the algorithm are selected. With this, data size reduction by reducing a number of features from 41 to 35 is shown in Table 1.

TABLE I. Features selection

Feature groups	Features
Basic	duration,protocol_type,service,flag,src_bytes,dst_bytes,wrong_fragment
Content based	hot,num_failed_logins,logged_in,num_compromised,root_shell,num_root,num_file_creations,num_access_files,is_guest_login
Time based	count,srv_count,serror_rate,srv_serror_rate,rerror_rate,srv_rerror_rate,same_srv_rate,diff_srv_rate,srv_diff_host_rate,dst_host_count,dst_host_srv_count,dst_host_same_srv_rate,dst_host_diff_srv_rate,dst_host_same_src_port_rate,dst_host_srv_diff_host_rate,dst_host_serror_rate,dst_host_srv_serror_rate,dst_host_rerror_rate,dst_host_srv_rerror_rate

As the dimension of the dataset is quite different which makes the running time longer, it needs to be standardized. The most common standardization method is Z-score which is called zero-mean normalization. After preprocessing, the data conform to the standard normal distribution that the mean is 0 and standard deviation is 1. This is given by:

$$z(x) = \frac{x - \bar{x}}{s(x)} = \frac{x - \bar{x}}{\sqrt{\frac{(x - \bar{x})^2}{n}}} \quad (9)$$

Where x is the original data, $\bar{x}$ is the mean of all data, and n is the number.

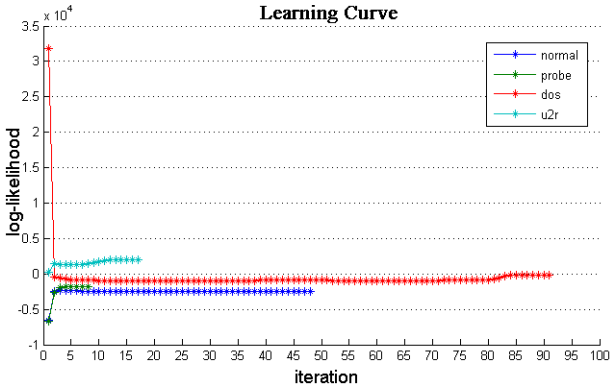
C. Results in preliminary identification layer

After feature selection, 35 features are used to form attack feature set S, which represents the type of attack. According to the classification of features, 35 features can be divided into three groups: Basic, content based, time-based features. According to the four attack types and normal of KDD Cup99, the sample set of each feature subspace is formed. Table 2 shows the number of records for each attack type in the training and test datasets, respectively. The network attack sample set we used is established as R.

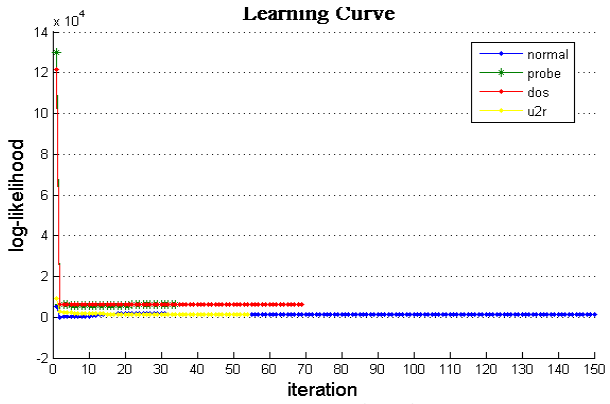
TABLE II. CONNECTION RECORDS OF TRAINING AND TEST SETS

Type of connection	Available training set	Training set	Available test set	Test set
Normal	972780	8000	60593	605
Probe	41102	4000	4166	1359
DoS	3883370	8000	229853	2230
R2L	1126	1126	16189	1618
U2R	52	52	228	228

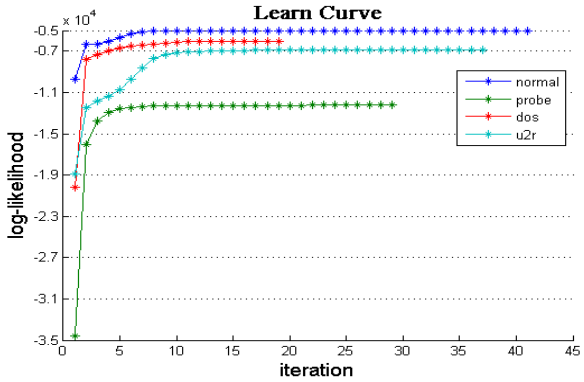
The network connection feature parameter space S is divided into three sub spaces S_j ($j=1, 2, 3$), S_1 is basic features, S_2 is content based features and S_3 is time based features. According to the definition of the feature subspace, feature parameter values are chosen from attack sample set R to consist of attack training sample set of feature subspace FR_{ij}. For each feature subspace S_j, the HMM has been trained for learning. While training the model, it is necessary to initialize appropriate values $\lambda_0 = \{\pi_0, a_0, b_0\}$, as the performance of the model mainly depends on these values. In this paper, initial parameters are generated randomly. After initialization of parameters λ_0, the model selection is a major issue. Standard Baum-Welch algorithm and EM algorithm are used to train the model. The forward algorithm is suitable to test the network traffic. Then the model parameters are $\lambda_1 = \{\pi_1, a_1, b_1\}$ after training. The learning curve of model training is shown in Figure 3.



(a) HMM training process of sub feature space S_1



(b) HMM training process of sub feature space S_2



(c) HMM training process of sub feature space S_3

Fig. 3. The learning curve of model training of sub feature space

After training the parameters, the model has the ability to learn, and is tested by test sample data. A test data of DoS attack is chosen to be evaluated with four kinds of sub-HMMs respectively. The results are shown in Table 3.

TABLE III. LOG LIKELIHOOD VALUE IN PRELIMINARY IDENTIFICATION

	S_1	S_2	S_3	loglik
Normal	-54.57	0.12	-32.4	-86.85
Probe	-33.79	5.31	-11.13	-39.61
DoS	-6.22	5.55	-25.82	-26.49
R2L	-62	-1.3	-31.2	-94.5
U2R	-16.12	-0.97	-55.53	-72.62

Log likelihood(loglik) represents the match value between the test data and the HMM (three parameters: prior1, transmat1, obsmat1). The bigger loglik value means matching better. In table 2, the maximum of the sum of the log-likelihood probability of all sub features is DoS. So, the initial judgment for the test sample is DoS.

D. Results in fusion decision layer

The frame of attack discernment is established as $\Theta = \{A_1, A_2, A_3, A_4, A_5\}$, where A_i is attack type : Normal, Probe, Dos, U2R or R2L.The forward algorithm of evaluation function of HMM is used to calculate the probability of an observed sequence with the given hidden Markov model. The output of each feature model H_{ij} is as the body of evidence. The equation (7) is used to obtain basic probability assign of the proposition

A_i distributed by all evidences. Table 4 shows the performance comparison of this method with classical DS, Yager method, Sun Quan method, and Murphy method. Calculate the average results of 10 times using DoS test dataset. Classification performance of several fusion methods is showed as follows.

TABLE IV. PERFORMANCE COMPARISON OF THE PROPOSED APPROACH WITH OTHER FUSION METHODS

Threshold	Fusion method	Time/s	Classification accuracy%
$\varepsilon_1 = 0.8$ $\varepsilon_2 = 0.1$	DS	4.978446	60.50
	Yager	4.688622	36.30
	Sun Quan	4.872527	79.30
	Murphy	4.748692	83.43
	Our method	4.617725	88.42
$\varepsilon_1 = 0.7$ $\varepsilon_2 = 0.15$	DS	4.662461	72.30
	Yager	4.696050	45.25
	Sun Quan	4.788757	83.01
	Murphy	4.479943	88.40
	Our method	4.607600	93.36
$\varepsilon_1 = 0.6$ $\varepsilon_2 = 0.2$	DS	4.872562	79.00
	Yager	4.953292	53.75
	Sun Quan	4.919677	87.39
	Murphy	4.744361	86.20
	Our method	4.870531	95.83

From the table, it can be concluded that this evidence fusion method has the best classification accuracy with different thresholds compared with other methods. The time required for classification is almost the same. When ε_1=0.8 and ε_2=0.1, the classification accuracy can reach 88.42%, while ε_1=0.6 and ε_2=0.2, the accuracy can reach 95.83%. In addition, in order to compare with the classification results of pure HMM method, traditional hidden Markov model method is applied to classify attacks. Here, the model is trained with training samples from training data set as in the above case. Results of training process is in Figure 4, and Table 5 shows the comparison between HMM-DS and HMM.

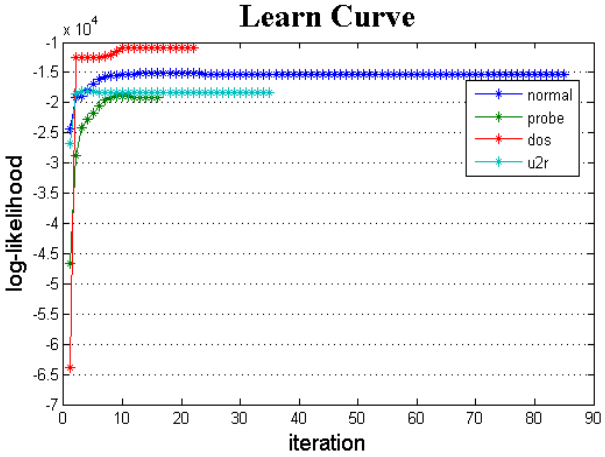


Fig. 4. HMM training process of feature space S

TABLE V. COMPARISON OF LOG LIKELIHOOD AND TIME IN DOS MODEL BUILDING

	Hmm_1	Hmm_2	Hmm_3	Hmm
loglik	-2460.63	-5307.02	-6342.99	-10850.15
Running time/s	20.7	101.56	78.2	273.88

Compare Figure 3 and Figure 4, log likelihood of feature sub space after training is less than the value of the whole feature space. Therefore, parameters are better optimized, and the total training time of HMM-DS is lower than of pure HMM with improved accuracy. Select test samples from each attack test data set. The classification performance of each attack type is shown in Table 6 and Table 7.

TABLE VI. RESULTS OF HMM AND DS

Attack type	Normal	Probe	DoS	R2L	U2R	Classification accuracy (%)
Normal	580	14	6	4	1	95.8
Probe	61	1280	17	1	0	94.1
DoS	122	58	2050	0	0	91.9
R2L	102	1	0	1490	25	89.6
U2R	19	2	0	4	195	88.6

TABLE VII. RESULTS OF PURE HMM

Attack type	Normal	Probe	DoS	R2L	U2R	Classification accuracy %
Normal	564	19	8	10	4	93.22
Probe	132	1359	7	0	0	89.80
DoS	128	100	2002	0	0	89.70
R2L	11	0	0	1125	2	80.90
U2R	91	2	0	11	124	54.40

Compared with pure HMM, HMM-DS system proposed in this paper can significantly improve the classification accuracy and speed. Other evidence fusion methods can improve the speed, but the classification accuracy is low. The comparison results of several common classification methods with the proposed approach are shown in Table 8.

TABLE VIII. COMPARISON BETWEEN THE PROPOSED APPROACH AND COMMON METHODS

Method	Normal	Probe	DoS	R2L	U2R
C4.5	97.08	87.62	96.08	8.12	23.69
LibSVM	91.83	85.26	97.30	18.29	25.88
NB	96.63	89.94	90.20	8.12	24.12
Our method	96.20	95.60	92.30	87.70	89.20

As shown in Table 8, all methods have high classification accuracy of Probe and DoS attack, but common methods are lower of R2L and U2R attack. For some business, and government networks, U2R and R2L attack have more damage than Probe and DoS attack. Thus, higher detection rate of U2R and R2L is equally important with the whole detection rate. Moreover, from the above discussion, it can be noticed the superiority of the proposed HMM-DS over other methods.

In conclusion, reasons for the above results are:

1) The network attack feature parameters space *S* is divided into several sub spaces which can reduce the dimension of the input vector for hidden Markov model. The training speed of each sub-model is accelerated, thus the classification speed of the HMM-DS method is improved.

2) The output results of each sub hidden Markov model are used as bodies of evidence. Some evidences are consistent

while some are conflicting. The evidence method proposed in this paper can effectively fuse these evidences.

3) The input of traditional HMM and other classification methods is ultra-high dimensional feature space. As some features interfere with each other, the speed and accuracy of classification is very low.

V. Conclusion

In this paper, the original feature parameters space of attacks were divided into several sub-feature spaces and a corresponding sub hidden Markov model for each sub-feature space was built. DS evidence theory method was applied to fuse the output of sub hidden Markov model, which can classify attacks effectively. The results show that this fusion system based on HMM-DS is obviously superior to the pure HMM or DS method, and combined the advantages both of HMM and DS. Hence, the proposed approach take advantage of HMM dealing with continuous dynamic signal, and calculate the match value between HMM model and unclassified data to form basic probability assignment which is provided for DS fusion decision. The advantage of DS can make up the shortage of HMM in making maximum probability judgment. the proposed approach proved to work well in combination of all kinds of evidence and to outperform other techniques in terms of classification accuracy. Experiments results show that the proposed approach can improve accuracy and speed of classification.

Although the proposed HMM-DS classification approach looks promising, there is still a large room to improve the classification accuracy for unknown attacks. In order to apply this scheme to other types of classification and recognition problems, a general framework for this approach needed to be constructed.

Acknowledgment

This work was supported by National Natural Science Foundation of China (No. 61202439), and the Fundamental Research Funds for the Central Universities of China.

References

[1] Panda, M., Abraham, A., Das, S., and Patra, M. R.,"Network intrusion detection system: A machine learning approach, " Intelligent Decision Technologies, vol. 5, no.4, pp. 347–356, 1955.

[2] BM Aslahi-Shahri, R Rahmani, M Chizari, A Maralani, M Eslami, MJ Golkar, A Ebrahimi, "A hybrid method consisting of GA and SVM for intrusion detection system" , Neural Computing & Applications, pp. 1-8, 1955.

[3] Cheng Xiang, Png Chin Yong, Lim Swee Meng, "Design of multiple-level hybrid classifier for intrusion detection system using Bayesian clustering and decision trees", Pattern Recognition Letters, vol. 29, no. 7, pp. 918–924 ,2008

[4] Levent Koc, Thomas A. Mazzuchi, Shahram Sarkani, "A network intrusion detection system based on a Hidden Naïve Bayes multiclass classifier", Expert Systems with Applications, vol. 39, no. 18, pp. 13492-13500 ,2012

[5] Yuk Ying Chunga, Noorhaniza Wahid, "A hybrid network intrusion detection system using simplified swarm optimization (SSO), Applied Soft Computing, vol. 12, no. 9, pp. 3014-3022 ,2012

[6] SJ Horng,MY Su,YH Chen,TW Kao,RJ Chen et al., "A novel intrusion detection system based on hierarchical clustering and support vector machines", Expert Systems with Applications, vol. 38, no. 1,pp. 306-313, 2011.

[7] Y Li, J Xia, S Zhang, J Yan, X Ai, "An efficient intrusion detection system based on support vector machines and gradually feature removal method", Expert Systems with Applications, vol. 39,no. 1,pp. 424-430,2012.

[8] D Ariu, R Tronci, G Giacinto, "HMMPayl: An intrusion detection system based on Hidden Markov Models", Computers & Security, vol. 30, no. 4, pp. 221‑241, 2011.

[9] S Jha, K Tan, RA Maxion, "Markov chains, classifiers, and intrusion detection", Computer Security Foundations Workshop, the 14th IEEE, Cape Breton, Novia Scotia, 2001, pp. 206-219.

[10] DEMPSTER A P, "Upper and low probabilities induced by a multi-valued mapping", Annuals of Mathematical Statistics, vol. 38, no. 6, pp. 325-339, 1967.

[11] SHAFER G A, "Mathematical theory of evidence," Princeton: Princeton University Press, 1976.

[12] FRADA B, CLYDE W H, " Handbook on decision support systems ," Heidelberg, Berlin,Springer, 2008.

[13] Chao Ning, Maoyin Chen, and Donghua Zhou, "Hidden Markov Model-Based Statistics Pattern Analysis for Multimode Process Monitoring: An Index-Switching Scheme", Industrial &Engineering Chemistry Research, vol. 53, no. 27, pp. 11084-11095,2014.

[14] W.Z. Wu, "Attribute reduction based on evidence theory in incomplete decision systems", Information Sciences, vol. 178, no. 2008, pp. 1355–1371.

[15] J Weisberg, "Dempster-Shafer Theory", International Journal of Approximate Reasoning, 2010.

[16] Yager R R, "On the Dempster-Shafer framework and new combination rule", Information Sciences, vol. 41, no. 2, pp.93-138, 1987.

[17] Yager RR, "On the fusion of non-independent belief structures", International Journal of General Systems, vol. 38, no. 5, pp. 505-531, 2009.

[18] SUN Quan, YE Xiuqing, GU Weikang, "A new combination rules of evidence theory", Acta Electronica Sinica, vol. 28, no. 8, pp.117-119, 2000.

[19] Thierry Denoeux. Conjunctive and disjunctive combination of belief functions induced by nondistinct bodies of evidence", Artificial Intelligence, vol. 172,no. 2–3 , pp. 234–264, 2008.

[20] Destercke S, Dubois D, " Idempotent conjunctive combination of belief functions: extending the minimum rule of possibility theory", Information Science, vol. 181, no. 18, pp. 3925-3945, 2011.

[21] Haenni R, "Are alternatives to Dempster's rule of combination real alternatives: comments on "about the belief function combination and the conflict management problem", Information Fusion, vol. 3, no. 4, pp. 237-239, 2002.

[22] Murphy C. Combining belief functions when evidence conflicts [J].Decision Support Systems, 2000, 29(1):1-9.

[23] Deqiang Han, Yong Deng, Chongzhao Han, "Sequential weighted combination for unreliable evidence based on evidence variance",Decision Support Systems, ,vol. 56, no. 6, pp. 387-393, 2013.

[24] Peida Xu, Yong Denga, Xiaoyan Sua, Sankaran Mahadevanc, "A new method to determine basic probability assignment from training data ",Knowledge-Based Systems, ,vol. 46, pp. 69-80, 2013.

[25] Xiaoyan Su, Sankaran Mahadevan, Peida Xu, Y Deng, "Handling of Dependence in Dempster‑Shafer Theory", International Journal of Intelligent Systems, ,vol. 30, no. 4, pp. 441‑467, 2015.

[26] HP Corporation, "Objective Attributes Weights Determining Based on Shannon Information Entropy in Hesitant Fuzzy Multiple Attribute Decision Making", Mathematical Problems in Engineering,vol. 2014, no. 1, pp. 1-7,2014.

[27] http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html

Cloud CRM: State-of-the-Art and Security Challenges

Amin Shaqrah
College of Business Administration
Taibah University
Medina, Saudi Arabia

***Abstract*—Security undoubtedly play the main role of cloud CRM deployment, since the agile firms utilized cloud services in the providers infrastructures to perform acute CRM operations. In this paper researcher emphasis on the cloud CRM themes, security threads the most concern. Some aspects of security discussed concern on deployment the cloud CRM like: Access customers' database and control; secure data transfer over the cloud; trust among the enterprise and cloud service provider; confidentiality, integrity, availability triad; and security hazard, future studies and practice are presented at the end.**

Keywords*—*Cloud computing; CRM; Security; Cloud Security

I. INTRODUCTION

Cloud computing begin changing IT industry and suitable technology at the current era according to its mass advantages and active usage of the resources. Cloud computing can be utilized under environments where the availability of resources is limited [1]. Cloud computing has become a research hot issue among contemporary technologies and researcher jostle to purplish at this field. There are several techniques used in the cloud technology before the implementation of the cloud technology on CRM system. For storing the customer database in the traditional enterprise internally, these data stored will be confidential and even it has some security measures and it is protected from the unauthorized user. But in the cloud computing situation the storage of customer database is someplace from the client workplace and the data storage and security measures will be in the service provider of the cloud computing environment [2].

Many firms like Google, Amazon, and Microsoft embraced cloud computing extensively in different areas. For example, Google or Dropbox have become everyday tools for millions of people. More, many enterprises presently used CRM based cloud and provided Cloud CRM as services such as Salesforce, Amazon, and Microsoft Azure [3]. In the technology enhanced CRM domain, the use of cloud-based system has also been identified as a crucial trend that permits accessibility to online services anyplace and undertakings scalability, enhanced availability and minimize cost to zero.

At the moment that cloud computing applied in the field of CRM; security play the main challenge deserve studying [4]. This paper showed the pre-requirement of cloud computing deployment, growths and CIA security themes which rising when deploying cloud CRM. Many security challenges have been discussed in the literature review. These challenges had better be consider before the deployment of cloud computing in CRM field [5]; [6].

II. CONCEPT OF CLOUD COMPUTING

Cloud computing services defined as a utilized the Internet as diffusion media and transferring information technology resources into services for end-users. The idea overdue cloud computing is to deliver computing as a utility in the same way that other public utilities such as gas and electricity are provided [1]. Cloud computing like's physical building, bringing home requirements to residents of the home [7]. As demonstrated, the significance of cloud computing's primarily being in tolerating the end user accessibility to resources through the Internet, as shown in Fig. 1. Some researchers find cloud computing parallel with grid computing [1], but some also find similarities to utilities such as water and electrical power and refer to it as utility computing [8]. Because the use of resources can be autonomously modified, it is also from time to time referred to as autonomic computing [9]. The literature review contains many explanations of cloud computing [10].

After gathering academic definitions of cloud computing[11] suggested that cloud computing could be defined as the combination of cybernetic resources according to user requirements, adaptably resources with IT architecture and infrastructure containing software facilities, computing platform facilities, expansion platform services, and rental the required infrastructure to create cloud services.

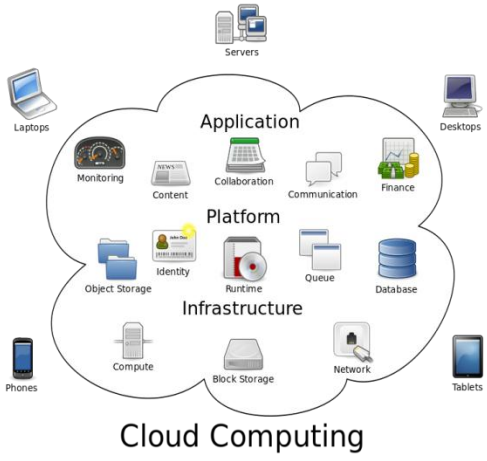


Fig. 1. Cloud Computing Metaphor (Adapted from Omni group Co.)

The distinct characteristics of cloud computing contain the capability of storing user information in the space and no necessity for application or software to install on the client side. By means of the user is able to link to the Internet, all of

the IT resources in the cloud can be used as client-side infrastructure. In the main, cloud computing applications are demand-driven, given various services regarding to user requirements, and service level agreement by service provider [12]. [13] Evaluated the importance of survival the enterprise and how move into a global market, trustworthy and well-organized infrastructure with cloud computing. [13] discussed the infrastructure as well as possible applications, cloud computing features can be figured-see figure 2- from the [13] comparison.

National Institute of Standards and Technology defined cloud computing as an on-demand access to a pooled of computing resources. All these computing resources - hardware, software, databases, networking, storage media, and so forth- are delivered rapidly to the clients [14]. No doubt the security is first priority to achieve an optimal allocation for immediate cloud services. Likewise, cloud base on demand access should be characterized by:

- Heterogeneous database system where apps. are stored in a cloud of distributed servers that can be reached through a Web browser.

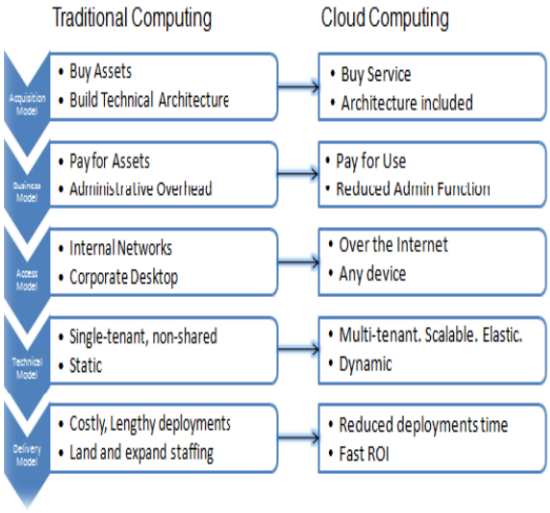


Fig. 2. From Traditional Computing to Cloud Computing (Adapted from [13])

- Robust technology infrastructures at the apps. and platform levels.
- On demand resources offered by the cloud allocated according to the need.
- Robust tolerance when one or several resources breakdown.
- Convincing business models where clients pay according to the resources used.

III. Cloud Deployment Models

The cost of cloud computing closed to zero maintenance since the service provider is responsible for the availability of services and clients are free from maintenance, unlike other computing models. By applying cloud computing technology, the enterprise doesn't need to pay attention or pay money for IT solutions, development, updating, and maintenance types from corrective, adaptive, preventive, and perfective [16]. Cloud service provider will change made to a system to fix or enhance its functionality [15] and [17] Suggested three commonly referenced cloud service models developed:

- Software-as-a-Service (SaaS): This type can use only when hosted app. are provisioned. By using this model enterprise can reduce the cost of IT infrastructure, also pre-and/or post-operations and all types of maintenance.
- Platform-as-a-Service (PaaS): In this model, the enterprise can customize his apps. on the provider platform. By using this model you can reduce the cost and raised management issues. The enterprise managed his required app. components of the platform. The development conditions are determined by the cloud provider according to the contract between parties. The cloud client has a full control over apps. and application environment settings of the platform.
- Infrastructure-as-a-Service (IaaS): The cloud provider hosts the clients' virtual machines and provides networks and storage. By using this model enterprise doesn't need to pay for procurement and managing the IT infrastructure components, all resources virtualized through a service interface.

In Cloud computing, resources can be either owned public cloud -like Google and Microsoft- or owned private cloud. Public cloud is contract with accessibility of external users whom pay-as you- use base. Whereas, the private cloud is built for accessibility within the enterprise where the users can utilize the facility without any charge. The third model is community cloud; community cloud shares infrastructure and architecture between enterprises from a particular community with mutual concerns, whether managed internally or by a third party and hosted externally. [18] Offered fourth model called hybrid cloud which encompasses two or more clouds "public, private, community" that residues a unique entities but is assured together, presenting the advantages of compound deployment models.

IV. Cloud Based CRM

Due to radically growth in the volume of information which must be under controlled, also the scale and scope of enterprise become grow rabidly. Effective CRM needs huge amount of investments in technology, whether business process reengineering, and training of users. Enterprise must utilize CRM technologies to analyze and synthesis enormous customer database information due to the interaction among them more often 24/7/365 days [19]. ICT allows customer database to be collected, consolidated, deployed, and analyzed on a unique scale. Web based CRM more than ICT. The customer must become the central point priorities of the enterprise, all management levels should recognize and support the shared values and vision required for successful CRM [20]; [21].

Enterprises agreement that web based CRM play a significant role of business development because the web based techniques permits exact analyzing, automated, and

classifying customers which is vital for focusing customer-centric space [22]. To compete in the digital era, focusing on the customer is becoming a key factor for enterprise. It is well-known that it takes up to five to eight times more money to obtain a new customer than to get current customer to make a new buying. Hence, customer retention and extension are important to transform enterprise into customer-centric space which can exploit the value of each customer [23] [24].

CRM is based on the capacity to assist interaction to offer steady, high quality, and cost-effective services to each customer [25]. [3] noticed CRM provides sales force more time to sell, increases customer response times and excellence of operation, and let's market share better. CRM attitude influences best option. Customer services on this field should be innovative, accessible, and confirmed by the cyber market needs. Hence, attempts to utilizing a new technology like cloud platforms, precisely on CRM value strategy whether any operational; analytical; or strategy side [24].

One main benefit of CRM packaged is delivered the customer database in the cloud. A cloud-based system is designed to be flexible with scalability so a business can scale up (or down) their CRM depending on their needs. Normally, the cost of CRM is often based on the number of users and storage, also as you requirements adjustment. In most cases the contacting and determine the service level agreement among cloud CRM vendor and enterprise are critical to your Cloud CRM deployment. Cloud CRM is often an optimal solution for enterprise that had a minimize experience in in-house IT deployment. With cloud CRM the vendor is responsible for managing and upgrading the software, so long as updates across the system and considering technical and non-technical problems, debugging and other issues which may be raised; one of the advantages of CRM base on the cloud is the combination process with universally standards and regulations. Furthermore, CRM integration with communities (Collaborative CRM) becomes prominent.

V. CLOUD CRM SECURITY CHALLENGES

Here researcher discussed some important cloud CRM security and challenges:

Access clients' database: The confidential client information transferring to provider of cloud computing database which mean a greater possibilities to be illegally accessed due to the accessibility controlling over the internet. As clients' information usually stored in the cloud for a lengthy years the threat of illegal accessibility is higher. So the cloud CRM provider should be secure database through coherent security policy and techniques [26].Customer database stored on the cloud must be saved private and the provider should not be able to conciliation the data privacy by any means.

Secure information transfer over the cloud: All data between the enterprise network and any service providers must pass Internet. Enterprise must sure that clients database is continuously moved on 'shttp' protocol and secure browser data also should be constantly encrypted and authentic [27].The data owner has full control over authorization of transferring data. This authority particular by the owner, the chosen user can at that point access the data kept on the cloud. Nonetheless, the process should not give the cloud provider any right to access the data.

Trust: Trust between enterprise and cloud service provider services supports the continuity relationship and commitment to encompass an inter-organizational relationship. Trust occurs when the enterprise certain degree of cloud service provider readiness and ability to deliver their responsibilities [28]. The importance of trust not only in the transaction stage –included detailed information about the services, ordering, purchasing, paying, and support the services), but also in the after transaction stage in the practice of warranties and money refunds. Hereafter, high levels of trust will likely result in high levels of security [28].

CIA TRIAD: Three requirements should establish to protect cloud CRM called CIA. Confidentiality, integrity, and availability, these pyramid are needed for authorized users. Confidentiality associated with privacy, that means the sender and its receiver should only share information, it is not able to secure data confidentiality while it flows in the system [29]. Such enterprise systems use long sequences of characters and complex algorithms to encode and decode information that exchanged among sender and receiver [27].

Integrity associated with the verification system against any kind of data loss, modification, and/or damage which caused intentional or un- intentional reasons, such as damaging actions of hackers or unauthorized person both internally or externally enterprise. Thus, the internet security system expected to assure that data received exactly at the right manner. Irrespective of the original cause of losing data integrity, this loss will definitely be a terrible for internet security system [30]. Last, authenticity is directly related to the techniques of security system which performs to establish data transfer when started and where end, thus trying to assure the data received was really originated where it says it is coming from and sent by the one mentioned on its label[31]. An Enterprise operation ruled by complex protocols which may unfortunately add some problems as far as system security is concerned [28].

Security hazard: There are two types of security mechanisms for conducting coherent cloud CRM. First, deploy physical security mechanism to minimize the hazard. The second type is the intangible protection security dealing with the system defense to enhance the capability of enterprise security, allowing conducting a successful business over network [30]. It's important to remember that awareness and responsibility of the underlying cloud CRM refer to the creating trust among enterprises rather the internet threats itself.

The mentioned of these security variables makes chances for effective deployment of cloud CRM. The suggested framework figure 3. increase understanding the main challenges of cloud CRM deployment. The research model illustrated the security variables which affects on cloud CRM deployment. The research model can be analyzed from not-technical perspective. For example, the security dimension of the below research model might be empirically tested to show the below independent variables are vital. In addition the above-mentioned, the research model reveals new academic

backgrounds and taking an early stage to entirely investigate cloud CRM deployment.

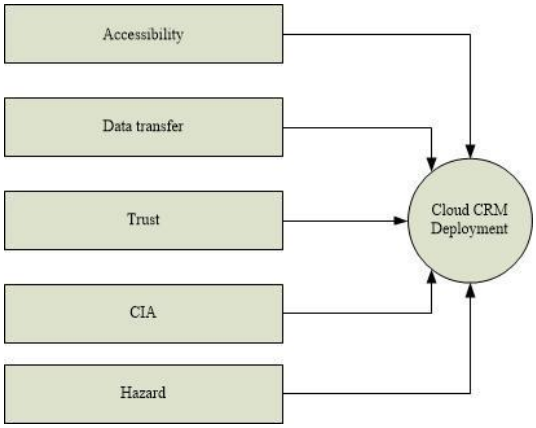


Fig. 3. Conceptual model of security dimension in cloud CRM

VI. Conclusion and Future Work

In this paper, researcher presented a broad discussion of cloud CRM challenges from security perspective. Researcher explained the cloud CRM and security challenges and worldwide web security. Cloud CRM has a fast step of development in the agile companies whether deploy public cloud, private cloud, hybrid cloud and community cloud. The cloud CRM is related to IT architecture and infrastructure and customized CRM area. This study explored that are five security variables could be affected on cloud CRM deployment; these variables were analyzed to perceive how the security dimension is critical for cloud CRM implementation. The results showed that the figured security dimensions have capabilities to fulfill of the pre- requirements of cloud CRM deployment, furthermore; the CIA issue becomes the most noticeable variable among the figured variables.

Nowadays, security is a core problem of cloud CRM, the challenges of keeping information and applying CIA pyramid become on the top priority; the most reason behind un-secure information is the architecture, infrastructure of the cloud CRM provider [32]. The cloud CRM security in Arab region needs to consider technical and strategically thinking, including but not limit to encryption scheme, resource provisioning, service level agreement, and accountability. Some organizations in Arab region don't believe utilize cloud CRM services for the reason that insecurity cloud world. This paper observed as a conceptual view paper; literature review supports our conceptual model of this study. A survey to validate and test the conceptual model needs initiative to add value of cloud CRM contexts.

References

[1] Sriram, I., and Khajeh, A. "Research Agenda in Cloud Technologies". 2010. (arXiv e-print No. 1001.3259). Retrieved from http://arxiv.org/abs/1001.3259.

[2] Weiss, A., "Computing in the clouds, networker", 2007, Vol. 11, No. 4, pp. 16-25.

[3] Marston, S., Li, Z., Bandyopadhyay, S., Zhang, J., & Ghalsasi, A., "Cloud computing -The business perspective". Decision Support Systems, 2011, Vol. 51, No. 1, pp. 176–189.

[4] Zhang, Q., Cheng, L., & Boutaba, R. "Cloud computing: state-of-the-art and research challenges". Journal of Internet Services and Applications, 2010, Vol. 1, No. 1.

[5] Chen, Y., Paxson, V., Katz, R., "What's new about cloud computing security?" Technical Report UCB/EECS-2010-5, Electrical Engineering and Computer Sciences, University of California at Berkeley, 2010.

[6] Christodorescu, M., Sailer, R., Schales, D., Sgandurra, D., Zamboni, D.,"Cloud security is not (just) virtualization security": a short paper. In: Proceedings of the 2009 ACM Workshop on Cloud Computing Security, pp. 97–102. London. Retrieved from http://link.springer.com/chapter/10.1007/978-0-85729-049-6_1, 2009.

[7] Cafaro, M., and Aloisio, G. Grids, "Clouds, and Virtualization". In M. Cafaro and G. Aloisio (Eds.), Grids, Clouds and Virtualization (pp. 1–21). 2011, Springer

[8] Chengyun, Z. "Cloud Security: The security risks of cloud computing, models and strategies", 2010. Programmer.

[9] Zhongze, Y. "The basic principles of cloud computing and its impact on education", Satellite TV and Broadband Multimedia, 2010.

[10] Hayes, B., "Cloud computing", Comm. ACM, 2008, Vol. 51, No. 7.

[11] Oredo, J., and J. Njihia, "Challenges of Cloud Computing in Business: Towards New Organizational Competencies", International Journal of Business and Social Science, 2014, Vol.5, No.3.

[12] Lin, A., and Chen, N., "Cloud computing as an innovation: Perception, attitude, and adoption". International Journal of Information Management, 2012, Vol. 32, No. 6.

[13] Delic, A., and Riley, J. "Enterprise Knowledge Clouds", Next Generation Km Syst. Int. Conf. Inform., Process, Knowledge Management, Cancun, Mexico, 2009,pp. 49–53.

[14] Mell, P., Grance, T., "NIST definition of cloud computing". 2010, National Institute of Standards and Technology.

[15] Pritesh, J., Dheeraj, R., Shyam, P., "A Survey and Analysis of Cloud Model-Based Security for Computing Secure Cloud Bursting and Aggregation in Renal Environment", 2011, IEEE

[16] Farhan, B., Haider, S., "The sixth International: Conference on Internet Technology and Secured Transactions", 2011, UAE.

[17] Wayne, A., Jansen, "Cloud Hooks: Security and Privacy Issues in Cloud Computing NIST", 2011, Proceedings of the 44th Hawaii International Conference on System Sciences.

[18] Buyya, R., Goscinski, A., and Broberg, J., "Introduction to Cloud Computing. In Cloud computing: principles and paradigms". 2011, Hoboken, N.J.: Wiley.

[19] Kim, W., "Cloud Computing: Today and Tomorrow". The Journal of Object Technology, 2009, Vol. 8 , No. 1.

[20] Rygielski,C.,Wang,J. C., and Yen,D. C., "Data mining techniques for customer relationship management". 2002, Technology in Society, Vol. 24, No. 1 .pp. 483–502.

[21] Piccoli,G., O'connor,P.,Capacciol i, C.,and Alvarez,R.,"Customer relationship management a driver for change in the structure of the US lodging industry". 2003,Cornell Hotel and Restaurant Administration Quarterly, Vol. 61, No. 1. pp. 61–73.

[22] Gurau, C.,Ranchh od,A.,and Hackney, R., "Customer-centric strategic planning: Integrating CRM in online business systems" 2013, Information Technology and Management, Vol. 4,No. (2–3), pp. 199–214.

[23] Baumeister,H., "Customer relationship management for SME's". 2002, Institut f˙ur Informatik,LMU,Oettinge nstr. 67,D-80538 Munchen, Germany,pp. 1–7.

[24] Skaates,M. and Seppanen, V., "Managing relationship-driven competence dynamics in professional service organizations". 2002, European Management Journal, Vol. 20, No. 4, pp. 430–437.

[25] Andrade,S.. "Using customer relationship management strategies". 2003, Applied Clinical Trials, Vol. 37, No. 1, pp. 37–41.

[26] Krautheim, F., "Building trusts into utility computing. Ph.D. dissertation", 2010, The University of Maryland, pp. 36-37.

[27] Wang, W., Rashid, A., and Chuang, H., "Toward the Trend of Cloud Computing", 2011, Journal of Electronic Commerce Research, Vol. 12, No.4.

[28] Shaqrah, A., "The Influence of Internet Security on E-Business Competence in Jordan: An Empirical Analysis", 2011, International Journal of business data communications and networking, Vol. 7, No. 4.

[29] Bishop, M., "Computer Security: Art and Science", 2002, Addison-Wesley.

[30] Ackermann, R. Schumacher, M. Roedig, U. and Steinmetz, R. "Vulnerabilities and Security Limitations of Current IP Telephony Systems" 2001, Proceedings of the Conference on Communications and Multimedia Security, PP.53–66.

[31] Jessup, L. and Valacich, J., "Information Systems Today: Managing in the Digital World" 2008, Pearson education, Inc. Upper Saddle River, NJ.

[32] Narzu, T., and Nova, A., "Efficient and reliable hybrid cloud architecture for big database" International Journal on Cloud Computing: Services and Architecture, Vol.3, No.6.

Improve Traffic Management in the Vehicular Ad Hoc Networks by Combining Ant Colony Algorithm and Fuzzy System

Fazlollah Khodadadi
Department of Engineering Software
Bushehr Branch, Islamic Azad
University, Bushehr, Iran

Seyed Javad Mirabedini
Department of Engineering Software,
Central Tehran Branch, Islamic Azad
University, Tehran, Iran

Ali Harounabadi
Department of Engineering Software
Central Tehran Branch, Islamic Azad
University, Tehran, Iran

***Abstract*—Over the last years, total number of transporter has increased. High traffic leads to serious problems and finding a sensible solution to solve the traffic problem is a significant challenge. Also, the use of the full capacity of existing streets can help to solve this problem and reduce costs. Instead of using static algorithms, we present a new method, ACO algorithm, combine with fuzzy logic which is a fair solution to improve traffic management in the vehicular ad hoc networks. We have called this the Improved Traffic Management in the Vehicular ad hoc networks (ITMV). Proffer method combines the map segmentation and assign to one server, calculate the instantaneous state of the traffic in roads with use fuzzy logic and distribute traffic for reduce traffic as much as possible by less time priority rather than shorter route. This method collects the vehicles and streets information to calculate the instantaneous state of the vehicle density. The proposed method through simulations were compared with some existing methods and in terms of speed, travel time and reduce air pollution improved by an average of 36.5%, 38%, 29% ,Respectively.**

***Keywords*—Traffic Management; Vehicular Ad-hoc Networks; Ant Colony Algorithms; Fuzzy System**

I. Introduction

Vehicular ad hoc network(VANET) consist of some Road Side Unit (RSU) and vehicles propagating messages. These networks aim to enable traffic and road condition information propagation to find independent mobile vehicles, increase driving safety, help drivers to find routes, accelerate routing, and finally, optimally predict and manage traffic.

A fuzzy system allows us to implement natural principles of life similar to what humans think. Predicting fuzzy logic rate is a technique that allows continues traffic monitoring. This logic can effectively identify vehicular congestion using the information collected from the environment and provide necessary information regarding traffic congestion features for road traffic management [10].

This article presenting a method based in ACO algorithm and fuzzy system for Improved Traffic Management in the VANETs (ITMV). In ITMV, traffic intensity is calculate using the fuzzy system. In addition to being effective in instantaneous optimal route selection, it is also effective in calculating the probability of selecting an optimal route through the corresponding procedure of the ant colony algorithm. Finally, it causes to use optimal routes more frequently and guide vehicles with high speed and short travel time to their destination. Traffic intensity is calculate using the fuzzy system.

In the following, an overview of the ACO algorithm, fuzzy system's and discusses the traffic management systems are presented in Section 2. ITMV and its operation are presented in Section 3. Section 4 includes the simulation and the results of the simulation. Finally, Section 5 concludes the paper.

II. Background and Previous Works

A. Vehicular Ad hoc Network (VANET)

In VANET (Figure 1), each vehicle is equipped with a technology that allows drivers to communicate with each other and with the road infrastructure. The road infrastructure, which is known as road side units, is placed at vital points of streets and roads, e.g. red lights at intersections or stop signs, to improve the traffic status and make driving safer [3].

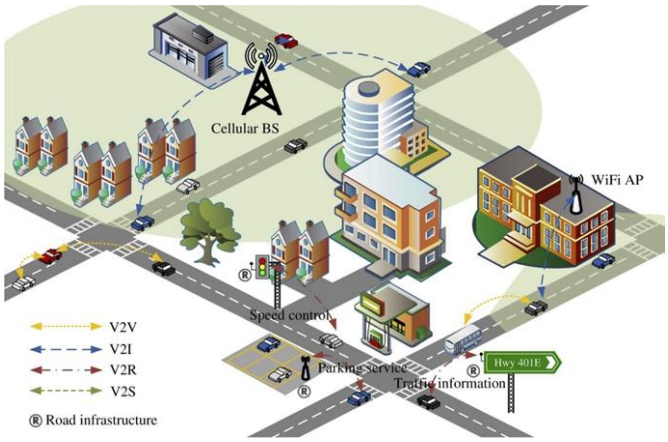


Fig. 1. A View of a vehicular ad-hoc network

Some of the defined applications of this technology include reducing traffic problems, preventing road accidents, improving the management of necessary operations after road accidents, license and registration checks and other statistical and entertainment applications.

B. Ant Colony Based Algorithm

The first algorithm to find an optimal path in a graph based on the behavior of ants in searching food from their colony was

introduced and developed by Dr. Marco Dorigo in 1992. Most features of actual ants are imitated by artificial ones to simulate the social behaviors of ants to solve optimization and distributed control problems.

This algorithm has changed in recent years and has been used in various fields. Some applications and fields where ant colony algorithm used include routing, image processing, scheduling and etc.

C. Fuzzy Systems

Fuzzy logic was first introduced after the development of the fuzzy set theory by pprofessor lotfizadeh in 1965. The tools and applications created by the fuzzy set theory can support all stages of a pattern analysis or knowledge discovery process. Most commercial and economic applications of fuzzy logic is related to process control [4].

The starting point of a fuzzy system is obtaining a set of if-then fuzzy rules from the experts` knowledge or the knowledge of the considered field. The next stage is to combine these rules in an integrated system [10].

D. Previous Works

Reference [11] propose an Alleviating Traffic Congestion (ATC) method in which each vehicle uses a VANET to collect traffic information of a limited region depending on its traffic conditions. Consequently, this provides higher speed and shorter travel time for vehicles` information and communication system for temporary and transient traffic in urban transportation.

Reference [2] propose an Urban Traffic Control Aware Routing Protocol (UTCARP), which includes two modules, i.e. node selection by sending a package to the destination and greedy exchange strategy for sending a package between two adjacent nodes. They asserted that their proposed method has better performance in delivering packages, end-to-end delay, and routing overhead.

Reference [6] present a method in which each vehicle independently collects local information of the congested region and distributes traffic from crowded to non-crowded areas. In locations, where traffic has temporary and distant changes, this method provides higher speed and shorter travel time in comparison to that of current systems.

Reference [9] have also proposed an optimal routing approach using a machine learning algorithm to reduce vehicles` travel time by combining the ant colony algorithm and path length based learning methods.

Reference [12] introduce a novel fleet system in which a unique strategy provides a rapid query response at each intersegment by inter-vehicular communication through forming a local label to predict a non-crowded route. This system enables vehicles to find a route, which can be navigated with a relatively high speed.

Reference [13] includes two models for Route Guidance Systems (RGS), one based on propagating traffic flow in the network and another based on tow flow capacity at different times on related road links. They claim that both models reduce prediction error to 52% and travel time average to 70% in comparison to other methods.

Jabbarpour et al. [8] have also published a paper in which congestion avoidance is improved by traffic prediction. Their method combines predicting the average speed of traffic movement in roads with segmented maps and finding the shortest path with minimum congestion using the ant colony algorithm. Consequently, average travel time is reduced by 11.5% and average travel speed is increased by 13%.

III. The Proposed Method

Proper traffic management means predicting traffic occurrence and providing solutions to prevent heavy traffic formation in the streets and highways. On this method, street maps were divided into several parts and a fuzzy-ant based

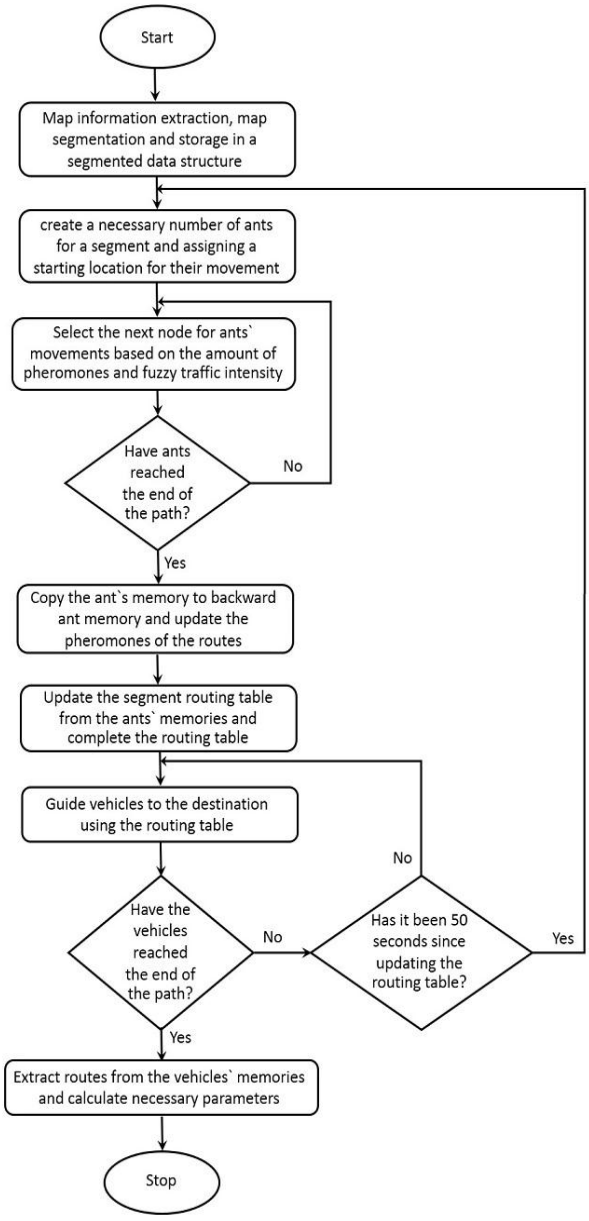


Fig. 2. The overall performance of the proposed method

algorithm was used to find best route and complete the routing table for every part of map. In addition, a method based on fuzzy logic is used to predict traffic patterns, which increases the probability of selecting optimal and non-crowded routes by ants through computing fuzzy traffic intensity. The overall performance of the proposed method is presented in figure 2.

A. Problem environment depiction:

In order to traffic management, problem environment depiction includes four steps, which are explained in the following:

1) Preparation of the real map: First step shows the roads map, streets and intersections of real map. Firstly, a part of the city of Bushehr, Iran map used in simulation which is shown in Figure 3.

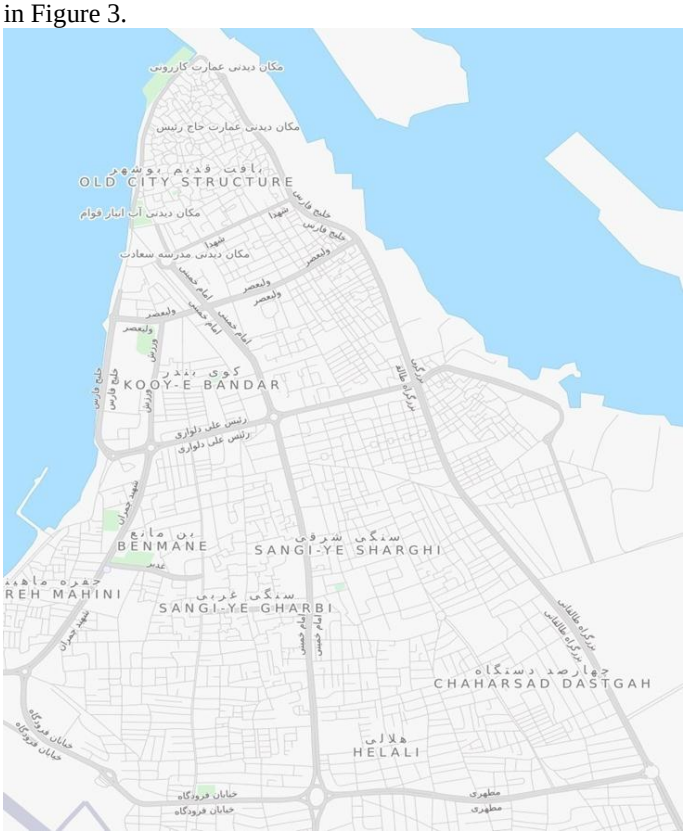


Fig. 3. Real map exported from OpenStreetMap database

2) Map Segmentation: The segmentation happens in step 2 and real map is divided into several parts with almost identical sizes which this segmentation can be based on the number of streets or intersections in each part.

With map segmentation we can manage the dynamic and quick changes of vehicular environments and routing is accomplished for each segment individually instead of the whole map [8].

3) Draw the problem graph: In this step, the real map is converted to a graph and this graph is given by $G_M = (N_M, L_M)$, where N_M and L_M are the set of nodes and links, respectively.

4) Draw the segment routing table: The routing table for each segment formed and updated by the same server, which is called the Segment Routing Table (SRT(i)), where i is the segment number (or identifier). $G_S=(N_S,L_S)$, where N_S and L_S is the set of nodes and links which assigned to each segment. The SRT are used by same segment server.

The ant colony algorithm is used separately for each segment, updates all edges and segments, and performs routing and traffic control separately by servers dedicated to each segment. The routing table of different segments are updated using a fuzzy-ant algorithm, which is talk about later.

B. Initialization

We use the different types of ants in our system, that's mean forward ant and backward ant, used in this subsection. The navigation servers is first provided with the number of vehicles simultaneously passing a road (the number of lines) and the link length.

Max_NV_{ij} is the maximum number of vehicles which can be on the road and computed using:

$$Max\_NV_{ij} = \frac{LL_{ij}}{L_V + \Delta L} \times NL_{ij} \quad (1)$$

Where LL_{ij} is the length and NL_{ij} is the number of lanes of street between node i and j. ΔL is the average space between two vehicles and Finally, L_V is the average length of vehicles.(L_V and ΔL are considered as 5 m and 3 m in this paper) In addition, vehicle density (D_{ij}) can calculate with using Equation (2):

$$D_{ij} = \frac{NV_{ij}}{Max\_NV_{ij}} \quad (2)$$

After this level, forward ants are used to find the shortest low traffic street between source and destination.

1) Forward ant: Forward ants can travel around the each segments to find the shortest and optimal between source and destination. This ants using a new probability function to choice the next node to move forward. Equation (3) shows the new probability function which can used by forward ants:

$$p_{ij}^{k}(t) = \begin{cases} \dfrac{\alpha(\tau_{ij}) + \beta(1-\eta_{ij})}{\sum_{h \notin tabu_k} \alpha(\tau_{ij}) + \beta(1-\eta_{ij})} \times \left(\dfrac{1}{1+\frac{1}{N_j}}\right) & if\ j \notin tabu_k, \\ 0 & otherwise, \end{cases} \quad (3)$$

Where τ_{ij} is the pheromone value an ant in node i to move to node j and is calculated by backward ants using Equation (7). η_{ij} is the instantaneous state of the fuzzy value on the link from i to j and calculated by vehicle as ant. α and β weight the importance of τ_{ij} and η_{ij}, and are called pheromone and fuzzy data power, respectively. $tabu_k$ also are the set of nodes connected to node i that an ant k has not visited yet. N_j is the number of neighbors for node j. The calculation methods of τ_{ij} and η_{ij} are discussed in the next sections.

2) Backward Ant: When a forward ant reaches its destination, it changes its tasks and becomes a backward ant, instead of destroy and copying its memory to backward ant. In this case, the overall system is reduced. The backward ant

returns to the source by using forward ant memory and updates the links pheromone intensity using the pheromone update equation, which is examined in the following sections.

C. Calculate fuzzy traffic intensity

Fuzzy instantaneous congestion state is a technique that constantly monitors traffic conditions. This effectively detects the traffic density of the road using vehicles` communications with the infrastructure and provides valuable information to manage urban traffic under the detected congestion conditions [10]. The fuzzy detection mechanism is based on two input parameters, which reflect the congestion level: the number of lines (NL_{ij}) and current traffic on the same link. Moreover, its output parameter provides the current instantaneous congestion state on the corresponding link (D_{ij}).

The fuzzy system is represented by linguistic variables, Lines and Traffic, for inputs and Intensity for output. Fuzzy input variables are mostly divided into different fuzzy sets. Fuzzy linguistic values or fuzzy sets are as follows:

- Lines = {L, M, H, VH}
- Traffic = {VL, L, M, H, VH}
- Intensity = {VL, ML, HL, LM, HM, LH, MH, VH}

Where, L=Low, VL=Very Low, ML=Middle Low, HL=High Low, LM=Low Medium, M=Medium, HM=High Medium, H=High, LH=Low High, MH=Middle High and VH=Very High. Figures 4 to 6 present fuzzy sets corresponding to fuzzy input and output variables.

Since each fuzzy set consist of elements with a degree of membership, the value of a fuzzy input may simultaneously belong to different fuzzy sets.

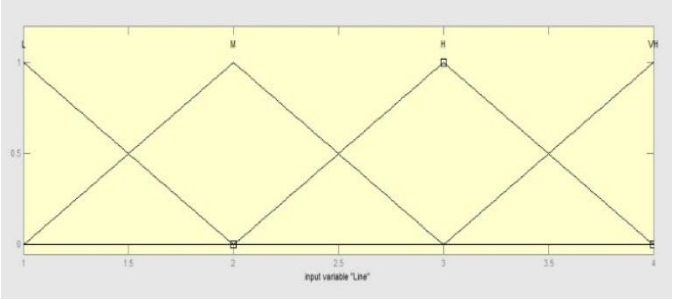


Fig. 4. Fuzzy set of Lines. (the number of lines in the road)

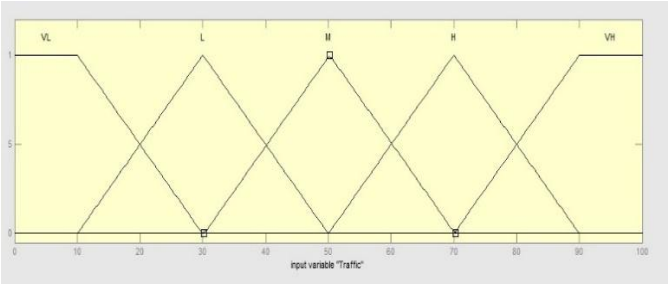


Fig. 5. The fuzzy set of Traffic. (the existing traffic on the road)

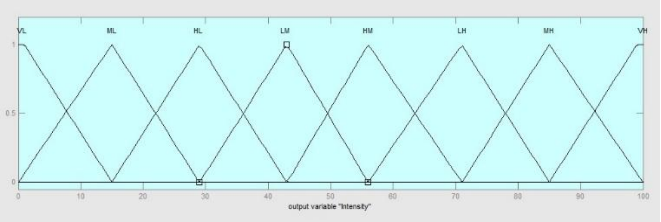


Fig. 6. The fuzzy set of Intensity. (instantaneous congestion state or traffic intensity)

As we can see in table 1, fuzzy engines consist of 20 rules for two fuzzy inputs and one fuzzy output. The main application of this system is that traffic management can adjust the rules according to current requirements. Instantaneous congestion state or fuzzy traffic intensity allows navigation systems to constantly monitor the instantaneous fuzzy traffic input parameters. This value affects the probability of route selection by ants and thus, vehicles. When instantaneous traffic intensity increases, the probability of selecting the path is reduced in proportion to the congestion.

TABLE I. DEFINED RULES TO PRESENT INSTANTANEOUS TRAFFIC INTENSITY

Lines	**Traffic**				
	VH	***H***	***M***	***L***	***VL***
L	VH	H	AM	M	M
M	H	AM	M	M	BM
H	AM	M	M	BM	L
VH	M	M	BM	L	VL

Figure 7 presents fuzzy outputs according to the values of fuzzy input and output sets and the table above. As we can see, traffic intensity is increased when traffic is increased or the number of lines is reduced.

In calculating instantaneous congestion, μ_{Lines} indicates the fuzzy number of lines, $\mu_{Traffic}$ shows the fuzzy value of traffic and $\mu_{Intensity}$ is the instantaneous traffic intensity, whose values are mapped to a value between 0 and 1.

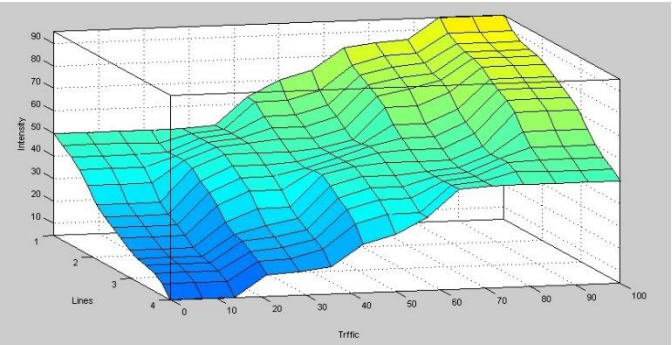


Fig. 7. Fuzzy output according to the set of defined rules

The inference engine used in this research is Mamdani based, which is applied as follows:

$$\mu_{Lines \cap Traffic} = \min\left(\mu_{Lines}, \mu_{Traffic}\right) \quad (4)$$

The representative of each fuzzy output value μIntensity is defined as follows according to its chart:

$$\bar{y}_i = \{0.01, 0.15, 0.3, 0.5, 0.7, 0.85, 1\} \quad (5)$$

Now, the center average defuzzifier is used to calculated traffic intensity. The required equation to calculate fuzzy traffic intensity using center average defuzzifier is as follows:

$$\eta_{ij} = y_i^* = \frac{\sum_{i=1}^{20} \bar{y}_i \times \mu_{Lines \cap Traffic}}{\sum_{i=1}^{20} \bar{y}_i} \quad (6)$$

D. Pheromone update

When the backward ants arrived, the pheromone value of links is updated. The amount of pheromone value is increased or decreased with Equation (7), as follows:

$$\tau_{ij}^{new} = (1-\rho)\tau_{ij}^{old} + \sum_{k=1}^{m} \Delta\tau_{ij}^{k}, \quad (7)$$

Where ρ∈A (0, 1] is a pheromone evaporation value, and n is the number of nodes in the same segment. The value of pheromone placed on links i and j by ant k is calculated using:

$$\Delta\tau_{ij}^{k} = \begin{cases} \frac{1}{LL_{ij}^{k}} + \frac{1}{\mathrm{TT}_{ij}^{k}} + \frac{1}{\mathrm{D}_{ij}^{k}} & if\ the\ kth\ ant\ pass\ link\ i-j, \\ o & otherwise, \end{cases} \quad (8)$$

Where TT_{ij}^{k}, D_{ij}^{k} and LL_{ij}^{k} are the travel time, vehicle density and length of each link which traverse by ant k, respectively.

E. Stopping Procedure

Ant colony algorithm should be terminated in a suitable situation. The number of repetitions, run time, a certain number of visited nodes and etc. are some examples of the algorithm termination condition. However, the terminated condition of ITMV is scrolling all the links to each segment by ants.

F. Update Routing Table

In this method, at specified intervals, γ, a certain number of forward ants to be re-created, N_a, in order to update routing table and randomly placed on different nodes per segments as start points. Then, they start to explore links using Equation (3) and trying to exploit all possible routes in segments. After that, backward ants return to source points and update the pheromone value of links by using Equation (7). Segment Routing Table or SRT is updated using routes have been found by ants. SRT(i) includes m smaller tables where m is the number of nodes in segment i. In each of these smaller tables there are m×n rows as a possible route, where n is the number of segments. Moreover, there are 3 columns in each table: destination node, next node and price route.

After forming the segment routing table, using border nodes of each segment and connecting them to its neighbors in the adjacent segment, the information of the routing table is extended and the vehicles' destinations as an ant are expanded beyond a segment.

G. Guiding Vehicles to Their Destination

The server's that is assigned to each segment, using the cable for communicate with each other and using segment routing table, through RSU navigate vehicles to your destinations.

Vehicles perform routing until the recommended path is the most optimal route to the destination; otherwise, another is recommended. The Cross Multiplication in equation 9 is used to determine the optimality of the recommended path:

$$p_n = \begin{cases} 1 & if\ \eta_{ij}^{n} < \omega \\ 0 & otherwise \end{cases} \quad (9)$$

Where, P_n is the probability of selecting a possible path from i to j by a vehicle, which should be guided through an optimal path (less distance and less traffic). Moreover, ω is the maximum instantaneous congestion, which can exist on a single link and has a value in range 0 and 1.

The closer this value is to the actual value, in addition to preventing congestion, vehicles navigate to shorter distance in less time with higher speed from their origin to destination and maximum capacity of paths is also optimal use.

IV. Simulation and Evaluation

This segment discusses simulation results and compares the proposed method with other existing systems. NS2 and SUMO simulators are two software applications, which are used for the simulation and performance presentation of the proposed method. First, the ant colony algorithm in combination with fuzzy logic which named Fuzzy and Ant Colony Optimization (FACO) is implemented in NS2 and their output, which is the scenario of vehicles` movements, as well as road maps, are inserted into SUMO after applying necessary changes. Finally, the simulation output, which consist of different information and statistics regarding vehicles, trips, simulation times, etc. are received through separate files and employed to compare and evaluate the proposed method.

We must note that simulation results also provide the maximum tolerable fuzzy instantaneous congestion on a single link.

A. Simulation Setup

NS2 is a simulator based on object-oriented and discrete event simulation, which supports different cable and wireless computer network simulations. This simulator is based on C++ language and OTCL interpreter to run user instructions. Using this tool, users can define different network topologies and protocols [7].

SUMO is an open-source, portable, and microscopic software application with a traffic simulation package to deal with very large networks. This software package includes several useful tools to help simulating urban networks. These tools include a graphical user interface for simulation, network converter and constructor, network graphic editor, routing scenario builder with different approaches, and several other useful tools [1].

In order to manage the traffic of vehicular networks, we exploit these two software packages to simulate the considered

vehicular ad hoc network in which routing is performed based on the proposed method that is a combination of the ant colony algorithm and fuzzy systems.

1) Map Preparation:

Netconvert tool embedded in SUMO software package are used to convert the primary map into a usable format in NS-2.35 and SUMO 0.12.3. The output of netconvert is an xml file, which is used as a road map in NS2. The FACO algorithm is implemented in NS2, using this map and its output, i.e. the scenario of vehicle routing, are transferred to SUMO simulator. Figure 8 presents the overall simulation process.

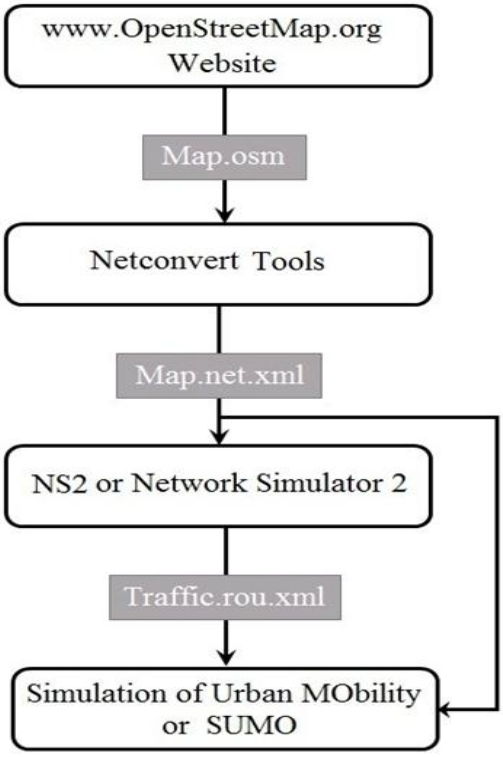


Fig. 8. The simulation process of the proposed method

The map file extracted from netconvert tools contains complete information of the used map, including street location and length, number of lines, maximum speed of each line or street, location and state of traffic signs (e.g. red lights), etc.

This map, according to the figure 3, as a real road map used in our simulation is shown in Figure 9 and the Table 2 represents the different statistic specifications for this road map.

TABLE II. STATISTIC SPECIFICATIONS OF ROAD MAP

Specification	Value
Size	3.5 km × 5 km
Map area	18 km^2
Streets/km^2	460
Junctions/km^2	219
Avg. street length	210.75 m
Avg. lanes/street	2.03

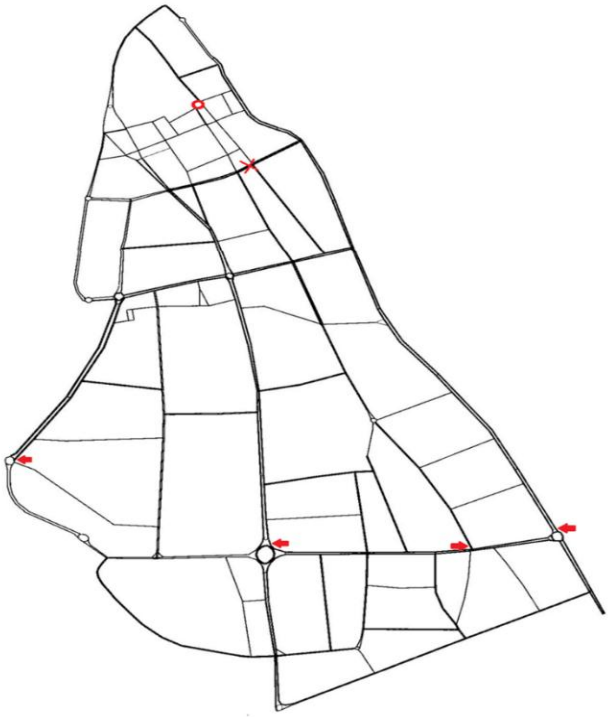


Fig. 9. Real road map of selected part of Bushehr city

2) FACO Algorithm Implementation:

In order to implement FACO algorithm using NS2, different data structures and functions are used to define and integrate constant values of the algorithm, specifications of the segmented road map, segment routing table, ants` information, fuzzy values of real-time data, etc. subsequently, by creating a number of predefined ants, the path mining and finding the most optimal route was initiated by ants.

After that, the segment routing table were updated. These two operations were repeated for specific intervals. Using these routing tables and a fuzzy system monitoring the road traffic, vehicles navigate their route to the destination. Finally, the ID, number of navigated streets, total navigated distance, and sum of stop times are stored at the memory of each vehicle.

To find the most appropriate value for the parameters of FACO algorithm, a simple map was used, which contains 16 two-way street with 12 intersection in 2 segment and is shown in Figure 10. In this figure, arrows indicate the starting point and the solid red circle indicate the end point of vehicle's movement. The average travel time was used as a measurement criterion in this subsection.

a) Pheromone power (α): This parameter indicate the possibility of a link to being selected based on the amount of pheromone by forward ants. Decreasing the value of α will reduce the power of the forward ants in search of new paths.

b) Fuzzy data power (β): This parameter specifies the effect of fuzzy data in path selection by forward ants.

There must be a appropriate balance between α and β (i.e., $\alpha + \beta = 1$). The best state happen when $\alpha=0.3$ and $\beta=0.7$ in this simulation. Figure 11 illustrates the average travel time of the found route by FACO algorithm.

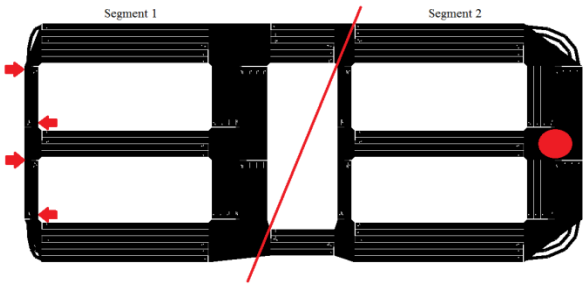


Fig. 10. Map used for finding FACO algorithm parameters' values

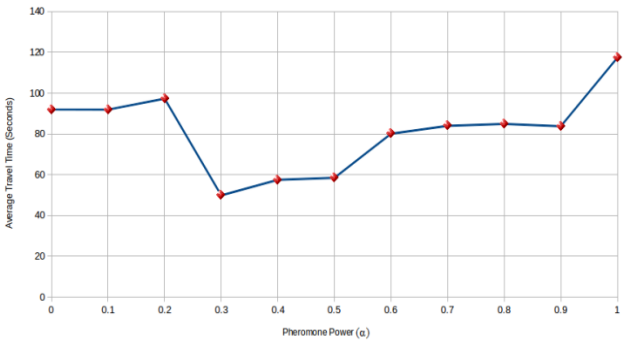


Fig. 11. Average travel time for FACO algorithm with different amounts of α. (ρ=0.5, N_a=12, γ=50 s)

c) Pheromone evaporation rate (ρ): This parameter very important when there are various path for choose and when the environment change quickly like vehicles environment. Since ρ can have direct impact on finding new path, different values are used through the our simulation and its result is exhibit in Figure 12. The lowest average travel time occur when ρ=0.6.

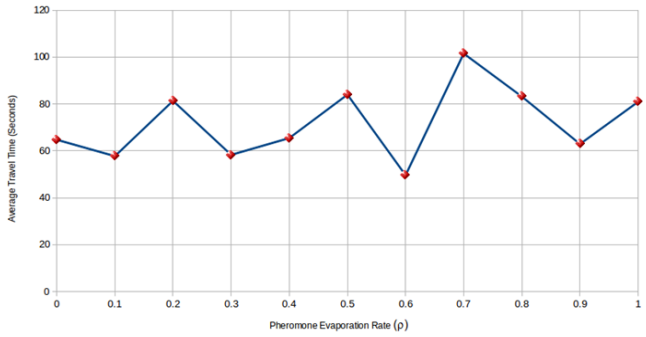


Fig. 12. Average travel time for FACO algorithm with different amounts of ρ from 0 to 1. (α=0.3, β=0.7, Na=12, γ=50 s)

d) Maximum fuzzy instantaneous congestion (ω): This parameter is used when maximum capacity of a street is occupied. FACO algorithm investigates this parameter to prevent excessive and intolerable traffic on the streets. It recommends less crowded and closer routes to vehicles and thus, manages the traffic of urban streets.

Through simulation, different values are investigated in range 0.5 to 1 with a 0.05 step to find the best value for this parameter. Figure 13 presents the results.

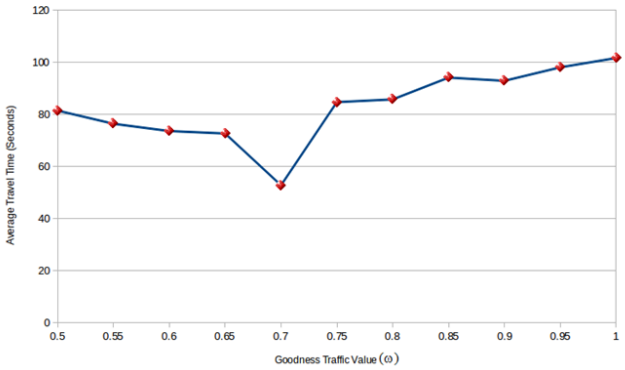


Fig. 13. Average travel time for FACO algorithm with different amounts of ω. (α=0.3, β=0.7, ρ=0.6, N_a=12, γ=50 s)

The best value for maximum tolerable instantaneous traffic on the streets is considered ω=0.725.

e) Number of ants (N_a) and re-create period (γ): In FACO algorithm, at regular intervals, the number of ants re-created and used for finding new path . In fact, set a less value for γ and a furter value for Na will cause the algorithm to perform better but increase the cost of the algorithm.

According the above fact, 50 second is set to the re-create period of ants in FACO and considering the number of streets in each segment, the number of ants was chosen 12 in our scenario. The configuration parameters of FACO algorithm in NS2 are listed in Table 3.

TABLE III. CONFIGURATION PARAMETERS OF FACO ALGORITHM IN NS2

Parameter	α	β	ρ	γ	ω
Proper value	0.3	0.7	0.6	50 s	0.725

3) Simulation:

In our simulation, create the number of vehicles from 100 to 1000 and put on 4 start points which is show in Figure 9 with red arrows. Also, the vehicle's destination specified with red circle at top of this figure. We have use the default value for some of the characteristic of SUMO but a lot of them changed with new value is shown in table 4.

TABLE IV. DEFAULT AND NEW VALUES FOR A NUMBER OF FEATURES OF SUMO

Characteristic	**Default value**	**New value**
Acceleration	2.6 km/h	0.8 km/h
Deceleration	4.5 km/h	-
Sigma (driver imperfection)	0.5	-
Minimum gap	2.5 m	3 m
Vehicle Length	5 m	-
Max speed	70 km/h	50 km/h

B. Comparing proposed approach

After finding the appropriate values for the various parameters of the FACO algorithm, the efficiency of improved method was appraised by comparing with Dijkstra and Pure Ant Colony Optimization (PACO) method. The Dijkstra algorithm was selected because it is a simple and widely used algorithms. Because the Ant Colony Optimization was used in the ITMV system, for this reason, the pure ant colony optimization was selected to compare this system with ITMV.

These three system were compared, based on the average travel time, waiting time, speed, distance, and reaction for accident, air pollution and fuel consumption by applying 100 to 1,000 vehicles.

1) Average travel time:

This parameter was computed and the results are shown in Figure 14 for Dijkstra, PACO and ITMV. As you can see, with the increasing number of vehicles, travel time for Dijkstra and PACO algorithm dramatically increased. In simulation, ITMV system improved the average travel time significantly. ITMV had the best average travel time in different vehicle densities and it improved travel time up to 36% and 40% than Dijkstra and PACO, respectively. Street traffic problem using the proposed method is significantly reduced due to adoption of a fuzzy system for predicting traffic.

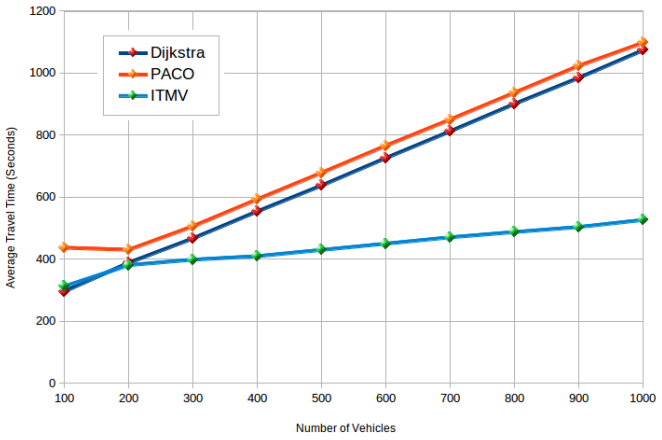


Fig. 14. Average travel time for Dijkstra, PACO and ITMV with the number of vehicles from 100 to 1,000

2) Average travel speed:

Average travel speed of Dijkstra, PACO and ITMV with different number of vehicles is shown in Figure 15. ITMV acquired the highest average than Dijkstra and PACO by traffic management before congestion happen. By increasing the number of vehicles, the average speed decreased smoothly from 41 to 27 km/h in ITMV and it increased travel speed up to 34% and 39% compared with Dijkstra and PACO, respectively.

The average travel speed for the lower number of vehicles, is almost at the same level but with the increasing number of vehicles, the average travel speed at Dijkstra and PACO compared to ITMV are greatly reduced.

3) Average travel distance:

This parameter was computed and the results are shown in Figure 16 for Dijkstra, PACO and ITMV. The greatest distance traveled by the proposed method because the ITMV approach seeks to achieve higher speed and lower travel time through a little farther routes with less traffic instead of the shortest routes with high traffic. It is worth noting that the average travel distance had increased at most 10.43% compared with Dijkstra. Since ITMV uses fuzzy system for predicating traffic, its average travel distance is higher than Dijkstra and PACO which haven't program for traffic management.

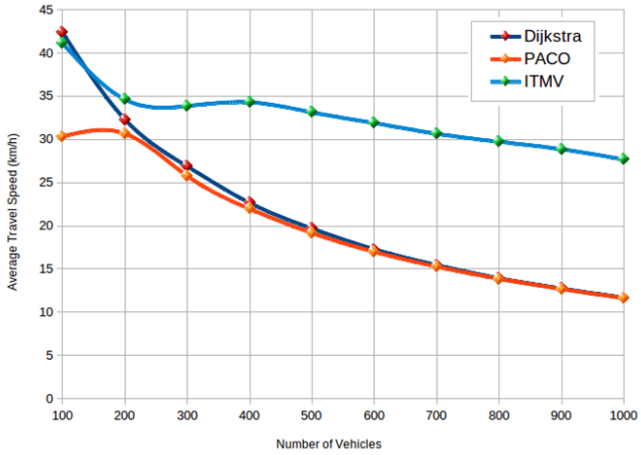


Fig. 15. Average travel speed for Dijkstra, PACO and ITMV with the number of vehicles from 100 to 1,000

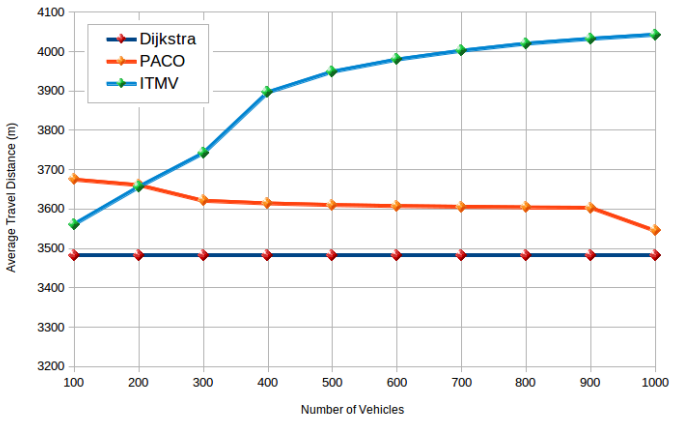


Fig. 16. Average travel distance for Dijkstra, PACO and ITMV with the number of vehicles from 100 to 1,000

4) Average waiting time:

The average waiting time of vehicles depend on the time they spend in traffic, which may be ensued by accidents, crowdedness, etc. This waiting time is computed for different routing algorithms, Dijkstra, PACO and ITMV management by SUMO. Figure 17 presents the results.

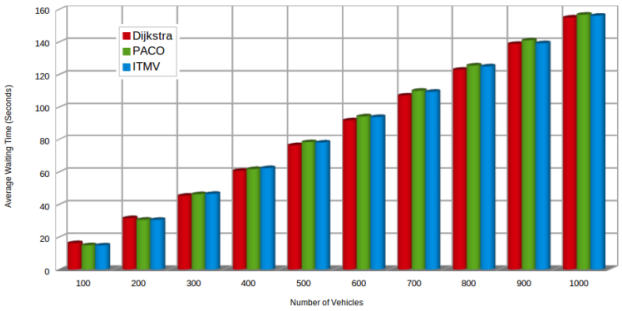


Fig. 17. Average waiting time for Dijkstra, PACO and ITMV with the number of vehicles from 100 to 1,000

As we can see, the waiting time of vehicles is almost the same for all three algorithms. Increasing the number of vehicles also increases their average waiting time. The proposed method only enhances travel time and vehicles` speed in navigating the distance.

5) Reaction to the incident

To evaluate the response of the proposed method in the incident as compared to other two methods, an additional simulation was performed for 1000 s. The incident was happened at one of the street, which is depicted via a cross sign in Figure 9, after 300 s and has lasted to the 700th second of simulation. In this case, vehicles are forced to stop until the end of time resolved the incident by using the stop element in SUMO. Average travel times were calculated for all three method in every 100 seconds of simulation and the results are displayed in Figure 18.

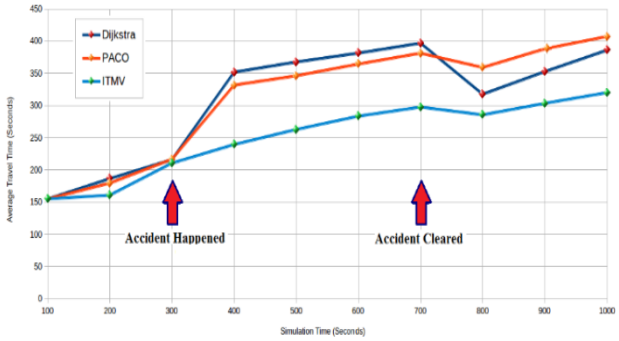


Fig. 18. Average travel time for Dijkstra, PACO and ITMV as a routing system as a function of simulation time

As you can see, before the incident happened, the average travel time for all methods is almost the same. From 300 to 700 seconds in the simulation, travel time is increased for all method due to traffic. However, in the proposed method, travel time takes its natural course due to traffic management.

At the 700th second when the incident was cleared from the street, the average travel time decreased for all of the methods and all the graphs smooth out to reach their initial values. ITMV had the best reaction for incident since it uses travel time and vehicle density and travel speed prediction for vehicle routing and uses fuzzy system from the beginning before incident happens. The improved traffic management approach shows 19.4% and 19.49% improvement on average in comparison to Dijkstra and PACO algorithm.

6) air pollution and fuel consumption

As you can see in figure 19, relationship between the vehicles speed and CO_2 emissions was measured by Barth and Boriboonsomsin [5] and they find that there is a U-shape relationship between these metrics. This means that at very low or very high average travel speed, the fuel consumption as well as CO_2 emissions increased by an average of 30%.

According figure 19 (the relationship between average speed and CO_2 gas propagation), results of average vehicle speed and travel distance, air pollution is computed for Dijkstra, PACO and ITMV and presented in figure 20.

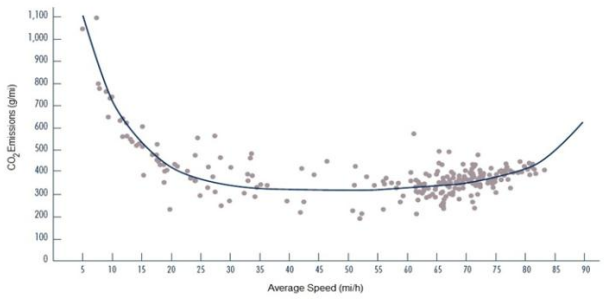


Fig. 19. Greenhouse gas propagation results according to average vehicle speed according to barth and boriboonsomsin research

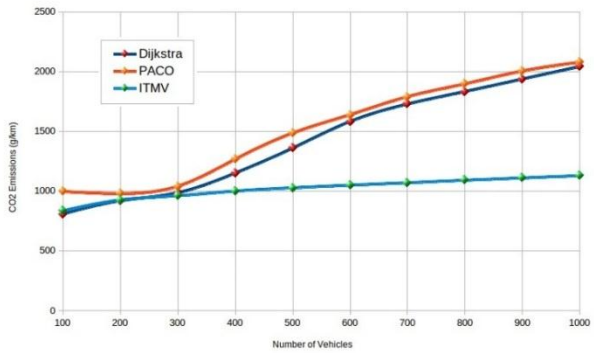


Fig. 20. CO_2 emissions for Dijkstra, PACO and ITMV with the number of vehicles from 100 to 1,000

As we can see, although the proposed method results in a larger average distance in comparison to other methods, it outperforms them regarding greenhouse gas production and fuel consumption, particularly when the number of vehicles is increased. This reduction in pollution and fuel consumption is ensued by high travel speed during navigation. The improved traffic management approach has effective in reducing air pollution and fuel consumption for 28.93% and 32.81% on average in comparison to Dijkstra and pure ant colony algorithm.

V. CONCLUSIONS

Traffic management will be discussed in this article, as one of the serious problems in the management of cities and mega cities. An ant-based algorithm was combined with fuzzy system and the map segmentation in order to derive an improved traffic management. Segmentation and fuzzy system were used to bring down complexity of computing. Applying an ant-based algorithm to our system required some modifications to the original ACO algorithm. These modifications include map segmentation, new probability function, new reinforcement and evaporation rules.

ITMV efficiency was computed by comparing it with other algorithms such as Dijkstra and PACO, taking into consideration the average travel time, waiting time, speed, distance, air pollution and fuel consumption as the evaluation metrics. The results from SUMO show that ITMV outperforms than others in the case of the average travel time, speed, air pollution and fuel consumption even when the number of vehicles is very high.

REFERENCES

[1] D. Krajzewicz, J. Erdmann, M. Behrisch and L. Bieker, "Recent development and applications of SUMO—Simulation of Urban MObility," Internationl Journal On Advanced Systems and Measurements, Vol. 5, Issue 3&4, pp. 128-138, December 2012.

[2] H. Mohammadzadeh and S. Joudi Bigdello, "UTCARP: Urban Traffic Control Aware Routing Protocol," International Journal on AdHoc Networking Systems (IJANS), Vol. 3, Issue 1, pp. 1-13, January 2013.

[3] J. Dallmeyer, R. Schumann, A.D. Lattner and I.J. Timm, "Don't Go With the Ant Flow: Ant-Inspired Traffic Routing in Urban Environments," Journal of Intelligent Transportation Systems: Technology, Planning, and Operations, Vol. 19, Issue 1, pp. 78-88, September 2015.

[4] L.X. Wang, "A Course in Fuzzy Systems and Control," International Edition, Prentice-Hall International Inc., 1997, pp. 1-126.

[5] M. Barth and K. Boriboonsomsin, "Traffic Congestion and Greenhouse Gases," ACCESS Magazine, University of California Transportation Center, pp. 2-9, October 2009.

[6] M. Kimura, S. Inoue, Y. Taoda, T. Dohi and Y. Kakuda, "A Novel Method Based on VANET for Alleviating Traffic Congestion in Urban Transportations," IEEE Eleventh International Symposium on Autonomous Decentralized Systems (ISADS), 2013, pp. 1-7.

[7] M. Korkalainen, M. Sallinen , N. Karkkainen and P. Tukeva, "Survey of Wireless Sensor Networks Simulation Tools for Demanding Applications," Fifth International Conference on Networking and Services(ICNS) 2009, pp. 102-106.

[8] M.R. Jabbarpour, A. Jalooli, E. Shaghaghi, R. Noor, L. Rothkrantz, R. Hafeez Khokhar and N. Badrul Anuar, "Ant-based vehicle congestion avoidance system using vehicular networks," Engineering Applications of Artificial Intelligence, Vol. 36, pp. 303–331, August 2014.

[9] P. Yousefi and R. Zamani, "The optimal routing of cars in the car navigation system by taking the combination of divide and conquer method and ant colony algorithm into consideration," International Journal of Machine Learning and Computing (IJMLC), Vol. 3, Issue 1, pp. 44–48, February 2013.

[10] R. Naja and R. Matta, "Fuzzy Logic Ticket Rate Predictor for Congestion Control in Vehicular Networks. Wireless Personal Communications," Wireless Personal Communications, Volume 79, Issue 3, pp. 1837-1858, December 2014.

[11] S. Inoue, Y. Taoda and Y. Kakuda, "An Alleviating Traffic Congestion Scheme Based on VANET with a Function to Dynamical Change Size of Area for Traffic Information in Urban Transportations," 10th International Symposium on Autonomous Decentalized Systems (ISADS), 2011, pp. 299-302.

[12] Z. Jiang, J. Wu and P. Sabatino, "GUI: GPS-Less Traffic Congestion Avoidance in Urban Areas with Inter-Vehicular Communication," The 11th IEEE International Conference on Mobile Ad hoc and Sensor Systems (IEEE MASS), 2014, pp. 19-27.

[13] Z. Liang and Y. Wakahara, "Real-time urban traffic amount prediction models for dynamic route guidance systems," EURASIP Journal on Wireless Communications and Networking, Volume 2014, Issue 85, pp. 1-13.

An Approach of Self-Organizing Systems Based on Factor-Order Space

Jin Li, Ping He*
Department of Information, Liaoning Police College
Dalian, 116036 Liaoning, China

***Abstract*—To explore new system self-organizing theory, it's urgent to find a new method in the system science. This paper combines factor space theory with system non-optimum theory, applies it into the research of system self-organizing theory and proposes new concepts as system factor space, object-factor and space-order relation. It constructs factor-space framework of system self-organizing based on factor mapping and object inversion, studies system ordering from a new perspective with optimum and non-optimum attributes as the basis of system uncertainty, and expands factor space theory from $f(0,\overline{o})$ to $f(o,0)$. The research suggests that the construction of system factor space is to build an information system capable of self-learning for system self-organization and better enhance functions of system self-organization by adopting information fusion of data analysis and perception judgment.**

Keywords*—Self-organization; factor-order space; extended order; extended entropy; systems level

I. Introduction

From the perspective of general system theory, system is a unity (complex, whole) composed by several interrelated elements (objects). The purpose of any system study is how to exert the function of the system, and to achieve acceptable goal. The main research content of the traditional system theory: determine the running rules and changing rules that fit the characteristics of system, and thus, make the system from disordered structure (random) into a certain behavior and target of ordered structure. The dissipative structure theory of I. Prigogine, the synergy theory of Herman Hake and the mutation theory of R. Thom have established a relatively complete theoretical system, and made a contribution to the development of system science. [1] It is not difficult to find out traditional research of self-organizing theory ignores the factor study which decides the existence and change of the objects. Actually, factors are roots of the existence and development of anything, like the genes of an organism. So, it is worthy of introducing the factor theory into the study of system self-organization. [2]

With the deepening of research on the theory and application of system self-organization, the researchers find out that the order of a system is influenced by related factors with positive and negative attributes, and the recognition and adjustment of different factor attributes is an important content of the system self-organization process. Literature [3] defines positive attribute as optimum attribute, negative attribute as non-optimum attribute. Literature [4] establishes a non-optimum diagnosis model of network system and proposes that non-optimum attribute is the risk source of network system by adopting non-optimum theory to the order of network system. The literature research on relevant information system discovers [5-8] that the order of information system comes from the order of information factors, that is to say, the ordered structure and the mode of an information system is realized in the process of self-organization of factor-space. Under this background, and based on non-optimum and factor space theory [9], this paper proposes is a kind of research method for self-organization from a new system thinking perspective, and introduces a new research content for the theory of factor space. Two main contributions of this paper are: to establish factor space theory of system self-organization; to propose orderly structured evaluation method based on factor-order.

The objective of this paper is to analyze non-optimum problems of the system's self-organization and to discuss a method for evaluating systems level based on the theory of non-optimum analysis. The paper also put forward a new concept called extended entropy with λ order (optimum order), η order (non-optimum order) and $\lambda-\eta$ order (meso-optimum order).

The remaining parts of the paper are structured as follows: We discuss the basic conception of factor space, as well as the factors level based on factor-order in Section 2. In Section 3 we present a practical notion to describe self-organizing systems based on extended order. Finally, Section 4 concludes the paper and sketches possible further research and applications based on the presented approach.

II. Factor Space and Factor Order

A. Non-optimum and Self-organization

The self-organization system arises during balancing that optimum and non-optimum attributes. Thus, if a system factor has optimum and non-optimum attributes at same time, it is called meso-optimum factor [10]. In traditional self-organization theory, the standards of system order are expressed by optimum attributes of system factor. But it is still difficult to solve the system analysis problem with non-optimum attributes. The primary cause is uncertainty of optimum attribute and the existence of non-optimum ones. In fact, the process of the self-organization is based on the comparison of optimum and non-optimum attributes.

The diagnosis of the system uncertainty is created by the reversed way of thinking. Therefore, in research self-

organization system with the uncertainty, we must be considering the optimum and the non-optimum attributes of the system factors, as well as changes of these attributes. In fact, the diagnosis of the system uncertainty is the most important parts of self-organization system. It consists of several processes: (1) Non-optimum attributes identification of self-organization system. (2) Non-optimum attributes evaluation of self-organization system.

B. Basic Conception

Definition 2.1 Let $P=\{p_1, p_2, \cdots, p_L\}$ be a projects set of the system, $F=\{f_1, f_2, \cdots, f_k\}$ be a factors set of the projects, $R(o,\bar{o})$ be a factor relation based on the optimum and the non-optimum attribute, then $\Omega=(P, f, R(o,\bar{o}))$ be called a factor space of the system.

The factor set F of the system is divided into three parts: optimum attribute, non-optimum attribute and unknown attribute. If it is a complete factor space of the system, the measurement of unknown attribute must be equal to zero. Otherwise it is an incomplete factor space of the system. Therefore a complete factor space can be regard as a special case of an incomplete factor space. [8] The basic frame of factor space is shown in Fig. 1.

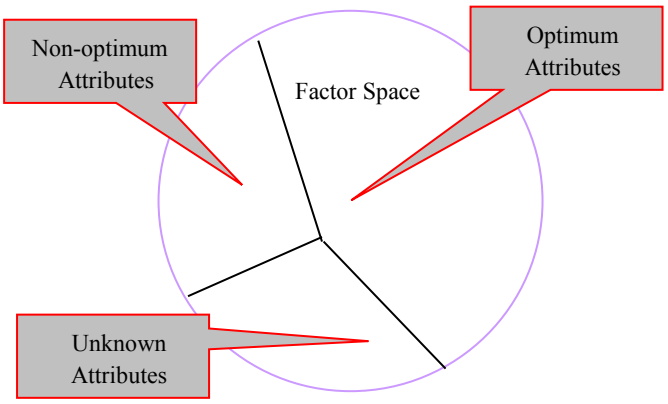


Fig. 1. Incomplete factor space

Definition 2.2 Let $P=\{p_1, p_2, \cdots, p_L\}$ be a projects set of the system, $F=\{f_1, f_2, \cdots, f_k\}$ be a factors set of the projects, and $f(o,\bar{o})$ a binary relation defined between O and $\bar{O}$, where $O=\{o_1, o_2, \cdots, o_n\}$ be a set of the optimum attributes of F, $\bar{O}=\{\bar{o}_1, \bar{o}_2, \cdots, \bar{o}_n\}$ be a set of the non-optimum attributes of F, then there exist to be a set $f(o,\bar{o})=\{<o_i,\bar{o}_i>: o_i \in O, \bar{o}_i \in \bar{O}\}$ where $i=1, \cdots, n$, and $R(o,\bar{o})=f(o,\bar{o})$ be called a relationship of the meso-optimum based on the optimum and the non-optimum attributes.

We use an ordered triad $f(o,\bar{o})$ as the meso-optimum relationship of the factor based on the optimum and non-optimum attributes for describing system, called the factor-order of the system.

C. The Analysis of Systems Level Based on Factor-Order

In the research of the self-organization of the uncertain system, the goals of the self-organization are the transformations on the factor level of the system. If the transformations are negative, we can call them friction non-optimum. If they are positive, we can call them optimum. Thus, meso-optimum relation $R(o,\bar{o})$ is decided by differences the optimum and non-optimum degree and if the results of decisions fall in this trusted classification, we can call them trusted, which decide the optimum degree and the non-optimum of the system factor.

Definition 2.3 Let $v(o,\bar{o})$ be a value set of factor-order $f(o,\bar{o})$ of the system, then there exist to be a mapping

$f_j : v(o,\bar{o}) \to [-1,1],\ f_j \in F, (j=1,2,\cdots,m)$, where the $v(o,0) \to [0,1]$, the $v(0,\bar{o}) \to [-1,0]$ and denote respectively the degree of the positive level, the degree of the negative level, if $v(o,\bar{o}) \to [-1,1]$ then denote respectively the degree of the neutrally level.

The major function of the self-organization of the system is to optimal the running of the system, develop its goals; they have to experience from minimum non-optimum to maximum optimum, that is

$$f_j(v(\text{Min}\ (o), \text{Max}(\bar{o}))) \to f(o,\bar{o})$$

$$\to f_j(v(\text{Max}\ (o), \text{Min}(\bar{o}))),$$

and from maximum optimum to minimum non-optimum, that is

$$f(v(\text{Max}\ (o), \text{Min}(\bar{o}))) \to f(o,\bar{o})$$

$$\to f(v(\text{Min}\ (o), \text{Max}(\bar{o})))$$

If the system is not featured with this attribute, it doesn't need self-organization either. The analysis shows that systems always stay on the border of optimum and non-optimum, that is, $f(v(o,\bar{o}))$ and the aim of self-organization is to bring the system from the satisfaction of the lower levels to the higher level. Of course, the actual angle of the system doesn't have the optimal criteria, and it is also not necessary to make sure what is optimal. As long as the system can shorten the time of moving from the non-optimum to the border and from the border to the optimum, the system is satisfactory [9].

The traditional self-organization theories discuss the system level problems under the condition of free borders and fixed borders, which actually reflect subjective and objective restrictions. Whatever the subjective and objective restrictions are, the satisfaction level of a system under non-optimum conditions is called meso-optimum. In the actual analysis,

under certain optimal interval of aims and results, the premise is to find out the correspondent non-optimum interval. Thus, a meso-optimum interval is decided by the optimum and non-optimum interval and if the results of decisions fall in this interval, we can call them trusted meso-optimum. All systems have optimum and non-optimum attribute if the optimum attributes of the system can be distinguished and the non-optimum attributes of the system can be controlled, moreover, the goal of the system can be selection, then the system goal exists in a maximize the satisfaction. [11]

Definition 24 Let λ is a parameter of $f(v(o,0))$ in factor space, and $\lambda=\{\lambda_1,\cdots,\lambda_n\}$, then

$$v^{\lambda}(o,0)=\{v(o_i,0)\geq\lambda:0\leq\lambda\leq 1,o_i\in O,i=1,\cdots,n\}$$

is called the λ level of positive or is called optimum order (λ order), and λ is order parameter.

Definition 2.5 Let η is a parameter of $f(v(0,\bar{o}))$ in factor space, and $\eta=\{\eta_1,\cdots,\eta_n\}$, then

$$v^{\eta}(0,\bar{o})=\{v(0,\bar{o}_i)\leq\eta:-1\leq\eta\leq 0,\bar{o}_i\in\bar{O},i=1,\cdots,n\}$$

is called η -level of negative or is called non-optimum order (η order), and η is a unordered parameter.

Definition 2.6 Let $\lambda-\eta$ is a parameter of $f(v(o,\bar{o}))$ in factor space,

$\lambda-\eta=\{\lambda_1-\eta_1,\lambda_2-\eta_2\cdots,\lambda_n-\eta_n\}$,

Then

$$v^{\lambda-\eta}(o_i,\bar{o}_i)=\{\eta\leq v(o_i,\bar{o}_i\leq\lambda:\lambda>0,\eta<0,$$
$$o_i\in O,\bar{o}_i\in\bar{O},\ i=1,\cdots,n\}$$

is called the $\lambda-\eta$ level of meso-optimum effects or is called meso-optimum order ($\lambda-\eta$ order), and $\lambda-\eta$ is an order parameter of the system self-organizing.

We will be called extension order about synthesize λ order, η order and $\lambda-\eta$ order, that is, $\{\lambda,\eta,\lambda-\eta\}$ order.

Definition 2.4 belongs to the issue of the traditional self-organizing, and it is a core issue that how to determine λ of factor-order space, thus there is a time limitation on the system's stay in the optimum category. Within a certain time, because the system is stable, it stays in the optimum category. Definition 2.5 shows that, when the factors of the system cannot determine the optimum attributes, this system belong to the category of the non-optimum, then the function of the self-organization system is how to determine the minimum non-optimum. Definition 2.6 shows that there are the optimum and non-optimum attributes to everything, and the final direction of the system can be only achieved through practice and the transition of the optimum and non-optimum attributes. The state of the system decides its goal by choosing between optimum and non-optimum attributes. [12]

According to definition 2.4-2.6, a factor-order of the self-organizing has three factor characteristics, that is, the factor-order of the positively level

$$A_{\lambda}=(F,c(o,0),v^{\lambda}(o,0)),$$

the factor-order of the negatively level

$$A_{\eta}=(F,c(0,\bar{o}),v^{\eta}(0,\bar{o}))$$

and the factor-order of the neutrally level

$$A_{\lambda-\eta}=(F,c(o,\bar{o}),v^{\lambda-\eta}(o,\bar{o}))\ .$$

Thus, the factor-order of the self-organizing can be as the following formulation:

$$A=\begin{pmatrix} F & c(o,0) & v^{\lambda}(o,0)\\ & c(0,\bar{o}) & v^{\eta}(0,\bar{o})\\ & c(o,\bar{o}) & v^{\lambda-\eta}(o,\bar{o})\end{pmatrix}$$

In the analysis of the self-organizing system, to some extent, the optimum and non-optimum attributes can be judged and controlled based on experience and intuition. In fact, the optimum and the non-optimum attributes depends on recognition degree in self-organizing process of system. Some factor-order of system are both optimum and non-optimum attributes. For instance, there is acceptable and unacceptable aspect when the decision is made. At the same time, there exist satisfactory results, as well as unsatisfactory ones, et al. The uncertainty decision arises during balancing those optimum and non-optimum attributes [13].

III. Self-Organization Based on Extended Entropy

A. Extended Order Based on Meso-optimum

The key to solving self-organizing level is the study of properties about factor-order of the system, from which the model of self-organization system is formed. In the course of solving self-organizing level, we must come out of the habitual domain; open up systems involved in the problems and put forward creative methods. In fact, a system factor-order with meso-optimum in self-organization system, which is including the maximum optimum order and the minimum non-optimum order, the minimum optimum order and maximum non-optimum order. The above meso-optimum, which makes it possible to open up systems from different viewpoints, is the basis for both meso-optimum thinking and solving self-organizing level.

The study on the self-organizing level of systems helps us understand the optimum and non-optimum between different parts more clearly. The theory of system self-organizing has given a kind of description of the order structure of systems, which studies systems from the composition and the relationship between optimum attributes and non-optimum

attributes of system factors. The analysis on lots of system has shown that the order structure of systems can be studied from meso-optimum nature of system factors. From the view of meso-optimum nature of system factors, a system factor can be divided into two attributes: optimum attributes and non-optimum attributes. It is saying that the optimum attributes are the base, and the non-optimum attributes are what we used. Every system is the entity of the optimum and the non-optimum attributes.

Thus, every factor, and every system can be seen as factor-order with goals and meso-optimum order aiming to reach those system goals. The meso-optimum order of the factor-order can affect (positively, negatively, or neutrally) the fulfillment of the goals of other factor-order, thereby establishing a relation. The measurement of the system's goals of a factor-order can be represented by using meso-optimum order $v(o,\bar{o})$ based on optimum and non-optimum attributes.

The attainability of the objective of the system shows that the distance between the recognized goal of the system and the actual goal of the system is acceptable. The achievability of the function of the system refers that the actual functional resources are near to the objective-required resources. The controllability of the environment of the system refers to the self-organizing capacity or the order parameters achieving the permitted value. [14]

In the self-organization system, every factor and every system, can be seen as extended factor-order $A=(F,c,v)$ with goals and behaviors aiming to reach those goals. The characteristics of factor-order can affect (positively, negatively and neutrally) the fulfillment of the goals of other factor-order, thereby establishing a relation. The satisfaction or fulfillment of the goals of a factor-order can be represented using the set of the extended order parameter, which is:

$$\sigma=\{\sigma^{\lambda},\sigma^{\eta},\sigma^{\lambda-\eta}\}=\{v^{\lambda}(o,0),v^{\eta}(0,\bar{o}),v^{\lambda-\eta}(o,\bar{o})\}.$$

Relating this to the higher, the satisfaction of a system σ_{sys} can be recursively represented as a function

$$f_j: v(o,\bar{o})\to[-1,1]\ \ (f_j\in F,\ j=1,\cdots,m)$$

And, $E=\{e_1,\cdots,e_n\}$, thus

$$\begin{aligned}\sigma_{sys} &= f_j(\sigma_{sys}^{\lambda},\sigma_{sys}^{\eta},\sigma_{sys}^{\lambda-\eta})\\ &= f_j(v^{\lambda}(o,0),v^{\lambda-\eta}(o,\bar{o}),v^{\eta}(0,\bar{o}),\\ &\quad W(w^{\lambda},w^{\eta},w^{\lambda-\eta}))\end{aligned}$$

We have:

$$\begin{aligned}\sigma_{sys}^{\lambda} &= f_j(\sigma_1^{\lambda},\sigma_2^{\lambda},\cdots,\sigma_n^{\lambda},w_1^{\lambda},w_2^{\lambda},\cdots,w_n^{\lambda})\\ &= f_j(v_1^{\lambda}(o,0),v_2^{\lambda}(o,0),\cdots,v_n^{\lambda}(o,0),w_1^{\lambda}(o,0),\\ &\quad w_2^{\lambda}(o,0),\cdots,w_n^{\lambda}(o,0))\end{aligned}$$

$$\begin{aligned}\sigma_{sys}^{\eta} &= f_j(\sigma_1^{\eta},\sigma_2^{\eta},\cdots,\sigma_n^{\eta},w_1^{\eta},w_2^{\eta},\cdots,w_n^{\eta})\\ &= f_j(v_1^{\eta}(0,\bar{o}),v_2^{\eta}(0,\bar{o}),\cdots,v_n^{\eta}(0,\bar{o}),w_1^{\eta}(0,\bar{o}),\\ &\quad w_2^{\eta}(0,\bar{o}),\cdots,w_n^{\eta}(0,\bar{o})\end{aligned}$$

$$\begin{aligned}\sigma_{sys}^{\lambda-\eta} &= f_j(\sigma_1^{\lambda-\eta},\sigma_2^{\lambda-\eta},\cdots,\sigma_n^{\lambda-\eta},w_1^{\lambda-\eta}(o,\bar{o}),\\ &\quad w_2^{\lambda-\eta}(o,\bar{o}),\cdots,w_n^{\lambda-\eta}(o,\bar{o}))\\ &= f_j(v_1^{\lambda-\eta}(o,\bar{o}),v_2^{\lambda-\eta}(o,\bar{o}),\cdots,v_n^{\lambda-\eta}(o,\bar{o}),\\ &\quad w_1^{\lambda-\eta}(o,\bar{o}),w_2^{\lambda-\eta}(o,\bar{o}),\cdots,w_n^{\lambda-\eta}(o,\bar{o}))\end{aligned}$$

where $w_i^{\lambda}(o,0),w_i^{\eta}(0,\bar{o}),w_i^{\lambda-\eta}(o,\bar{o})(i=1,2,\cdots,n)$ are determined tautologically by the importance of σ_i^{λ}, of σ_i^{η}, and of $\sigma_i^{\lambda-\eta}$ of each factor to the satisfaction of the systems $\sigma_{sys}^{\lambda},\sigma_{sys}^{\eta}$ and $\sigma_{sys}^{\lambda-\eta}$.

B. Measurement of Self-organization Level

1) η order and η entropy

In reality, every uncertain system belongs to the non-optimum category. It meets the recognition and realization of mankind to analyze the causes of non-optimum category and the ways to reach optimum from the viewpoint of the non-optimum category. According to the concept of the factor-order analysis, when $\sigma=v^{\eta}(0,\bar{o})\in[-1,0]$, we have

$$\begin{aligned}\sigma_{sys}^{\eta} &= f_j(\sigma_1^{\eta},\sigma_2^{\eta},\cdots,\sigma_n^{\eta},w_1^{\eta},w_2^{\eta},\cdots,w_n^{\eta})\\ &= f_j(v_1^{\eta}(0,\bar{o}),v_2^{\eta}(0,\bar{o}),\cdots,v_n^{\eta}(0,\bar{o}),\\ &\quad w_1^{\eta}(0,\bar{o}),w_2^{\eta}(0,\bar{o}),\cdots,w_n^{\eta}(0,\bar{o})\end{aligned}$$

The satisfaction of this system σ_{sys}^{η} belongs to the issue of the self-organizing in the non-optimum category, where $w_i^{\eta}(0,\bar{o})$ is determined tautologically by the share of σ_i^{η} ($\sigma_i^{\eta}=v_i^{\eta}(0,\bar{o})$, the characteristics values of the non-optimum) of each factor to the satisfaction of the systems σ_{sys}^{η}. Thus, we can be design a entropy of the non-optimum (or is called η entropy), that is

$$H(P(\sigma_{sys}^{\eta}))=-\sum_{i=1}^{n}P(w_i^{\eta}\sigma_i^{\eta})\log P(w_i^{\eta}\sigma_i^{\eta}) \qquad (1)$$

The goals of the system self-organizing are to achieve minimum η entropy in the non-optimum category.

2) Extended entropy based on $\lambda-\eta$ order

The groundwork of the self-organization theory of the system is the systematic non-optimum analysis doctrine. For any uncertain system, whether it has entered the non-optimum category or gone out of non-optimum category are judged

through factor-order. In order to hold the system in the optimum category under certain degree and stage, we have to recognize and control the non-optimum attributes of the system factor through extended order. As we know, the non-optimum attributes of the system factor are not only dynamic, but also evaluative. In order to measure the degree of the evolution of the system's meso-optimum, the criteria of evolution have to be set up (meso-optimum criteria). According to the concept of the factor-order space, when $\sigma = v^{\lambda-\eta}(o,\bar{o}) \in [-1,1]$, we have

$$\sigma_{sys}^{\lambda-\eta} = f_j(\sigma_1^{\lambda-\eta}, \sigma_2^{\lambda-\eta}, \cdots, \sigma_n^{\lambda-\eta}, w_1^{\lambda-\eta}(o,\bar{o}), w_2^{\lambda-\eta}(o,\bar{o}), \cdots, w_n^{\lambda-\eta}(o,\bar{o})) = f_j(v_1^{\lambda-\eta}(o,\bar{o}), v_2^{\lambda-\eta}(o,\bar{o}), \cdots, v_n^{\lambda-\eta}(o,\bar{o}), w_1^{\lambda-\eta}(o,\bar{o}), w_2^{\lambda-\eta}(o,\bar{o}), \cdots, w_n^{\lambda-\eta}(o,\bar{o}))$$

The satisfaction of this system σ_{sys}^{η} belongs to the issue of the self-organizing in the optimum category and non-optimum, where $w_i^{\lambda-\eta}(o,\bar{o})$ is determined tautologically by the share of $\sigma_i^{\lambda-\eta}$ ($\sigma_i^{\lambda-\eta} = v_i^{\lambda-\eta}(o,\bar{o})$, the characteristics values of the optimum and non-optimum, at the same time) of each element to the satisfaction of the systems $\sigma_{sys}^{\lambda-\eta}$. Thus, we can be design extension entropy of the optimum and non-optimum (or is called $\lambda-\eta$ entropy), that is

$$H(P(\sigma_{sys}^{\lambda-\eta})) = \frac{H(P(\sigma_{sys}^{\lambda} - \sigma_{sys}^{\eta}))}{H(P(\sigma_{sys}^{\lambda})) + H(P(\sigma_{sys}^{\eta}))} = \frac{-\sum_{i=1}^{n} P(w_i^{\lambda}\sigma_i^{\lambda} - w_i^{\eta}\sigma_i^{\eta}) \log P(w_i^{\lambda}\sigma_i^{\lambda} - w_i^{\eta}\sigma_i^{\eta})}{-\sum_{i=1}^{n} P(w_i^{\lambda}\sigma_i^{\lambda}) \log P(w_i^{\lambda}\sigma_i^{\lambda}) - \sum_{i=1}^{n} P(w_i^{\eta}\sigma_i^{\eta}) \log P(w_i^{\eta}\sigma_i^{\eta})} \quad (2)$$

In the (2), the satisfaction of this system σ_{sys}^{λ} belongs to the issue of the traditional self-organizing [14], where is w_i^{λ} determined tautologically by the share of σ_i^{λ} ($\sigma_i^{\lambda} = v_i^{\lambda}(o,0)$, the characteristics values of the optimum) of each factor to the satisfaction of the σ_{sys}^{λ} .

The extended entropy is useful because it gives an analysis method of the factor-order representation for the self-organization level of the system, which is measurement of an uncertain system. An extended order would assume that the extended entropy of the factors of a system would be also extended entropy of the self-organization level of the system. However, this is not always the case, since some order parameters can "take advantage" of other order parameters. Thus, we need to concentrate also on the cooperation of the factor-order. [15]

C. Measurement of Self-organization Level

If the self-organization level of a system considers more than two levels, then $\sigma_{sys}^{\lambda-\eta}$ of higher levels will be recursively determined by the extended entropy of lower levels. However, f_i 's most probably will be very different on each level. Certainly, an important question remains: how do we determine the function f_i and the extended order parameter? To this question, there is no complete answer. One option would be to approximate numerically f_i . An explicit f_i may be difficult to find, but an approximation can be very useful. Another method consists of function the system: removing or altering factors of the system, and observing the effect on $\sigma_{sys}^{\lambda-\eta}$. [16]

Through analyzing the effects of different lesions, the function f_i can be reconstructed and the $\{\lambda, \eta, \lambda-\eta\}$ obtained. If $\Delta(w_i^{\lambda}\sigma_i^{\lambda} - w_i^{\eta}\sigma_i^{\eta})$ is a small change in any σ^{λ} produces $|\Delta\sigma_{sys}^{\lambda-\eta}| \geq |\Delta\sigma_i^{\lambda}| + |\Delta\sigma_i^{\eta}|$, the system can be said to be the level of the satisfaction. What could then be done to maximum $\sigma_{sys}^{\lambda-\eta}$? How can we relate $\sigma_i^{\lambda-\eta}$'s and avoid conflicts between factors? This is not an obvious task, for it implies bounding the factor-order's characteristics that reduce other σ_i^{η} 's while preserving their functionality. [17]

Not only should the optimum or the non-optimum between $\Delta(w_i^{\lambda}\sigma_i^{\lambda} - w_i^{\eta}\sigma_i^{\eta})$ of factors be minimized, but the synergy or "positive interference" should also be promoted. Dealing with complex systems, it is not feasible to tell each factor what to do or how to do it, but their behaviors need to be constrained or modified so that their goals will be reached, blocking the goals of other factors as little as possible. These constraints can be called mediators. They can be imposed from the top down, developed from the bottom up, be part of the environment, or be embedded as an aspect of the system. Mediators are determined by an observer, and can be internal or external to the system (depending on where the sub-optimum sets the boundaries of the system) (He Ping, 2011).

Confusion may arise when people describe systems as the lower level causing change (that is $w_i^{\lambda}\sigma_i^{\lambda} \leq w_i^{\eta}\sigma_i^{\eta}$) in the emergent properties of the same system. Vice versa, downward causation is the idea that higher level properties (that is $w_i^{\lambda}\sigma_i^{\lambda} \geq w_i^{\eta}\sigma_i^{\eta}$) constrain or control components at the η order. Speaking about causality between λ order and η order is not accurate, in fact, they are dynamic. What we could say is

that when we observe certain conditions in the η order (lower level), we can expect to observe certain properties at the λ order (higher level) and vice versa. There is correlation function, but not actual causation.

IV. Conclusion

Factor space theory is a new theory of Information Science and System Science, and the basic principle of the system non-optimum science is to discuss the choice of uncertainty from the crossing perspective between optimum and non-optimum. Based on the theory of space factor, this thesis adopts the use of non-optimum to explore system self-organization, and proposes several new concepts about system self-organization by applying factor-order into the discussion of system uncertainty. The research shows that the system factor space establishes the information system, and the factor-state of optimum and non-optimum attribute is an effective parameter to describe the system ordering. The key to self-organization of uncertain system is the realization of the two mappings, and the transformation of disorder and order can be achieved by factor mapping and object inversion. The research content of this paper will get further development in the following several aspects: Firstly, how to find factor of object is an important part of the study of factor space in the design and operation of the system. It should be made clear that the premise of system ordering and optimization is not to find knowledge, but factor. Because only after the finding of factor can find knowledge. Secondly, factor-order is the key to study optimum and non-optimum. In the study, it is necessary to introduce the human-computer interaction algorithm to solve the problem between the perception, and how to effectively improve the data analysis and judgment, collaborative perception is also a research subject.

Acknowledgments

The authors are grateful for the support given by National Natural Science Foundation of China (Grant No. 61272170). We also thank reviewers for insightful and helpful suggestions.

References

[1] P Jaroslaw, P Filip, S Tomasz, Theoretical model for mesoscopic-level scale-free self-organization of functional brain networks, IEEE Transactions on Neural Networks, 2010, vol.21,no.11, pp.47-58.

[2] K Nobuhiko, T Shoji, S Yasuyuki, Establishment of self-organization system in rapidly formed multicellular heterospheroids, Biomaterials, 2011, vol.32, no.26,pp.6059-6067.

[3] He Ping, Theories and Methods of Non-optimum Analysis on Systems, Journal of Engineering Science, 2004, vol.2, no. 1, pp. 73-80.

[4] He Ping, Risk Assessment of Network Security Based on Non-optimum Characteristics Analysis, International Journal of Advanced Computer Science and Applications, 2013, vol.4, no.10, pp.73-79.

[5] Ping He, System Non-optimum Analysis and Extension Optimum Theory, In: Guangya Chen, ed, Proc. of the Int'l conf on Systems Science and Systems Engineering, 2003, pp.131-137.

[6] He Ping, Method of System Non-optimum Analysis in Crisis Management, Proc. of Second International Conference of Information System for Crisis Response and Management, 2007, pp. 640-645.

[7] Ping He. Maximum Sub-Optimum Decision-Making Based on Non-Optimum Information Analysis. Advanced Science Letters, 2012, vol.5, no.1, pp. 376-385.

[8] Ping He. Characteristics Analysis Of Network Non-Optimum Based On Self-Organization Theory. Global Journal of Computer Science and Technology, 2010, vol.10, no.9, pp.62-77.

[9] Wang Peizhuang, Factor spaces and data science, Journal of Liaoning Technical University（Natural Science）, 2015, vol.34, no.1, pp.273-280.

[10] Ping He, Kaiqi Zou, Fuzzy Meso-optimum Sets and Trusted Optimum Analysis, ICIC EXPRESS LETTERS, Part B: Applications, 2015,vol.6, no.8, pp.2079-2086.

[11] Weidong Tao, Studies of human-computer interaction system based on trust intuition learning theory, 2010 2nd International Conference on Computer Engineering and Technology, 2010, pp.455-459

[12] Jiantong He, A New Intelligence Analysis Method Based on Sub-optimum Learning Model, 2009 ETP International Conference on Future Computer and Communication, 2009, pp.116-119.

[13] Zengtang Qu, Sub-optimum evaluation on incomplete network information system, 2010 International Conference on E-Health Networking Digital Ecosystems and Technologies (EDT), 2010, pp.465-468.

[14] HJ Morowitz, CHAPTER V–The Second Law of Thermodynamics, Cuad Geogr, 2014, vol.23, no.1, pp.109-124.

[15] ET Jaynes, On the rationale of maximum-entropy methods, Proceedings of the IEEE, 2010, vol.70, no.9, pp.939-952.

[16] S Cano-Andrade, GP Beretta, MRV Spakovsky, Steepest-entropy-ascent quantum thermodynamic modeling of decoherence in two different microscopic composite systems, Phys.rev.a, 2015, vol.17, no.3, pp. 57-73, 2015.

[17] A García-Hiernaux, J Casals, M Jerez, Estimating the system order by subspace methods, Computational Statistics, 2012, vol.27, no.3, pp.411-425.

Incorporating Multiple Attributes in Social Networks to Enhance the Collaborative Filtering Recommendation Algorithm

Jian Yi
Chongqing University of Posts and Telecommunications
Chongqing, China
Chongqing Engineering Laboratory of Network and Information Security

Xiao Yunpeng
Chongqing University of Posts and Telecommunications
Chongqing, China
Chongqing Engineering Laboratory of Network and Information Security

Liu Yanbing
Chongqing Engineering Laboratory of Network and Information Security
Chongqing University of Posts and Telecommunications
Chongqing, China

***Abstract*—In view of the existing user similarity calculation principle of recommendation algorithm is single, and recommender system accuracy is not well, we propose a novel social multi-attribute collaborative filtering algorithm (SoMu). We first define the user attraction similarity by users' historical rated behaviors using graph theory, and secondly, define the user interaction similarity by users' social friendship which is based on the social relationship of being followed and following. Then, we combine the user attraction similarity and the user interaction similarity to obtain a multi-attribute comprehensive user similarity model. Finally, realize personalized recommendation according to the comprehensive similarity model. Experimental results on Douban and MovieLens show that the proposed algorithm successfully incorporates multiple attributes in social networks to recommendation algorithm, and improves the accuracy of recommender system with the improved comprehensive similarity computing model.**

Keywords*—Recommender System; Social Networks; Collaborative Filtering; Comprehensive Similarity

I. Introduction

Social networks and recommender system are quickly becoming popular. Collaborative filtering is treated as a technique in assisting users to locate what they are interested in a timely manner [1]. Collaborative relationships in recommender systems can be represented as a social network [2], the growth of social networks and the development of personalized recommendation techniques have evidently improved users' experiences and delivered higher quality of services [3]. However, social recommender systems are significantly challenged by the data sparsity issue -- the social network topology structure shows that only a small number of users have relatively many connections with other users, and most of the users have very few or no connections. In other words, the number of users and the fan relationship follow the long tail distribution [4, 5, 6]. Also, users sharing similar interests in social networks generally have a tendency to contact with each other [7]. Traditional personalized recommendation methods fail to take into account users' social relationships and the fact that a user's interests may be affected by another user's interests through the social relationship, resulting in inferior recommendation quality.

The objective of this paper is to propose a new comprehensive similarity model to determine neighbors set and top-N items list recommended to the target user, thereby making a new contribution towards the solution of the data sparsity problem. Experiments of the approach were expounded and proved on both MovieLens and Douban dataset.

In short, the main contributions of the paper can be summarized as follows: A new comprehensive similarity measurement is proposed by devising and integrating user attraction similarity and user interaction similarity within a friend-user-item framework.

The newly proposed method outperforms some of the peer collaborative filtering algorithms.

II. Related Work

Extensive researches have been done in terms of dealing with both the data sparsity issue in collaborative filtering and the fact that traditional recommendation methods are not directly suitable for social networks.

The most frequently-used recommender systems are collaborative filtering [8, 9], they analyze users behavior in the past and mine correlations between users and items. The similarity calculation of items or users is a core problem of collaborative filtering. In [10] and [11], both users and items are considered in the determination of user similarity to improve the prediction accuracy. In [12], researchers incorporate user-based and item-based methods to reduce the computational costs. Huang presented a graph-based approach in which users' tastes are assumed to be "transitive", this approach enhances the information matrices and thereby contributes to the resolution of the data sparsity problem [13].

Social recommendation algorithms are also very popular in order to address limitation of the collaborative algorithms caused by the data sparsity. The research of [14] fuses the collaborative filtering and social network information into one model and the ensuing strategy is able to dynamically adjust the weight of each attribute resulting in a notable balance in terms of accuracy and coverage. Liu proposed a user similarity model which effectively improves the quality of recommender

systems by combing users' Weibo contents, social networks, and users' activities on Weibo [15]. Also, Roth suggested and verified an improved friend recommender system which generates groups of friends by mining implicit graphs in social networks and reportedly leads to an increase in user satisfactions for recommender systems [16]. Konstas effectively integrated social networks and the recommender system by studying additional relationships [17], and as social community discovery algorithm [18, 19, 20], Bayesian personalized ranking model [21] are proposed with social information aiming to enhance the accuracy of recommendations. There are also studies [22, 23] in the literature that investigate the roles played by social elements such as friendship and trust in the context of collaborative filtering. Surprisingly, it is interesting to note that social network information based recommendation methods may excel mathematical algorithms [24]. Rong studied how to predict a user's social connections by means of some public data involved in e-mails for purpose of recommending friends for users [25]. No matter what needs to be recommended, friends or items, it seems certain that the social network information is finding wider and deeper applications in the construction of recommender systems.

III. Incorporating Social Multiattribute for Recommenders

A. Problem Description

The selection of the nearest neighbors set for the target user is the critical task of the collaborative filtering method, and is always determined by the similarity among users. We assume that users' behavior dataset and users' social network information dataset are both available, where the former describes users' various behaviors and interests in the past, and the latter shows the following and/or being followed relationships among users.

We consider both interest similarity and social connections of users, and propose the notion of comprehensive similarity, which is defined by combining users' attraction similarity with users' interaction similarity.

The attraction similarity is a measurement of users' likeness in terms of their interests. Two users sharing the same kind of taste on most items with a small interest-gap (defined formally in Section 3.3.1) will have a high attraction similarity. In addition to the attraction similarity, we also define the interaction similarity among users to measure their interactive similarity degree in social networks.

Since users in general tend to trust more on items recommended by friends with high interaction similarities, this mechanism may provide new users who do not have any user behaviors with some high quality recommendations thereby alleviating the data sparsity problems. As such, the comprehensive similarity delivers a more accurate measurement of the analogy between users and indicates users' interests more precisely.

B. Algorithmic Framework

The proposed algorithm SoMu is based on the memory-based collaborative filtering algorithm. According to regular procedure of collaborative filtering, note that various types of information can be used to calculate the user similarity in social networks. We choose to use the information of users ratings on items to devise the attraction similarity, and the information about users interdependency in social networks to devise the interaction similarity. SoMu completes its task by the following steps.

Step No.1 is to collect the available data including social relationship, user profiles, and item profiles. And social relationship contains all users following and followed relationship. User profiles contain user id, user preference and so on, and item profiles contain item id, item association attributes and so on.

Step No.2 is to clean the above data and develop the user node adjacency matrix and item rating matrix. The two cleaned matrix respectively represent users all neighbors and all the cleaned ratings. These basic data are sources of the following similarities.

Step No.3 is to calculate attraction similarity for all users, complying with the formula defined in section 3.3.1.

Step no.4 is to calculate interaction similarity for all users, complying with the formula defined in section 3.3.2.

Step No.5 is to combine attraction similarity and interaction similarity to get the comprehensive similarity, which is the measure of dividing neighbors set.

Step No.6 is to output the personal top n recommendations for all target users. As shown in the dotted line, all the final results would react to user profiles and item profiles for further research. Figure 1 depicts the algorithmic framework of this approach.

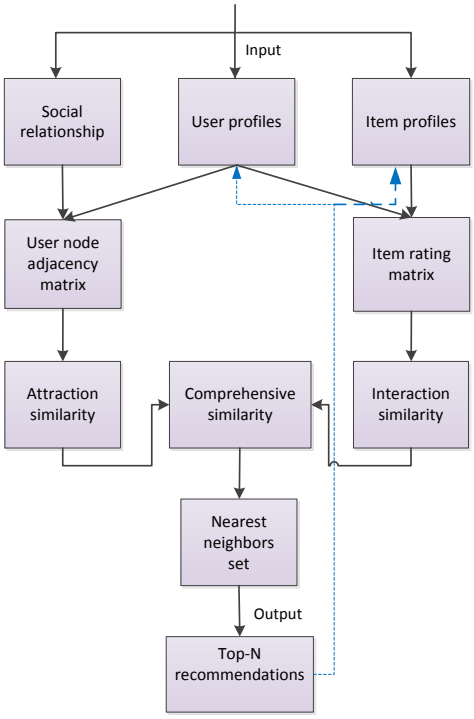


Fig. 1. Framework of the proposed algorithm SoMu

C. Comprehensive User Similarity Model

As indicated in Figure 2 (where capital letters in circles represent users and lower case letters in squares represent items), the proposed comprehensive similarity model is obtained by considering and integrating the followed-following relationship among users and the rating relationship among users and items. Specifically, the comprehensive similarity $W(u,v)$ between two users u and v is calculated as follows:

$$W(u,v) = \alpha * W(u,v)_{att} + \beta * W(u,v)_{int} \quad (1)$$

Where $W(u,v)_{att}$ and $W(u,v)_{int}$ denote the attraction similarity and the interaction similarity between and the users u and v, respectively. Also, α and β are weights satisfying $\alpha + \beta = 1$.

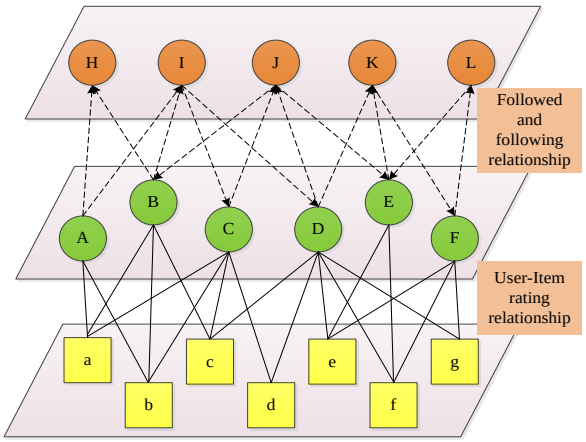


Fig. 2. The model of the comprehensive user similarity

Attraction Similarity

In user attraction similarity, the distance called interest-gap between two users is measured by their ratings on common items. A smaller interest-gap indicates a greater attraction that common items have on these two users. We describe the construction of the user attraction similarity. Considering that users and items in social networks can be regarded as nodes in graphs, weighted bipartite graphs are a natural choice for modeling the behaviors of users with respect to items.

A weighted bipartite graph is a 4-tuple (U, I, E, w), where U is the set of user nodes, I is the set of item nodes , E is the set of edges connecting user nodes and item nodes, and $w: E \to Z^+$ is a function from E to the set Z^+ of positive integers. If user (node) u has a rating for item (node) i, then there would be an edge $e \in E$ connecting node u and node i, and $w(e)$ would be the value of the rating. For example, Figure 3 shows the situation where the user set is {A, B, C} and the item set is {a, b, c, d} with A having ratings for {a, b, c}, B for {b, d}, and C for {b, c, d}.

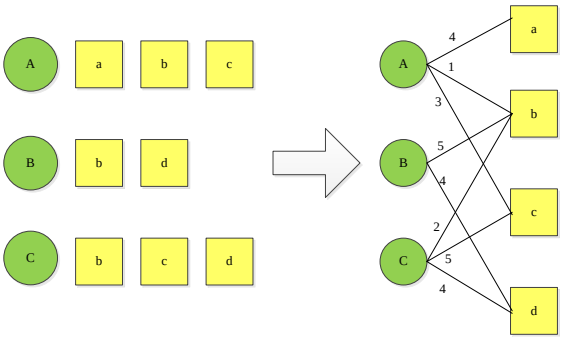


Fig. 3. User-Item rating relationship represented by a weighted bipartite graph

The attraction similarity $W(u,v)_{att}$ between users u and v is computed as follows. Let $N(u)$ be the set of items which user u has rated, $M(i)$ be the set of users who has rated the item i, r_{max} be the maximum possible rating on items, r_{min} be minimum possible rating on items, and γ be the normalized factor. Then, for any pair of users u and v, when $N(u) \cap N(v) \neq \varnothing$,

$$W(u,v)_{att} = \gamma \cdot \frac{\sum_{i \in N(u) \cap N(v)} \frac{1}{\log\left(1 + |M(i)|\right)}}{R(u,v)^2}$$

$$R(u,v) = \frac{A + B}{|N(u) \cap N(v)|}$$

$$A = \sum_{\substack{i \in N(u) \cap N(v) \\ r_{ui} \neq r_{vi}}} \left(r_{max} - min(r_{ui}, r_{vi}) + r_{min}\right)$$

$$B = \sum_{\substack{i \in N(u) \cap N(v) \\ r_{ui} = r_{vi}}} r_{min} \quad (2)$$

otherwise,

$$W(u,v)_{att} = 0. \quad (3)$$

Interaction Similarity

For popular commodities, we assign a penalty parameter to them to adjust the calculation since users will naturally have a high rating on popular commodities. The interaction similarity is computed by considering a target user's follower set and the set of users that this target user follows in a cosine-like setting. Note that the former set is the set of users who actively make friends with this target user, and the latter set is the set of users with whom this target user makes friends. Similar to that the user-item rating relationship can be modeled by weighted bipartite graphs, users following-followed relationship can be modeled by directed graphs. If user u follows user v, there would be an arrow from user node u to user node v in the graph. For any user u, we use $out(u)$ to denote the set of users whom user u follows, and $in(u)$ to denote the set of

users who follow user u. In other words, $out(u)$ represents the connections that user u has, and $in(u)$ represents the influences that user u has exerted. Figure 4 shows the following-followed relationship among users $A, B, ..., K$, where user B follows users D, E and G, and user J is followed by users D, F and G. That is, $out(B) = \{D, E, G\}$ and $in(J) = \{D, F, G\}$.

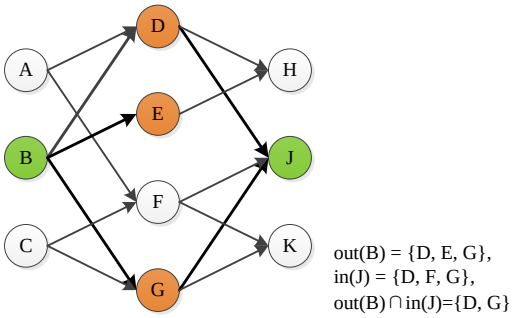


Fig. 4. Users' following-followed relationship model

Based on the observation that users tend to trust more on items recommended by friend that they follow,

$$W(u,v)_{int} = \frac{|out(u) \cap in(v)|}{\sqrt{|out(u)||in(v)|}} \quad (4)$$

we define the interaction similarity $W(u,v)_{int}$ between two users u and v to be formula 4.

D. Top-N Recommendations

In recent years, researches show that forecasting the items that will attract the user is more meaningful than predicting about what scores the user will rate on items. That is, the top-N prediction may be considered more valuable than the score prediction. Many existing recommendation algorithms are based on top-N prediction, and have great performance [26, 27]. We take the top-N approach in this paper. To generate a top-N item list for a user u, we calculate a predicting score p_{ui} for each candidate item i as follows:

$$p_{ui} = \sum_{v \in N(i) \cap S(u,k)} W(u,v) * r_{vi} \quad (5)$$

where $S(u,k)$ is the set of k nearest neighbors of user u which is decided by the integrated similarity of u to other users, and then rank items according to the scores. The algorithm for the top-N recommendation, named SoMu, is shown below. (see Algorithm 1).

This algorithm computes the attraction similarity by setting up a user-item reversal list first and then constructing a matrix W of size $|U| \times |U|$ which will be used as the numerator in the computation of the attraction similarity.

ALGORITHM 1. SoMu collaborative filtering algorithm

Algorithm SoMu Collaborative Filtering Algorithm

Input: M_{rating}, R_{follow}, K, N

Output: L_{top-N}

Procedure begin

1: Define bipartite graph mode $G(U,I,E,w)$ by the user-item rating matrix M_{rating};

2: **for** $u, v \in G$ and $u \neq v$ **do**

3: Compute $W(u,v)_{att}$;

4: **end for**

5: Define the out-degree matrix and in-degreed matrix O and I by the user following and followed-matrix R_{follow};

6: **for** $u \in O$, $v \in I$ and $u \neq v$ **do**

7: Compute $W(u,v)_{int}$;

8: **end for**

9: Generate $W(u,v)$ as the liner combination of $W(u,v)_{att}$ and $W(u,v)_{int}$;

10: Compute the K nearest neighbors by $W(u,v)$, then compute recommendation list L_{top-N} of length N;

11: **return** L_{top-N};

Procedure end

$W[u][v]$ and $W[v][u]$ will be incremented by 1 if users u, v both rate an item a. Iterating through the list of all items will give rise to the matrix W. For the denominator of the attraction similarity computation, an interest-difference gap matrix R is constructed by using a hash table which is of $O(1)$ time.

The entire computation of the attraction similarity costs time $O(n^2)$ where n is the number of nodes in U. Although the time complexity of SoMu is general, it plays well on evaluation metrics such as precision, recall, coverage, and popularity, which would be mentioned farther below.

E. Identify the Headings

We use the following metrics to evaluate the quality of the proposed top-N recommendation algorithm described in the previous section: *Precision*, *Recall*, F_{PR}, *Coverage and Popularity*.

In all formulas below, $R(u)$ represents the set of the top-N items recommended for user u; $T(u)$ represents the set of items that user u actually rates in a testing bed; U is the set of users; and I is the set of items. *Precision* and *Recall* can be seen as a pair of quality assessment for information retrieval [30], and the same for recommendations. Figure 5 shows the original definition of precision and recall.

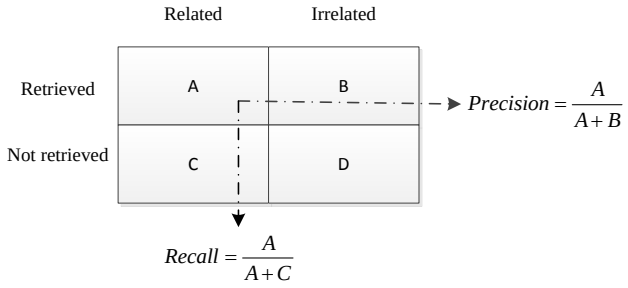


Fig. 5. Precision and Recall description

Considering that *Precision* and *Recall* are individual measurements and are related to each other, we devise another measurement F_{PR} as follows to indicate the effects of both *Precision* and *Recall*.

$$Precision = \frac{\sum_{u \in U} |R(u) \cap T(u)|}{\sum_{u \in U} |R(u)|}$$

$$Recall = \frac{\sum_{u \in U} |R(u) \cap T(u)|}{\sum_{u \in U} |T(u)|} \quad (6)$$

$$F_{PR} = \frac{(\lambda^2 + 1) Precision * Recall}{\lambda^2 (Precision + Recall)} \quad (7)$$

In (7), λ is weight, usually $\lambda = 1$ is frequently-used. In addition to *Precision*, *Recall* and F_{PR} which evaluate the accuracy of the recommendation algorithm, the notion of *Coverage* is used to indicate the long-tail exploration capability of the algorithm. In [33], data from social network was analyzed in order to find what information could improve the diversity and coverage of recommendations. the notion of *Coverage* is used to indicate the long-tail exploration capability of the algorithm.

$$Coverage = \frac{|\cup_{u \in U} R(u)|}{|I|} \quad (8)$$

Finally, *Popularity* is used to indicate whether the recommendation results are new. A smaller *Popularity* value means that most of the recommended items are not very popular and suggests that the algorithm works better. The definition of *Popularity* is given as follows [29].

$$Popularity = \frac{\sum_{u \in U} \sum_{i \in T(u)} \log(1 + |M(i)|)}{\sum_{u \in U} |T(u)|} \quad (9)$$

IV. Experiments

We in this section present the experimental testing result for the algorithm SoMu discussed in the previous section.

A. Datasets

We tested the SoMu algorithm on two publicly available datasets: MovieLens (http://www.grouplens.org) and Douban (http://datatang.com). The practical scale of experimental datasets are shown below TABLE.1.

TABLE I. Experimental Data Scale

dataset	user	item	rating	relationship
Douban	383033	80008	3648104	100000
MovieLens	934	1682	100000	\

Note that the following-followed social links in Douban dataset is unidirectional and thus can be understood as the in-degree and out-degree of the user nodes in directed graphs (as we discussed in the previous section). While a user's in-degree is the indicator of his/her social status and influences, a user's out-degree is the indicator of the number of other users that he/she cares and follows. It can be seen clearly that both users' in-degrees and out-degrees are in line with the long-tail distribution.

B. Design of the Experiment

We randomly divide the set of the user behavioral data into two parts as follows: 80% of the data is used as the training set and 20% of the data is used as the testing set. The algorithm is applied to the training set to obtain the top-N recommendation list for the user and is used on the testing set for the purpose of performance evaluation. Specifically, we set up the following three experiments:

- Using the MovieLens 100k dataset, compare SoMu algorithm with a peer collaborative filtering algorithm to see which one has a higher comprehensive measurement F_{PR}.
- Using the dataset from Douban, observe the performance of the SoMu algorithm to determine if the addition of social information into a collaborative filtering algorithm can improve the quality of the algorithm. If so, go ahead and pursue the values of the parameters K, N, α, β which may enable the best performance of the algorithm.
- Based on the result of experiment 2, compare the performance of SoMu with that of another peer social recommendation algorithm in terms of the recommendation quality.

C. Experimental Results

Experiment 1. In this experiment, compare the algorithm SoMu with one of the traditional collaborative filtering algorithms UserCF [11] in terms of the comprehensive evaluation metric F_{PR}. The comparison result is shown in Figure 6 with $\lambda = 1$ and number of neighbors ranging from 5 to 100. Given the fact that a larger F_{PR} indicates a superior algorithm, it can be seen clearly that the proposed algorithm outperforms UserCF, and would be a preferred choice when a special requirement such as finding an equilibrium between *Precision* and *Recall* is needed.

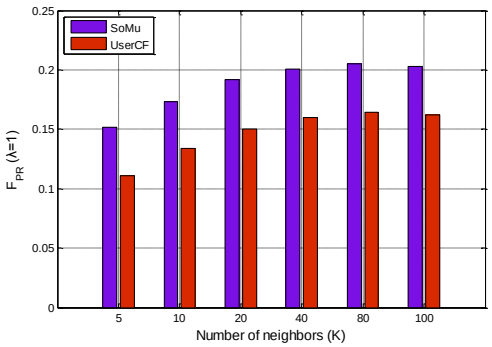


Fig. 6. Comparison between SoMu and UserCF with respect to F_PR

Experiment 2. In order to see the effectiveness (or non-effectiveness) of different similarities on the evaluation metrics *Precision*, *Recall*, F_{PR}, *Coverage* and *Popularity*, we in this experiment conduct three sub-experiments: (1) SoMu@1: implement the recommendation solely by the attraction similarity, (2) SoMu@2: implement the recommendation solely by the interaction similarity, and (3) SoMu@3: implement the recommendation by the combination of the attraction similarity and the interaction similarity.

Since the data sparsity of the user rating matrix for items may affect the result of recommendations, we conduct the experiments on the basis of various instances of data sparsity. In order to give a quantitative measurement for data sparsity, define the notion of user sparsity as follows:

$$D_{u\_spa} = S(u) \tag{10}$$

Where u denotes a user and $S(u)$ denotes the number of rated items by user u. In the experiments, D_{u_spa} is set to be many different values with any two consecutive values differentiated by 30.

Also, note that experimental results can be affected by the following parameters: *K, N, α, β.* (Recall that *K* denotes the number of the nearest neighbors to the target user in the process of recommendation, *N* is the length of the recommendation list, and α and β are weight factors in the computation of the comprehensive similarity.) As such, different experiments are devised to examine these possible and potential impacts. Specifically, algorithm SoMu@1 corresponds to the computation of formula (1) with $\alpha \neq 0$ and $\beta = 0$, and thus is completely determined by the attraction similarity; algorithm SoMu@2 corresponds to the computation of formula (1) with $\alpha = 0$ and $\beta \neq 0$, and thus is completely determined by the interaction similarity; algorithm SoMu@3 corresponds to the computation of formula (1) with $\alpha \neq 0$ and $\beta \neq 0$, and thus is completely determined by the integrated similarity proposed in this paper. Figures 7 – 10 show the comparisons of SoMu@1, SoMu@2, and SuMu@3 in terms of *Precision*, *Recall*, *Coverage* and *Popularity*.

All experiments shown in Figures 7-10 are conducted for some given and fixed user sparsity. Figure 7 illustrates the correlation between *K* and *Precision*; see that SoMu@3 outperforms SoMu@1 slightly, but beats SuMo@2 to a large extent. Figure 8 shows the correlation between *K* and *Recall*. Again, we are able to observe that SoMu@3 is superior to both SuMo@1 and SuMu@2. Figure 9 exhibits the correlation between *K* and *Coverage*, and clearly indicates that SuMu@3 has a stronger long-tail item mining capability than both SuMu@1 and SuMu@2. Figure 10 depicts the correlation between *K* and *Popularity*. A low *Popularity* of a recommender system means that the items recommended by this system are not those hot, popular commodities on the market, which indicates that this recommender system has a certain degree of novelty. A high *Popularity* of a recommender system would mean the opposite. In Figure 10, that the *Popularity* of SoMu@3 is lower than that of SoMu@1 but higher than that of SoMu@2, resulting in a balanced state in terms of recommendation novelty and item recognitions.

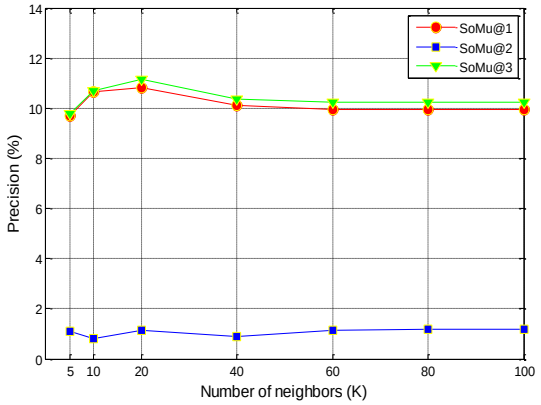


Fig. 7. Correlation between K and Precision

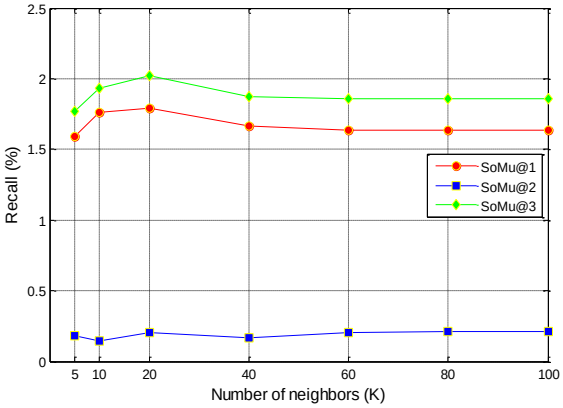


Fig. 8. Correlation between K and Recall

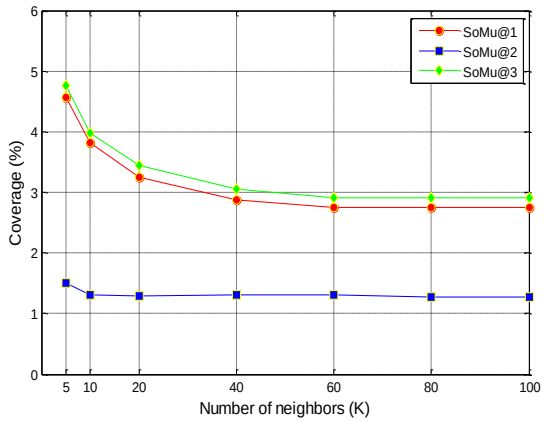


Fig. 9. Correlation between K and Coverage

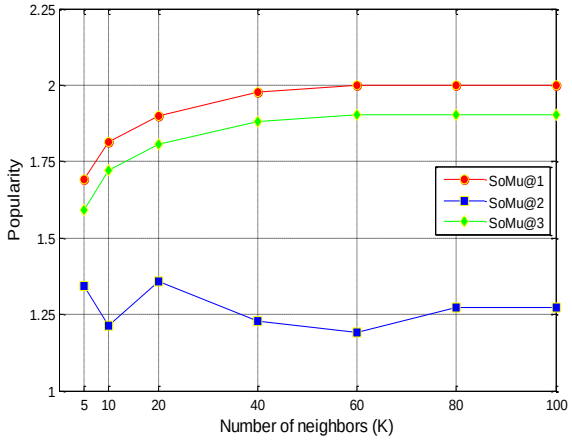


Fig. 10. Correlation between K and Popularity

Figures 7-10 clearly and heuristically indicate that the algorithm tends to be stable for all aspects when K is sufficiently large, although the rigorous such argument needs to be proved mathematically. Also, we can determine by these figures that the optimal values for K, N, α, β are K=20, N=24, α=0.988, β=0.012.

Experiment 3. In this experiment, we compare the SuMo algorithm with one of the typical social recommendation algorithms Neighbor [29]. The typical algorithm used Pearson correlation to calculate user similarity. Pearson correlation is defined as below, and the meaning of the symbols and letters is the same as above formulas, and need not be repeated here.

$$w_{ij} = \frac{\sum_{u \in U} \left(r_{ui} - \bar{r}_i\right)\left(r_{uj} - \bar{r}_j\right)}{\sqrt{\sum_{u \in U} \left(r_{ui} - \bar{r}_i\right)^2 \sum_{u \in U} \left(r_{uj} - \bar{r}_j\right)^2}} \quad (11)$$

Based on Experiment 2, set the values of parameters in SuMo to the optimal ones, i.e., K=20, N=24, α=0.988, β=0.012. The comparison results are depicted in Figures 11 and 12 and TABLE 2.

TABLE II. COMPARISON BETWEEN SUMO AND NEIGHBOR ALGORITHMS WRT PRECISION AND RECALL

K	Neighbor		SoMu	
	Precision	**Recall**	**Precision**	**Recall**
5	9.378%	1.759%	9.776%	1.769%
10	10.652%	1.927%	10.684%	1.934%
20	**11.111%**	**2.011%**	**11.165%**	**2.021%**
40	10.256%	1.856%	10.363%	1.876%
60	10.150%	1.837%	10.256%	1.856%
80	10.150%	1.837%	10.256%	1.856%
100	10.150%	1.837%	10.256%	1.856%

The data in Table 2 clearly show that SoMu exceeds Neighbor in terms of performance metrics *Precision* and *Recall*, and also show that both SoMu and Neighbor reach their own best performance at the optimal parameter setting ($K = 20$). Figure 11 demonstrates the comparison between SoMu and Neighbor with respect to evaluation metric *Coverage*. Evidently, SoMu outperforms Neighbor in *Coverage* although the two methods' performances trend in the same manner. The comparison between SoMu and Neighbor in regards to *Popularity* is given in Figure 12. We are able to note that SoMu has a higher *Popularity* than Neighbor prior to the stabilization of these two algorithms, indicating that SoMu, during this period of time, primarily recommends recognized and fashionable items to the users. However, as the algorithm tends to stabilize with the increase of K, SoMu exhibits a lower *Popularity* than Neighbor, indicating that SoMu starts to recommend non-fashionable items to the user with a sense of novelty.

In summary, Table 2, Figure 11, and Figure 12 suggest that the proposed algorithm SoMu outperforms the algorithm Neighbor in terms of all evaluation metrics. Also, we can see that all evaluation metrics tend to become a constant as the number of neighbors K is sufficient large.

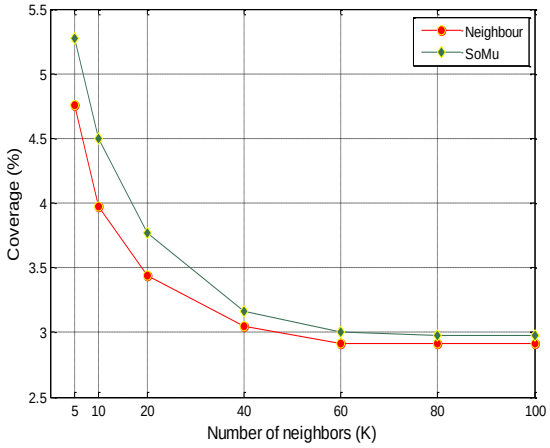


Fig. 11. Comparison between SuMo and Neighbor Algorithms wrt Coverage

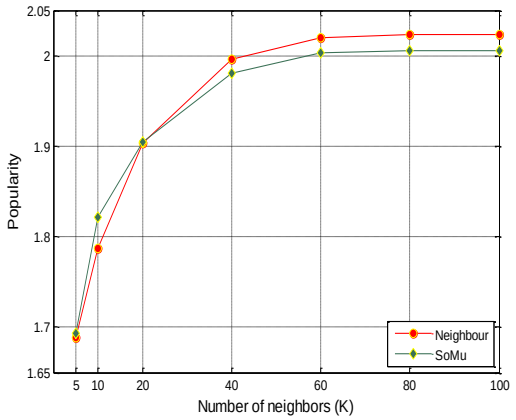


Fig. 12. Comparison between SuMo and Neighbor Algorithms wrt Popularity

V. CONCLUSIONS

We have in this paper proposed a new collaborative filtering recommendation algorithm SoMu which leverages multiple attributes in social networks to improve the recommendation result. By applying proposed to the popular datasets obtained on MovieLens and Douban and comparing the outcomes with that obtained from other peer recommendation algorithms, we have found that SoMu excels other peer algorithms in terms of recommendation evaluation metrics. As the further work, we plan to deepen the study on the correlations between recommender systems and the social networks by further investigating the relations formed among various groups on the social networks and by associating items recommendations with friend recommendations. We plan to parallelize the algorithm, and increase the amount of experimental data.

REFERENCES

[1] G Adomavicius, A Tuzhilin. "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions". Knowledge and Data Engineering, IEEE Transactions on, 2005, 17(6): 734-749.

[2] R Burke, M P O'Mahony, N J.Hurley "Robust collaborative recommendation". Recommender Systems Handbook. Springer US, 2011: 805-835.

[3] P Kazienko, K Musial, T Kajdanowicz. "Multidimensional social network in the social recommender system". Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on, 2011, 41(4): 746-759.

[4] H J Ahn. "A new similarity measure for collaborative filtering to alleviate the new user cold-starting problem". Information Sciences, 2008, 178(1): 37-51.

[5] Y J Park, A Tuzhilin. "The long tail of recommender systems and how to leverage it". Proceedings of the 2008 ACM conference on Recommender systems. ACM, 2008: 11-18.

[6] Y Moshfeghi, B Piwowarski, J M Jose. "Handling data sparsity in collaborative filtering using emotion and semantic based features". Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval. ACM, 2011: 625-634.

[7] J Odonovan, B Smyth. "Trust in recommender systems". Proceedings of the 10th international conference on Intelligent user interfaces. ACM, 2005: 167-174.

[8] C Desrosiers, G Karypis. "A comprehensive survey of neighborhood-based recommendation methods". Recommender systems handbook. Springer US, 2011: 107-144.

[9] Y Koren, R Bell. "Advances in collaborative filtering". Recommender systems handbook. Springer US, 2011: 145-186.

[10] W Jun, A P De Vries, M J T Reinders. "Unifying user-based and item-based collaborative filtering approaches by similarity fusion". Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2006: 501-508.

[11] K Zhao, L Pengyu. "Improved collaborative filtering approach based on user similarity combination". Management Science & Engineering (ICMSE), 2014 International Conference on. IEEE, 2014: 238-243.

[12] L Qingwen, X Yan, H Wenchao. "Combining User-Based and Item-Based Models for Collaborative Filtering Using Stacked Regression". Chinese Journal of Electronics, 2014, 23(4).

[13] H Zan, H Chen, D Zeng. "Applying associative retrieval techniques to alleviate the sparsity problem in collaborative filtering". ACM Transactions on Information Systems (TOIS), 2004, 22(1): 116-142.

[14] A Bellogin, I Cantador, F Dier. "An empirical comparison of social, collaborative filtering, and hybrid recommenders". ACM Transactions on Intelligent Systems and Technology (TIST), 2013, 4(1): 14.

[15] L Qingwen, X Yan, H Wenchao. "Integrating social information into collaborative filtering for celebrities recommendation". Intelligent Information and Database Systems. Springer Berlin Heidelberg, 2013: 109-118.

[16] M Roth, A Ben-David, D Deutscher. "Suggesting friends using the implicit social graph". Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2010: 233-242.

[17] I Konstas, V Stathopoulos, J M Jose. "On social networks and collaborative recommendation". Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval. ACM, 2009: 195-202.

[18] T A Dang, E Viennet. "Collaborative filtering in social networks: A community-based approach". Computing, Management and Telecommunications (ComManTel), 2013 International Conference on. IEEE, 2013: 128-133.

[19] W Yu, G Lin. "Social circle-based algorithm for friend recommendation in online social networks". Chinese Journal of Computers. 2014 (4)

[20] Y Zhimin, Y Xiangzhan, Z Hongli. "Commodity recommendation algorithm based on social network". Advances in Computer Science and its Applications. Springer Berlin Heidelberg, 2014: 27-33.

[21] Z Tong, J Mcauley, I King. "Leveraging social connections to improve personalized ranking for collaborative filtering". Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management. ACM, 2014: 261-270.

[22] B Shapira, L Rokach, S Freilikhman. "Facebook single and cross domain data for recommendation systems". User Modeling and User-Adapted Interaction, 2013, 23(2-3): 211-247.

[23] D Odoherty, S Jouili, R P Van. "Trust-based recommendation: an empirical analysis". Proceedings of the Sixth ACM SIGKDD Workshop on Social Network Mining and Analysis SNA-KDD. Beijing, China, ACM. 2012.

[24] L Fengkun, H J Lee. "Use of social network information to enhance collaborative filtering performance". Expert systems with applications, 2010, 37(7): 4772-4778.

[25] R Huigui, H Shengxu, H Chunhua,. "User Similarity-based Collaborative Filtering Recommendation Algorithm". Journal on Communications. 2014, (2): 16-24.

[26] P Cremonesi, Y Koren, R Tumin. "Performance of recommender algorithms on top-n recommendation tasks". Proceedings of the fourth ACM conference on Recommender systems. ACM, 2010: 39-46.

[27] N Barbieri, G Manco. "An analysis of probabilistic methods for top-N recommendation in collaborative filtering". Machine Learning and Knowledge Discovery in Databases, 2011: 172-187.

[28] A Bellogin, I Cantador, P Castells. "A comparative study of heterogeneous item recommendations in social systems". Information Sciences, 2013, 221: 142-169.

[29] X Liang. Recommender Systems in Action. Beijing: Posts & Telecom Press, 2012:35-73.

Using Rule Base System in Mobile Platform to Build Alert System for Evacuation and Guidance

Maysoon Fouad Abulkhair
Department of Information Technology
Faculty of Computing and Information Technology
King Abdulaziz University
Jeddah, Saudi Arabia

Lamiaa Fattouh Ibrahim
Department of Information Technology
Faculty of Computing and Information Technology
King Abdulaziz University
Jeddah, Saudi Arabia
Department of Computer Science and Information
Institute of Statistical Studies and Research
Cairo University, Cairo, Egypt

***Abstract*—The last few years have witnessed the widespread use of mobile technology. Billions of citizens around the world own smartphones, which they use for both personal and business applications. Thus, technologies will minimize the risk of losing people's lives. Mobile platform is one of the most popular platform technologies utilized on a wide scale and accessible to a high number of people. There has been a huge increase in natural and manmade disasters in the last few years. Such disasters can happen anytime and anywhere causing major damage to people and property. The environment affluence and the failure of people to go to other safe places are the results of catastrophic events recently in Jeddah city. Flood causes the sinking and destruction of homes and private properties. Thus, this paper describes a system that can help in determining the affected properties, evacuating them, and providing a proper guidance to the registered users in the system. This system notifies mobile phone users by sending guidance messages and sound alerts, in a real-time when disasters (fires, floods) hit. Warnings and tips are received on the mobile user to teach him/her how to react before, during, and after the disaster. Provide a mobile application using GPS to determine the user location and guide the user for the best way with the aid of rule-based system that built through the interview with the Experts domains. Moreover, the user will receive Google map updates for any added information. This system consists of two subsystems: the first helps students in our university to evacuate during a catastrophe and the second aids all people in the city. Due to all these features, the system can access the required information at the needed time.**

Keywords—safety; Natural disasters; smartphone; rule-based system; Mobile Network; Smart Phone

I. INTRODUCTION

Damages and ripple effects can happen due to disasters without notification anytime and anywhere [1]. Disaster break presents many unique logistics challenges, such as damage to transportation infrastructure, limited communication, and coordination of multiple agents [2]. To build a solid disaster management system, three fundamental things need to be addressed [3]:

a) To prevent disaster, we must have strong communication.

b) Mobile user location and real-time pictures of events are very important to make decisions.

c) Effective analysis and reasoning engines help in the prediction model, and reduce, and prevent disasters.

Knowledge and personal expertise is represented in the form of rule-based systems that has the form of IF-Then rules. A number of applications are suitable to use Rule-based systems [4], [5] and [6]. New types of spatial computing applications and technologies use GPS and sensors installed on smartphones and other powerful smart devices [7] and [8].

In this paper, we use Rule-based systems, smartphones equipped with GPS and other technology to make our system. Our system is divided into two subsystems:

1) Calamity communicator (iCalamityGuide) will deliver the following in terms of Hardware and Software:

- Build a communicator mobile phone application that broadcasts notification, guiding and giving correct directions as exits from buildings and nearest safe building from an authorized source to guide and locate registered people (students and employees of King Abdulaziz University, Jeddah city) by using WIFI or 3G, and GPS locator.
- Engaging the locating services such as GPS and maps reflecting King Abdulaziz University.
- Establishing a connection using PHP to integrate a double communication link between the database at the hosting server with the iPhone application and XML files. We have the ability to edit this database and reflect the changes on user request from the application.

2) This subsystem is called Disaster smart road guidance DSRG. It solves the problem of real-time disaster and road constructions information. This system notifies the users with the updated information concerning fires, floods, and constructions in Jeddah city in Saudi Arabia. The users can also receive messages about the action they must take before, during, and after a disaster, on their mobile phones. This application provides the users with nearest short path safe location using GPS technology.

Moreover, this application sends Google map updates for any newly added information to the users. Feedback action is also sent to users in a time of trouble using Rule-based system

feature upon a specific situation. The interview with the Experts domains was the aid in building the Rule-based system. [8].

The rest of this paper is organized as follows. Section 2 discusses the related work. Section 3 introduces the iCalamityGuide and DSRG system. Comparison of related work is done in Section 4. The paper conclusion is presented in Section 5.

II. Related Work

GAIA GPS (FOR HAITIAN DISASTER RELIEF) [9]: The application used to download the map when an earthquake occurs in the area. Up-to-date overlays of disaster sites, hospitals, and other relevant waypoints showed. The application also records GPS tracks, waypoints, and geo-tagged photos. One can also import /export GPX tracks photos and guidance to waypoints and along tracks.

UBALERT– DISASTER ALERT NETWORK [10]: This application supports interactive map, shoots photos for the disaster, and reports the events with more descriptions. It also sorts events according to severity, location, and popularity.

DISASTER ALERT [11]**:** This application detects the most popular types of natural disaster from their occurrence time, and it announces the same to all citizens in the same region. The Pacific Disaster Centre (PDC) organization is the main source of the application information. This application presents its information on an interactive user map using global disaster info [11].

In [12], system works on the principles of the client-Server system, wherein the server responds to the requests of the Clients. This system is implemented in two parts: the EMS Client Application, Rescue Application and Server. The Client and Rescue Application are implemented on an Android Application. The server is implemented as a web-based Java Application. This web application is useful for generating area wise issues report, priority wise issues report, and location wise issues report.

III. Description of the System

When a calamity strikes, warning alerts become very important. The existence of some social networking apps as message boards to convey information, apps for receiving news updates, and breaking news help people become aware of a situation.

The process of calamity management involves four phases: mitigation, preparedness, response, and recovery. The mitigation phase attempts to reduce calamity risks by focusing on long-term measures of eliminating calamity. The preparedness phase is the development of an action plan for an upcoming calamity. The response phase is the mobilization of services and relief when calamity strikes and the recovery phase is the restoration of the affected area to its previous state.

A. Implementation of iPhone Calamity Guide iCalamityGuide iCG

This system is built to serve merely the students, staff, employees, and visitors of King Abdulaziz University. The goals of this system are:

- To make the university a safe place for all.
- To make the evacuation plan clear. Guiding a large number of students during the calamity is not an easy task. There are so many blockage problems because they do not know how to act during the calamity.
- To enable the security employees to check and detect the safe places for students and view them on a KAU map.

Figure 1 presents the architecture of the entire system of iCalamityGuide. This system expresses flow and relation through systematic division such as user, interface, web server, and database server, modules for interfaces and service.

In calamity information management, geographic locations are important. Using their mobile phones, they can be located using the mobile network system or using an integrated GPS included in their phone.

GPS can:

- locate,
- calculate the distance traveled,
- record the user's path as a set of waypoints,
- navigate routes,
- work as a compass,
- indicate the elevation above sea level,
- provide the accurate time.

Xcode

Xcode is an Integrated Development Environment (IDE) containing a suite of software development tools developed by Apple for developing software for OS X and iOS".

It was released in 2003; the latest stable release is version 4.3.2. The main application of the suite called Xcode. The Xcode suite also includes most of Apple's developer documentation, and built-in Interface Builder, an application used to construct user graphical interfaces. It includes modified version of the GNU Compiler Collection as well as, in Xcode 3.1 and later.

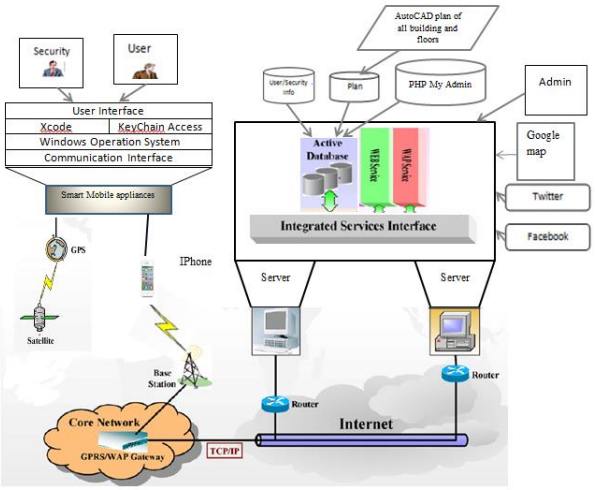


Fig. 1. iCalamityGuide Architecture

"Apple's LLVM-GCC compiler, with front ends of the GNU Compiler Collection and a code generator based on LLVM, and, in Xcode 3.2 and later, Apple's LLVM Compiler, with the clang front end and a code generator based on LLVM, and the Clang Static Analyzer. It supports C, C++, Objective-C, Objective-C++, Java, AppleScript, Python and Ruby source code with a variety of programming models, including but not limited to Cocoa, Carbon, and Java. Third parties have added support for GNU Pascal, Free Pascal, Ada, C#, Perl, and D. The Xcode suite used the GNU Debugger as the back-end for its debugger. As of version 4.2, the Apple LLVM Compiler became the default compiler. LLDB became the default debugger as of Xcode 4.3".

Keychain Access

Keychain Access is a Mac application that allows the user to access the Apple Keychain and configure its passwords for Websites, FTP servers, SSH accounts, network shares, wireless networks, groupware applications, encrypted disk images, etc. We used it to gain the keys required for obtaining certificates for Apple developer program to test the application on a real device, as well as distribute it through the Apple store after obtaining the approval of Apple Inc.

PHP My Admin

"Is a free and open source tool written in PHP intended to handle the administration of MySQL with the use of a Web browser? It can perform various tasks such as creating, modifying or deleting databases, tables, fields or rows; executing SQL statements, or managing users and permissions."[14]

Features:

"Features provided by the program include:

1) Web interface
2) MySQL database management
3) Import data from CSV and SQL.
4) Export data to various formats: CSV, SQL, XML, PDF (via the TCPDF library), ISO/IEC 26300 - Open Document Text and Spreadsheet, Word, Excel, LaTeX and others.
5) Administering multiple servers
6) Creating PDF graphics of the database layout
7) Creating complex queries using Query-by-example (QBE)
8) Searching globally in a database or a subset of it
9) Transforming stored data into any format using a set of predefined functions, like displaying BLOB-data as image or download-link.
10) Active query monitor (Processes)."[14]

AutoCAD

"Is a software application for CAD computer aided design and drafting, the software supports both 2D and 3D formats."[15]

The following section lists the functions used in our system.

- **Push notification:**

The application will consciously check if any broadcasts were newely added to the database, and will display if it found an Alert message. The user can then take two actions: either View, which will take them to the application, or cancel.

- **Guide me:**

This function describes acquiring guidance where to go. It will display a brief message saying, "Go to Building …." and the appropriate building based on what arranged from the safety center filled in the database.

This message itself will give two options. The Ok option will zoom out to the campus map as a whole and display the safe destination annotation with a bubble on the top with the tagged message pointing to the safe distention. The second option is Show Exits, which will provide exits for the current building and guidance for the way out in addition to the safe distention.

- **Share on Twitter:**

Connectivity to Twitter allows the User to share a picture or tweet so others can share their locations; any helpful information might help others in the same situation.

- **KAU:**

KAU is an action representing the zooming area that includes KAU Campus. Whenever a user goes away from that area, by this action he/she can get the view back to where it was at the beginning.

- **View Safe buildings:**

As a Security member, it can display all safe buildings users must go to and it helps to memorize them and keep up with any changes done.

- **Update Safe buildings:**

As a Security member, it can update the building's status. In the case of an emergency situation caused to any of the buildings and make them unsafe to stay in and need to be evacuated.

- **Detect users:**

For a Security member, the locations of users can be very helpful to rescue them or make sure they are at a safe building and away from danger, or react toward any crowded locations and take further action to make them safe.

When calamity occurs, the Admin starts the system by sending broadcasting warning news to security employees and users (students, staff, employs and visitors of KAU). Figure 2 (a) shows admin broadcasting an evacuation message to users received as an Alert.

Users respond to the system by push button "View". Then the system starts to detect the location of the user using the GPS location detection. This is followed by a message that represents the evacuation plan regarding that specific spot. Figure 2 shows the sequence in order of screens as viewed by a user. Figure 3 shows sequence in order of screens as viewed by security employees.

d) Implementation of DSRG Disaster Smart Road Guidance

This section will introduce the tools, technologies, and languages used to develop Disaster smart road guidance DSRG package.

Eb2a free Hosting [16]:

An Eb2a is a free hosting site that aids the user during building websites.

Types of panels used by Eb2a [17]:

CPanel (Vista panel): is an excellent hosting control panel on the net.

Fantastico: installs complex forums and website templates in an easy way.

Htaccess: our system uses the Htaccess features.

Eb2a features and offers [18]:

Eb2a servers use the latest versions of PHP, MySQL, and Apache Web Server.

Eclipse SDK platform:

To create our Android application, we use three key items in the project's root directory [19]:

- AndroidManifest.xml
- Category [20].
- A launcher icon [21].

Rule-based System: A Rule-based is used to collect information. Types of rule and knowledge in the system:

- Domain knowledge.
- Meta-knowledge.
- Common sense knowledge.
- Heuristic knowledge.
- Tacit knowledge.
- Explicit knowledge.

a) Admin sends broadcast message alert

b) Welcome page

c) iCG user registration

d) The first view for a user

e) After zooming in

f) the beginning of the guiding screen starts by tapping the Guide me button at top left corner

g) After choosing any of the KAU buildings iCG shows the selected floor (First-floor)

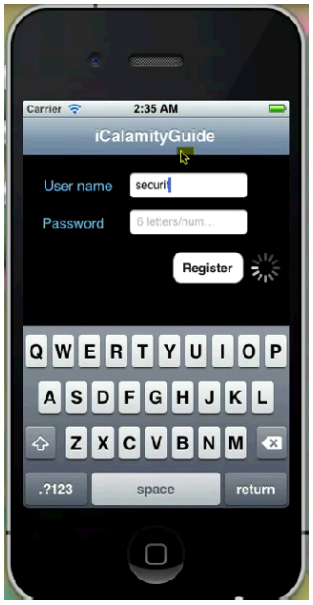
h) Second-floor

i) Third-floor

j) Show the user safe destination to go into wherever u goes the bubble will go with u

k) After twitter button has been chosen

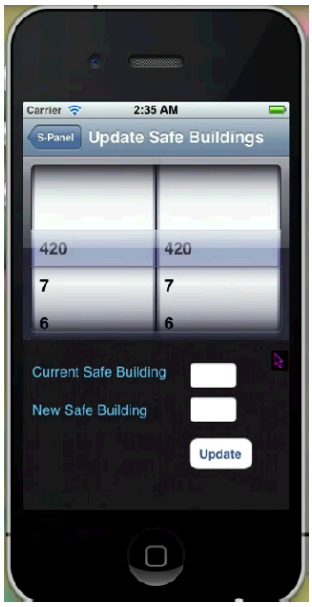
Fig. 2. Screen sequence User View



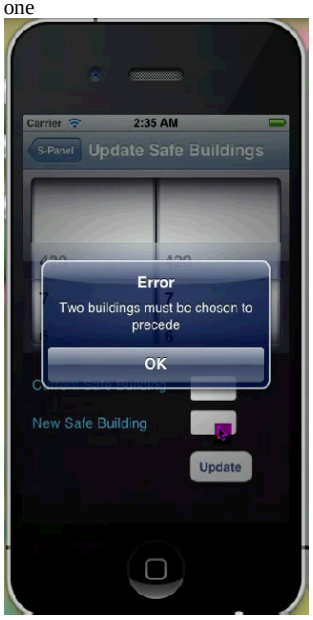
a) Enter the user name and password to recognize if the user is security or regular user

b) Welcome page

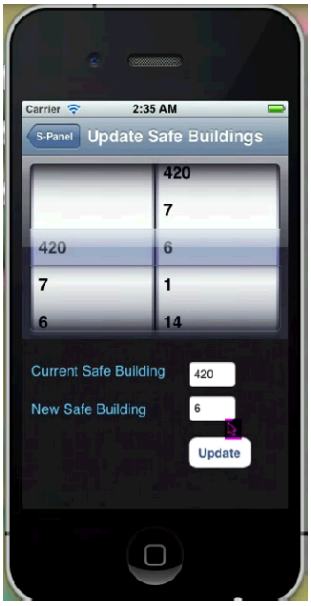
c) Security panel to list the three functions that security can apply one



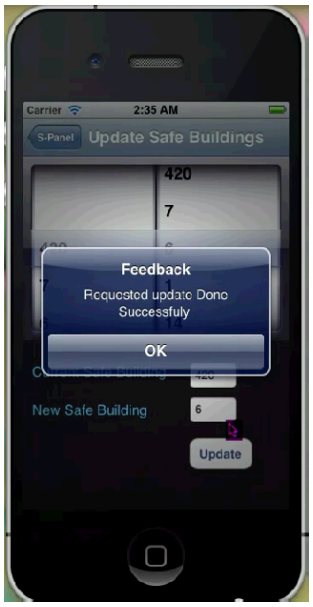
d) After tapping t first button (update the safe building) to replace the unsafe with safe



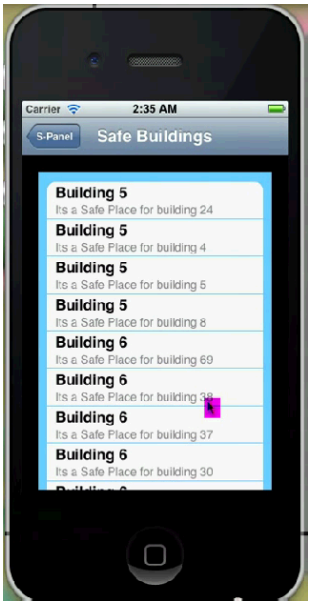
e) Appear if the text field was empty



f) The user can scroll the tow lists to fill out all



g) Feedback to tell u it has been successfully



h) Appear when the second button is chosen

i) The view when selecting the third button that views all pins in map representing all the registered users and their location in the same time

Fig. 3. Screen sequence Security View

Components for DSRG Application Implementation Process are described in Figure 4. As shown in Figure 5, the communication between server and Android is through the database. DSRG application operates on Android mobile phones. Figure **6** shows Android application structure.

Figure 7 shows Home Page and Figure 8 shows Map Screenshot of DSRG.

IV. Comparison of ICG, DSRG and Other Systems

Table 1 compares our systems with two other systems implemented on Android Operating System**.** iCG is the system used in the indoor environment. iCG and DSRG guide users have the highest reliability. iCG and DSRG use 3G and Wi-Fi in addition to the mobile network.

V. CONCLUSION

A disaster is an unexpected occurrence that happens anytime and anywhere. Most students and people are unaware about the prevention or safety responsiveness to face the disaster. When it happens, an evacuation process is conducted to save the victim. The Mobile platform Alerts Systems iCG and DSRG use the Rule-based system to detect evacuation places guidance mobile application, downloadable and ready to use by regular end users. It incorporates GPS technologies to location specific information and fed it to the citizens providing them with the nearest safe location and the shortest path to get there. It provides the user with evacuation tips depending on the current location and as provided from civil protection agency, weather status updates directly from the PME (Presidency of Meteorology and Environment) servers and map services.

This work also delivers real-time disaster and road constructions information and notification to the users such as fires, floods, constructions. Users receive warnings and tips on how to react before, during and after the disaster, directly on their mobile phones. Google map updates for any newly added information is received on the mobile user. In the time of a trouble, the user can benefit from the Rule-based system feature that will give him/her a feedback on what action should take in a specific situation. The Rule-based system was built through an interview with the experts domains. All these features, GPS, Google map, Wi-Fi, 3G and rule-based system, are improved access to the needed information at the needed time to save the life of people and students.

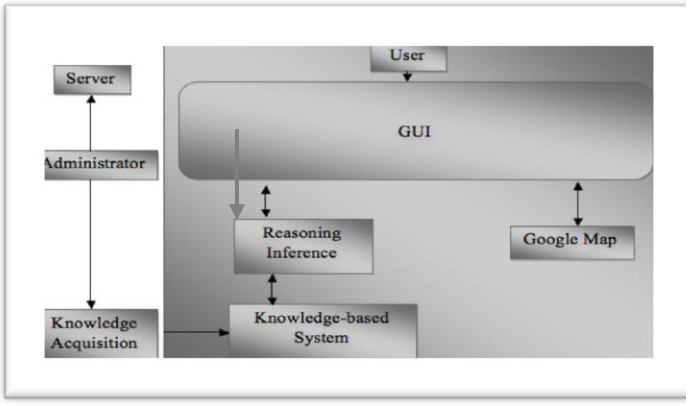


Fig. 4. Components for DSRG System

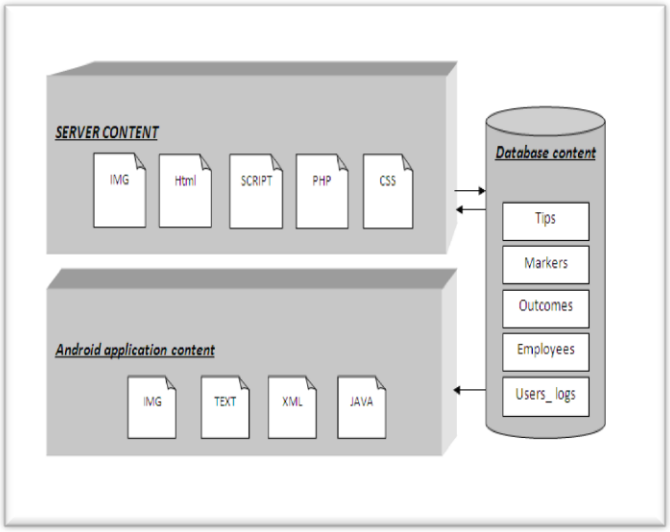


Fig. 5. DSRG Server and Android application content

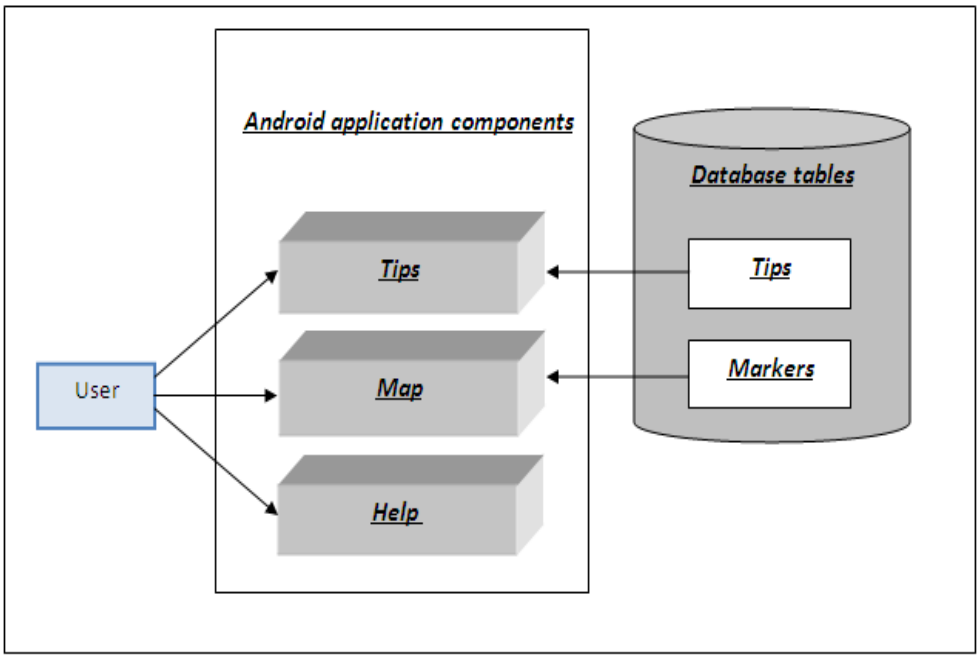


Fig. 6. DSRG Android application structures

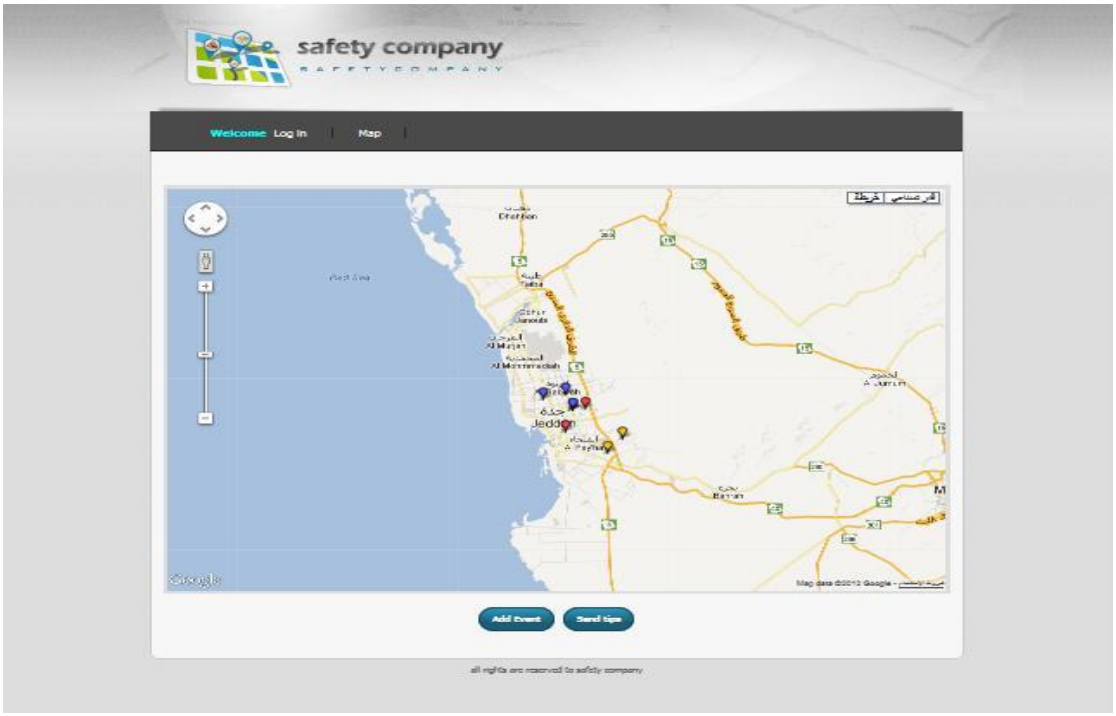


Fig. 7. DSRG Home Page

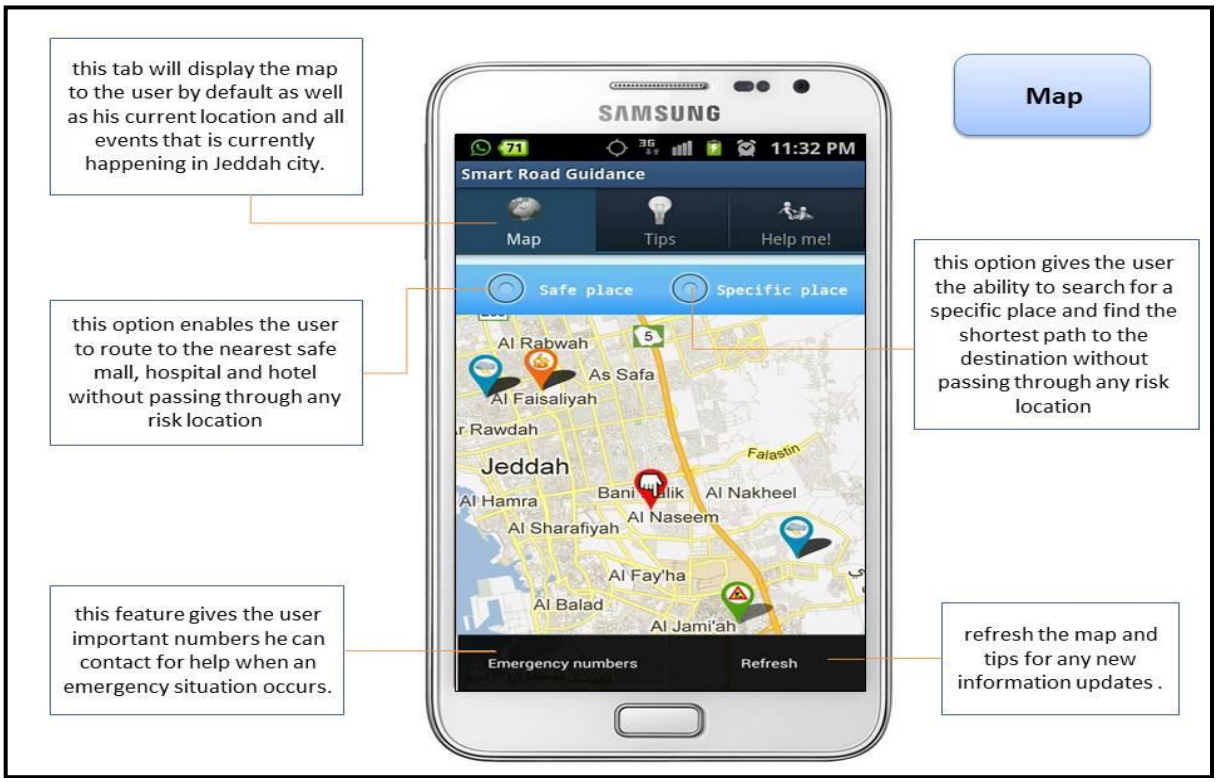


Fig. 8. DSRG Map Screenshot

ACKNOWLEDGMENT

This work was funded by the Deanship of Scientific Research (DSR), King Abdulaziz University, Jeddah, under grant no. (611-022-D1434). The authors, therefore, acknowledge with thanks, DSR technical and financial support.

REFERENCES

[1] Hyokyung Chang, Yongho Kang, Hyosik Ahn, Changbok Jang, Euiin Choi, "Context-aware Mobile Platform for Intellectual Disaster Alerts System", Energy Procedia, Volume 16, Part B, 2012, Pages 1318–1323.

[2] Luis E. de la Torre, Irina S. Dolinskaya and Karen R. Smilowitz, "Disaster relief routing: Integrating research and practice", Socio-Economic Planning Sciences, Volume 46, Issue 1, March 2012, Pages 88–97.

[3] Hsu-Yang Kung, Hao-Hsiang Ku, Che-I Wu, Ching-Yu Lin, "Intelligent and situation-aware pervasive system to support debris-flow disaster prediction and alerting in Taiwan", Journal of Network and Computer Applications Volume 31, Issue 1, January 2008, Pages 1–18.

[4] J.A. Bernard, "Use of a rule-based system for process control" IEEE Control Syst. Mag., Volume 8, Issue 5, October 1988, Pages 3-13.

[5] M. Nilashi, O. Ibrahim, "A model for detecting customer level intentions to purchase in B2C websites using TOPSIS and fuzzy logic rule-based system" Arab. J. Sci. Eng., 39 (3) (2013), pp. 1–16.

[6] G. Shobha, J. Gubbi, K.S. Raghavan, L.K. Kaushik, M. Palaniswami, "A novel fuzzy rule based system for assessment of ground water portability: a case study in South India", Magnesium (Mg), 30 (2013), pp. 35–41.

[7] Lamiaa F. Ibrahim, Reem Albatati, Samah Batweel, Rudainah Shilli, Mai Bakeer, Tsneem Abo Al Laban, "Safety of Natural Disasters", Design, User Experience, and Usability. User Experience in Novel Technological Environments, Book Series: Lecture Notes in Computer science, Volume 8014, 2013, pp 85-94, Springer Berlin Heidelberg.

[8] Nabil R. Adam, Basit Shafi q, Robin Staffin, "Spatial Computing and Social Media in the Context of Disaster Management", Intelligent Systems, IEEE, vol. 27, no. 6, pp. 90-96, Nov.-Dec. 2012 doi: 10.1109/MIS.2012.113.

[9] Gaia GPS. https://play.google.com/store/apps/details?id=com.trailbehind.android.gaiagps.pro&feature=search_result#?t=W251bGwsMSwxLDEsImNvbS50cmFpbGJlaGluZC5hbmRyb2lkLmdhaWFncHMucHJvIl0

[10] ubAlert- Disaster Alert Network. [Online]. http://itunes.apple.com/us/app/ubalert-disaster-alert-network/id455647397?mt=8

[11] Disaster Alert. [Online]. https://play.google.com/store/apps/details?id=disasterAlert.PDC

[12] Rehka Jadhav, Jwalant Patel, Darshan Jain, Suyash Phadhtare, "Emergency Management System Using Android Application", International Journal of Computer Science and Information Technologies, Vol. 5 (3) , 2014, 2803-2805.

[13] Jungki Lee; Niko, D.L.; Hyunsuk Hwang; ManGon Park; Changsoo Kim, "A GIS-based Design for a Smartphone Disaster Information Service Application," *Computers, Networks, Systems and Industrial Engineering (CNSI), 2011 First ACIS/JNU International Conference on* , vol., no., pp.338,341, 23-25 May 2011 doi: 10.1109/CNSI.2011.13

[14] (2016) PHPMYADMIN http://globaltechconsultants.org/?q=content/phpmyadmin

[15] Wikipedia (2012), The Free encyclopedia website, AutoCAD [Online] available from http://en.wikipedia.org/wiki/AutoCAD

[16] (2006) EB2A Internet Services. [Online]. https://www.eb2a.com/en/

[17] Sandip Patil. (2009, December) Android Project structure. [Online]. http://www.mobisoftinfotech.com/blog/android/android-project-structure/

[18] (2012, December) Intent Structure. [Online]. http://developer.android.com/reference/android/content/Intent.html

[19] (2012, December) Launcher Icons. [Online]. http://developer.android.com/guide/practices/ui_guidelines/icon_design_launcher.html

[20] (2012, December) Obtaining a Google Maps Android v1 API Key. [Online]. https://developers.google.com/maps/documentation/android/v1/mapkey?hl=ar-SA

[21] (2012, December) Knowledge base. [Online]. http://en.wikipedia.org/wiki/Knowledge_base

[22] Disaster Alert, (2015, Jan.), [online]. https://play.google.com/store/apps/details?id=disasterAlert.PDC

[23] Disaster Message Board Service, (2015, Jan.) [online]. https://www.nttdocomo.co.jp/english/info/disaster/disaster_board/

TABLE I. COMPARISON BETWEEN SYSTEMS

	iCG	DSRG	Disaster Alert[22]	Disaster Message Board [23]
Indoor	Yes	No	No	No
Guide user	Yes	Yes	No	No
Network type	3G or WIFI or Mobile network	3G or WIFI or Mobile network	Mobile network	Mobile network
Reliability	High	High	Low	Low
Update Information	Yes	Yes	Yes	No

Experimental Use of Kit-Build Concept Map System to Support Reading Comprehension of EFL in Comparing with Selective Underlining Strategy

Mohammad ALKHATEEB
Department of Information Engineering
Hiroshima University
Department of computer engineering
Tishreen University
Hiroshima, Japan

Taha RAJAB
Institute for effective Education
University of York
York,
UK

Yusuke HAYASHI
Department of Information Engineering
Hiroshima University
Hiroshima, Japan

Tsukasa HIRASHIMA
Department of Information Engineering
Hiroshima University
Hiroshima, Japan

***Abstract*—In this paper, we describe the effects of using Kit-Build concept mapping (KB-mapping) method as a technology-enhanced support for the Reading Comprehension (RC) in English as Foreign Language (EFL) contexts. RC is a process that helps learners to become a more effective and efficient reader. It is an intentional, active and interactive activity that language learners experience in their daily working activities. RC of EFL is a significant research area in technology-enhanced learning. In order to clarify the effect of KB-mapping method, we compared the results with that of selective underlining (SU) strategy through the Comprehension Test (CT) and the Delayed Comprehension Test (DCT) that performed two weeks later. As the results, it is clarified that there is a noticed difference in the DCT scores, while there is no significant difference in the CT scores. It indicates that the use of KB-mapping method helps learners to retain their information for longer period of time. By doing more statistical analysis for the results of the Kit- Build Conditions (KB-conditions) group and comparing them with the map scores, we found that the learners could answer 76% of the questions whose answers were included in their learner's maps. It was found that learners could recall 86% of the questions and that their answers were included in their learner's maps. It indicates that the use of KB-mapping method helps learners to retain and recall more information compared with the SU strategy even after two weeks elapsed. In a follow-up questionnaire after the end of all experiments, it was revealed that participants thought that using KB-mapping was similar to SU for the CT just after the use, but KB-mapping was more useful in remembering information after a while, and it was more difficult to carry out. Participants liked to use it in RC tasks, but asked for more time to do it.**

Keywords*—Technology-Enhanced Learning; Kit-Build Concept Map; Reading Comprehension; EFL

I. INTRODUCTION

Reading comprehension (RC) is one of the important learning activities, and it needs a special ability for learners to reap its benefits [1]. RC poses many challenges such as slow reading, insufficient vocabulary comprehension and bad recalling [HYPERLINK \l "Bro81" 2]. Researchers have always tried to support this learning activity by proposing methods or strategies. The main goal is to boost comprehension skills in the target subject area. When they are deployed in a language course, the main aims are to improve student RC skills and to contribute to the acquisition of the target language3].

The RC in an English as foreign language (EFL) context is a special case of RC; it is highly complex, dynamic, multi-componential and multi-dimensional task in the learning process. We can explain it as multiple interactions between the reader's background and knowledge in his or her Mother Language (ML) and that in English. Broadly speaking, the EFL RC is the same as the ML RC but it is slower and less successful than ML reading [HYPERLINK \l "And99" 4]. This can be explained in that the reading process is depending on many factors such as the level of reader's language proficiency, type of text, text difficulty and task demands.

In general, RC ability of EFL readers are not enough because of their low language proficiency, low awareness of RC strategies, limited vocabulary knowledge and potential comprehension failure5] [HYPERLINK \l "Pan08" 6]. So we can consider them as poor readers.

In our previous researches, we confirmed that KB-mapping method has the same effects of Scratch Build concept Mapping (SB-mapping) method in the CT, just after the use of method, but KB-mapping method has better effects in recalling after two weeks7]. There are many researchers have confirmed that SB-mapping method is a very effective supporting method for EFL RC [HYPERLINK \l "Man12" 8]9] [HYPERLINK \l "Sal13" 10]. Some researches have mentioned that SC-mapping method had two weak points. Firstly, poor readers lack the ability to construct knowledge and seldom utilize

learning strategies11] [HYPERLINK \l "Lun95" 12]. Secondly, poor readers have weaker meta-cognitive ability and have difficulties associated with not being aware of the reading process13] [HYPERLINK \l "Zen99" 14]. This means that SB-mapping method is not fit for EFL learners as poor readers. So, only comparing KB-mapping method with SB-mapping method is not enough. It is necessary to compare with reasonable method to support poor readers. Selective Underlining (SU) strategy is an authorized method that is useful for poor readers8] [HYPERLINK \l "Cla14" 15]. So in this paper, KB-mapping method is compared with SU strategy as a supportive method for EFL learners.

In this paper, "our hypothesis is that KB-mapping method has the same effects as SU strategy for the short term, and is more effective for the long term as shown in our previous researches based on SB-mapping method". In order to proof this hypothesis, we conducted an experiment to compare with SU strategy as the previous one.

In general, for long term learning process, the improvement of the ability is not easy to investigate, also the investigation process requires a special environment to keep the fare of the results, as an example to prove that there is no other method or strategy was used by the learner, during the learning time, is impossible, so in this research, we investigated the performance of the learners during the learning process16].

A. EFL Reading Comprehension

The Comprehension is defined as "the ability to understand something" in the Oxford Dictionary, the definition of Cambridge Dictionary is "the ability to understand completely and be familiar with a situation, facts, etc.", from these two definitions, we can define the RC in our research as "the learner's ability to understand completely and memorize the important information that is included in the text, that he is reading". Also RC is defined as the level of understanding of a text/message. This understanding comes from the interaction between the words that are written and how they trigger knowledge outside the text/message [HYPERLINK \l "And99" 4].

Fergus and Lockhart17] had proposed a theory that the RC involves two levels of processing, shallow (low-level) processing and deep (high-level) processing. This theory describes memory recall of stimuli as a function of the depth of mental processing. Deeper levels of analysis produce more elaborate, longer lasting, and stronger memory traces than shallow levels of analysis. Depth of processing falls on a shallow to deep continuum. Shallow processing (e.g., processing based on phonemic and orthographic components) leads to a fragile memory trace that is susceptible to rapid decay. Conversely, deep processing (e.g., semantic processing) results in a more durable memory trace. Many researches have used the memory recalling as indictor to measure the level of comprehension, and they have proposed the delayed comprehension test for the depth of comprehension [HYPERLINK \l "Par81" 18]19]. As an example, Mayer and Bluth [HYPERLINK \l "Mey80" 20] have proposed to check the recalling immediately, and one week later. Due to this theory, we measured the level of RC by using two kinds of test, the Comprehending Test (CT), which was done just after the practical use, to check the comprehended information from the text (Shallow processing), and the Delayed Comprehension Test (DCT), which was done two weeks after, in order to measure the recalled information (Deep processing).

In general, the RC is a very difficult task for learners in all the stages of study especially when they are reading text in a foreign language. Thus, the EFL reading is one of the most common research topics in the learning field.

Many researchers have tried to solve this problem by proposing sstrategies to help the EFL learners in the RC task; SU strategy15] [HYPERLINK \l "Pio051" 21], Note-Taking Skills Reading21], SQ3R (Survey, Question, Read, Recite, Review) [HYPERLINK \l "Hub04" 22] and PORPE (Predict, Organize, Rehearse, Practice, Evaluate)23]. However, most of them only had slight improvements in the EFL learners' comprehension just after the use. Several investigations of RC strategies have specifically addressed challenges related to reading expository text. Positive outcomes have been found for learners who were taught strategies to help learners identify main ideas [HYPERLINK \l "Pan08" 6].

B. Selective Underlining Strategy

SU strategy is one of the most important strategies commonly used in classroom. Generally, it is used to help learners to organize the text, which they are reading by selecting the important sentences. This strategy teaches learners to highlight/underline only the key words, phrases, vocabulary, and ideas that are essential to understanding the text15] [HYPERLINK \l "Hub04" 22]. It is very useful for comprehending the text, because it is a flexible strategy that may be tailored to fit various types of information and different skill-levels. This strategy can also be integrated with the use of technology and electronic information such as eBooks. As students study, it helps them learn to pay attention to the essential information within a text15]. In general, this strategy is focusing on the vocabularies and comprehension of the text during the reading time, also it helps the learners to identify the important points in text, helps them to pay close attention to what they are reading and also, allow greater learning and deeper comprehension.

Typically, in the SU strategy, the learner starts by reading the text to understand the main content of the text, after that, the learner rereads and begins to underlining the main ideas and their supporting details, then he selects the important facts and the key vocabularies [HYPERLINK \l "Cla14" 15].

By using the underlined part of the text, the learner can give a summary for the important information in the text that he has read. This strategy demands the learner to capture the main ideas, key concepts and details; also it helps learning by reducing the needed information in text, so it reduces the learning time and in the same time strengthens the RC.

As SU strategy is widely used in usual classroom, and it is suitable strategy to support the poor readers, as the learners of EFL RC. So we planned to compare KB-mapping method with SU strategy in order to confirm the usability of KB-mapping method for supporting EFL poor reader too.

II. KB-MAPPING METHOD

We have been developing learning tools to help both the students and the teachers in the learning process. One of these tools is KB-Map24] [HYPERLINK \l "Yam10" 25]26], and we have found the following results; (1) this tool is very useful for learning different topics of sciences in learner's mother languages [HYPERLINK \l "Sug12" 27]28], and (2) this method has the same effect of SB-mapping method just after the building task, but it has better effects in recalling after a while [HYPERLINK \l "Alk15" 7]16].

A. Overview of KB-map (Kit-Build Concept Map)

As a definition of KB-map we can find "a framework to realize automatic diagnosis of concept maps built by learners and to give feedback to their errors in the maps" [HYPERLINK \l "Yam10" 25]. KB-map is a special kind of concept map. Generally, the concept map creation consists of two steps; (1) the extraction of the concepts and the relations from a specified text, and (2) the making-connection between two extracted concepts by using the extracted relation. In KB-mapping method, the supervisor of learners performs the first step to create the goal map, and generates a kit by dividing the goal map. Then, the learner is asked to make connections by using the kit until he or she connects all of components in the kit.

B. KB-mapping System

We have already developed a system based on the KB-map explained in the previous section. This system is called "KB-mapping System". It is a web-based application with two client systems: "KB-map Editor" and "KB-map Analyzer", and a server system "KB-map DB". KB-map Editor provides an environment for teacher, or supervisor, to make a goal map, a kit, and for learner to make a learner's map. This system has been implemented by Java (version 1.6). KB-map Analyzer has functions to gather learner's maps online, generate a group map and diagnose the maps. This system has implemented by Flash and is used with Flash Player 10. KB-map DB has a function to store and share maps. This system has been developed by Ruby (version 1.8.7) on Rails (version 1.2.3) and MySQL (version 5.1.30)25].

III. EXPERIMENT METHODOLOGY

We investigate the effects of the KB-mapping method as a supportive tool for RC task using the CT for the short term and the DCT for the long term described in Section X, by comparing with those of the SU strategy under the same conditions. In this section, we describe the following four points of experiment; (1) the subjects, (2) the procedure of RC task, (3) the learning materials and (4) the experimental system.

A. The subjects

The subjects of our experiment were eight Japanese students of third grade of information engineering faculty. Their scores of TOEIC exam vary from 430~625, so they had different reading abilities; we prepared an aptitude test to check their reading abilities, this test was a simple reading test in the same level of the used in all the sessions. By using the information of their TOEIC records and aptitude test results, we grouped them into two groups A and B, which are almost commensurate with the reading ability as shown in Table I.

TABLE I. PARTICIPANTS GROUPING

	Group A	Group B
TOEIC (SD)	513.75(75)	537.5 (71)
Aptitude Test	70	65

B. Procedure of Experiment

We performed this experiment in six sessions as RC task for six different English texts; At the beginning of the first session, we gave instructions for subjects about the procedure of experiments including how to use KB-mapping system. As for the rest of sessions, we started with the DCT of the previous session, followed by a learning activity for thirty minutes in which subjects tried to improve their English level. At the end of the sixth session, we asked subject using questionnaire described later.

C. Procedure of One Session:

In our experimental use, we compared the effects of using KB-mapping method and the effects of SU strategy in the RC process. We were measuring these effects from two points of view, just after using and two weeks later.

To eliminate the effect of other supporting strategies, we designed our experiment learning activity under strict and limited time constrains. The process of one session, as shown in Table II, consists of 4 steps: Firstly, within 10 minutes, the both conditions groups were requested to comprehend the whole text by reading it generally and translating the difficult words in the text by using dictionary, then in the next 10 minutes the KB-conditions group was required to build the KB-map of the text by using KB-map editor and in the same time the SU-conditions group was required to do the underlining of the important sentences in the text. After that both groups did the CT within 5 minutes, and finally, after two weeks both groups did the CT again as a DCT.

TABLE II. PROCEDURE OF ONE SESSION

time	KB-conditions	SU-conditions
10 min	Reading materials (using dictionary is allowed)	
10 min	Making the KB-map by using KB-map editor	Underlining the important parts of the text
5 min	Doing Comprehension Test	
5 min	Doing DCT(2 weeks after)	

IV. EXPERIMENTAL USE

This experiment was done in 6 sessions with two groups of participants (A, B) both of them had almost the same reading ability. For each group they used the KB-mapping method for 3 sessions and the SU strategy for 3 sessions too. To conduct this experiment we used computers with Intel core i3-3240 processor, 4 GB of RAM and 20″ monitor. The used platform was windows 7. In this chapter we introduced 3 points: the preparation of the used materials, example of the real experimental use and example of the materials that used in the experiment.

A. Materials Preparation

The participants of this experiment were information engineering students, they were interested in the topics of information engineering, so firstly we selected 6 text from Wikipedia in the information engineering field, and checked their grammatical and semantic structure. We created the corresponding KB-maps (Goal maps) that covered the main concepts and relations of the texts, also we did the SU for the important and essential phrases in the texts. Then we prepared the CTs. Finally, we checked all of the material again to make sure that materials did not contain any error. In addition, we checked if the answers of questions of CT were included or not in the concepts extracted by the KB-mapping method and the SU strategy, to marking the questions that not covered.

B. Example of the experimental use

To investigate the effects of using KB-mapping method to improve the EFL CT, For the first session, we started with a learning activity to improve the English level of student, and in the last 30 minutes we did our experiment, firstly, within ten minutes, the learners were requested to read the text to get the main idea of it and to translate the unknown words for them by using a web dictionary. After that within 10 minutes too, the KB-conditions group tried to build the learner's map by using the KB-map editor, in the same time, the SU-conditions group were doing the SU of the important phrases of the text. In the next 5 minutes they answered the CT. Finally we collected all the materials about this session (text, notes and test papers). Two weeks after, the participants did the DCT. Figure2 shows the process of the first session.

For the other sessions, they did the DCT of the previous session, as we mentioned before, and after that, they started the new session as explained in the previous paragraph.

C. Example of experimental material

In this section, we introduce one example of the materials that was used in the first session of our experiment. Fig. 1 shows a part of the text that was used as the original text, which learners tried to comprehend it. After that, the KB-conditions group tried to build learner's map, within 10 minutes, by using the kit that provided by the system, which generated from the corresponding goal map. Fig. 3 shows the goal map of the text that was prepared by the supervisor. It contains most of the information of the original text, this goal map was divided by the system to generate the kit that shown in Fig. 4. Fig. 5 shows one example of the learner's map which was built by one of the learners.

In the same time, the SU-conditions group tried to do the SU for the important sentences in the text within 10 minutes too. The underlined text contains the important information of the text. Fig. 2 shows one example of the underlining of the same paragraph.

> "A general purpose computer has four main components: the arithmetic logic unit (ALU), the control unit, the memory, and the input and output devices. These parts are interconnected by busses, often made of groups of wires.
>
> Inside each of these parts are thousands to trillions of small electrical circuits which can be turned off or on by means of an electronic switch. The circuits are arranged in logic gates so that one or more of the circuits may control the state of one or more of the other circuits."

Fig. 1. Example, Part of First Session's Text

> "A general purpose computer has four main components: the arithmetic logic unit (ALU), the control unit, the memory, and the input and output devices. These parts are interconnected by busses, often made of groups of wires.
>
> Inside each of these parts are thousands to trillions of small electrical circuits which can be turned off or on by means of an electronic switch. The circuits are arranged in logic gates so that one or more of the circuits may control the state of one or more of the other circuits."

Fig. 2. Example, First Session's Underlined Text

After that all the learners did the same CT, that lasted for 5 minutes and which was a set of multi-choices questions scored from 100. All of these questions were asking about information included in the original text, some of them were asking about information included in the goal map. For the underlined text all of them were included. Fig. 6 shows a part of the CT of this session.

By the end of this test, they have finished the experimental use of that day and 2 weeks later, they did the comprehension test again as a DCT.

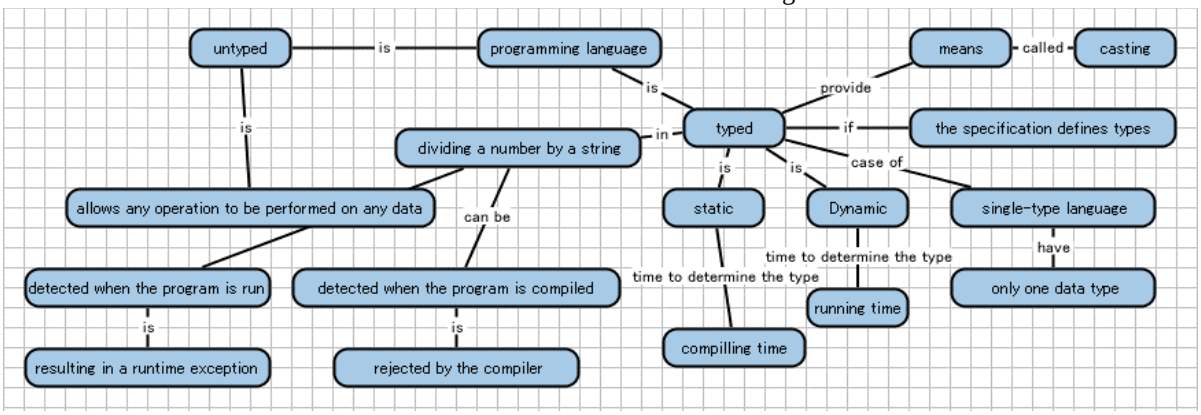


Fig. 3. The Goal Map of First Session

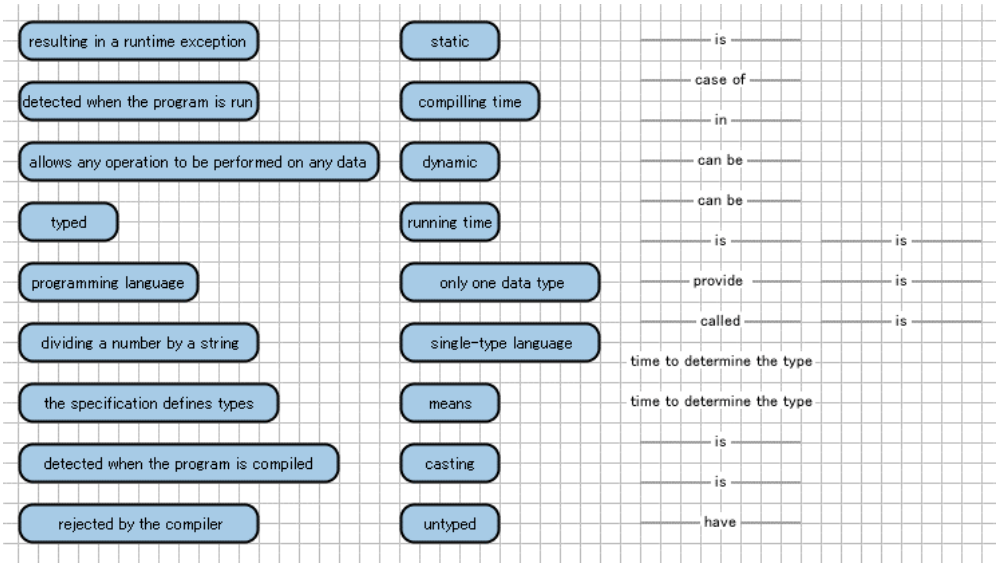


Fig. 4. The Kit of firest session

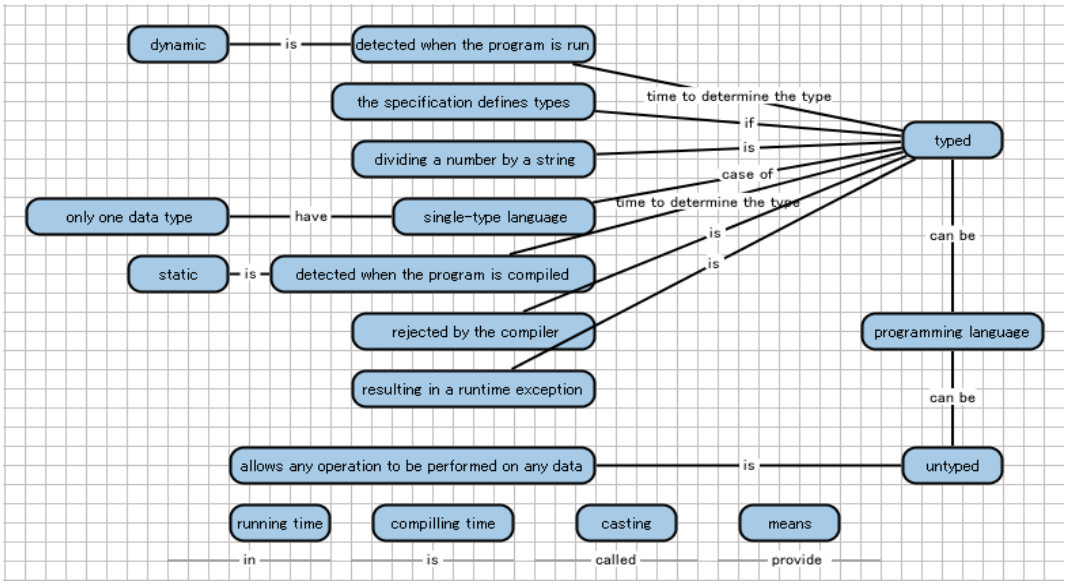


Fig. 5. Example of the learner's map

Q3: What is the component of busses?

a. Electrical circuits b. Logical gates

c. Wires d. Wires and logical gates

Q4: What are inside the control unit?

a. Electrical circuits b. Logical gates

c. Wires d. Electrical circuits and logical gates

Q5: what can the electric circuits control?

a. Electrical circuits b. Logical gates

c. Wires d. The control Unite

Fig. 6. Sample of the questions in comprehension test

V. RESULTS

We performed our experiment with 8 students in 6 sessions. In the first session, we had 4 participants as the KB-conditions (Group A) and 4 as the SU-conditions (Group B), then the participants were shifted to the alternate conditions in the next and so on. The groups are made by balancing in the scores of TOEIC as shown in Table1. We had 6 sessions with different 6 texts. Goal maps and tests are prepared for each text. Table III summarizes the details of every session.

TABLE III. DETAILS OF EVERY SESSION

Session No.	**1**	**2**	**3**	**4**	**5**	**6**
KB-Conditions	Group A	B	A	B	A	B
SU-Conditions	Group B	A	B	A	B	A
Text	T1	T2	T3	T4	T5	T6

A. Effects of Using KB-mapping Method in CT and DCT

For every session, we compared the CT average scores, the DCT average scores and the differences (DCT-CT) for the two conditions group. The scores are shown in Table IV. We found that the KB-conditions have a better CT average score in one session and nearly worse average in the remaining five. For the DCT, we found that the KB-conditions have a better DCT average score in five sessions and nearly worse average scores in the remaining one. And for the difference DCT-CT (Diff), we found that the KB-conditions have a higher average in all the sessions.

TABLE IV. AVERAGE COMPREHENSION TEST & DELAYED COMPREHENSION TEST SCORES FOR THE TWO CONDITIONS

Session No.	1	2	3	4	5	6
KB-CT	80	42.5	50	62	60	80
SU-CT	75	80	41	87	65	90
KB-DCT	90	55	65	58	65	70
SU-DCT	65	60	45	45	35	55

In the remaining one case where KB-condition is worse than SU-condition in DCT, we found that the learners failed to complete a concrete learner's maps in comparison with other cases. As they could only get about 37% as a map score, in comparison with the other session average was 74%, because of the complexity of goal map (map size was big) and the limitation of time (10 minutes only for building). In general when learners failed to build learners' map they could not comprehend the whole text as structured form. This may be case a misunderstanding for the information.

In this experiment, the two groups of subjects are balanced by their TOEIC scores and aptitude test scores, both groups had followed KB-conditions for three times and SU conditions for three times too. The process of KB-conditions and SU-conditions were same in all phrases except the phrase of building learner's map and doing the SU for the text, which had the same time to do. We analyzed the data by using win-lose-tie statistical analyze to show which is better among the CT, DCT and the difference (Diff=DCT-CT). Table V is the summary of the win-lose data.

TABLE V. WIN-LOSE TABLE AND BINOMINAL TEST OF THE KB-CONDITIONS SCORES

Session No.	1	2	3	4	5	6	∑	*P*(value)
CT	1	0	1	0	0	0	2	0.234
DCT	1	0	1	1	1	1	5	0.094
Diff	1	1	1	1	1	1	6	0.016

To evaluate the results of the win-lose-tie table, we used the binominal test to calculate the probability mass function of the number of KB-mapping method winnings in all the sessions. By using the binominal test, We found that the KB-conditions won in the CT for 2 of six sessions (2 wins +4 lose); the probability that there were two successes in six trials was (p(2/6)=0.234>0.05). This result indicated that the KB-mapping method did not show better effects in the CT just after being used. For the DCT, we found that the KB-conditions won in five of six sessions (6 wins); the probability that there were 5 successes in six trials was (p(5/6)= 0.094>0.05); This was slightly significantly different (p<0.10); this result indicated that the KB-mapping method showed better effects in the DCT two weeks after being used. Also for the differences (DCT-CT), we found that the KB-conditions won in six of six sessions (six wins), and the probability that there were six successes was (p(6/6)= 0.016<0.05); this probability was significantly different. This result indicated that the KB-mapping method had better effects on recalling the comprehended information 2 weeks after being used.

B. Effects of Using KB-mapping Method in CT and DCT

To investigate the effects of using KB-mapping method, we did more detailed analysis about the included questions in learner's map of our experiment.

The included questions were the ones that can be answered by using the learner's map, as example, in Fig. 6 the questions 3 and 4 were included in the learner's map that shown in Fig. 5, they could be answered by using the components of the learner's map.

For the CT, we calculated the score of the included questions, that their answers were correct, in the learner's map and compared them with the total number of the included questions in the learner's map. We found that the KB-conditions learners could answer 76% of the questions that their answers were included in their learner's maps.

$$Av\left(\frac{\text{correct answer in comprehension test included}}{\text{included questions in learner's map}}\right) * 100 = 75.625 \quad (1)$$

In (1), we calculated the number of the correct answered questions in the CT, which their answers were included in the learner's map and compared them with the total number of questions, which their answers were included in the learner's map, finally we calculate the average of all learners.

For the DCT, we calculated the score of the included questions in the learner's map and compared them with the total number of the included questions. We found that the KB-conditions learners could remember 86% of the questions that their answers were included in their learner's maps.

$$Av(\frac{\text{correct answer in delayed comprehension test included}}{\text{included questions in learner's map}}) * 100 = 85.625 \quad (2)$$

In (2), we calculated the number of the correct answered questions in the DCT, which their answers were included in the learner's map and compared them with the total number of questions which their answers were included in the learner's map, finally we calculated the average.

From the (1)&(2), we could say that the using of KB-mapping method helped the learners to use and recall, after two weeks, most of the information that was included in their learner's map.

C. Effects of Text Without KB-map in CT and DCT

To ensure the reliability of our findings about the effects of using KB-mapping method, by checking the effects of the other parts of the text that were not included in the learner's map, we did more detailed analysis about the questions that

were not included in learner's map called as "not included questions".

The "not included questions" were the questions that cannot be answered by using the learner's map, as example, in Fig. 6 the questions 5 is not an included questions in the learner's map is shown in Fig. 5, they could not be answered by using the components of the learner's map.

For the CT, we calculated the score of the" not included questions" that their answers were correct, in the learner's map and compared them with the total number of the "not included questions". We found that the KB-conditions learners could answer only 45% of the questions that were not included in their learner's maps.

$$\mathrm{Av}\left(\frac{\text{correct answer in comprehension test not included}}{\text{not included questions in learner's map}}\right) * 100 = 45.069 \quad (3)$$

In (3), we calculated the number of the correct answered questions in the CT which their answers were not included in the learner's map and compared them with the total number of questions which their answers are not included in the learner's map", finally we calculate the average.

For the DCT, we calculated the score of the "not included questions" in the learner's map and compared them with the total number of the not included question. We found that the KB-conditions learners could remember only 42% of the questions that were not included in their learner's maps.

$$\mathrm{Av}\left(\frac{\text{correct answer in delayed CT not included}}{\text{not included questions in learner's map}}\right) * 100 = 42.0139 \quad (4)$$

In (4), we calculated the number of the correct answered questions in the DCT which their answers were not included in the learner's map and compare them with the total number of questions which their answers were not included in the learner's map, finally we calculated the average.

From the (3)&(4), we could say that the effects of the text without KB-mapping method was not so helpful for the learners to answer or to remember the information that not included in their learner's maps.

As a summarization for section 5.3 and 5.4, we found that if the learner could build a good learner's map, he would get a good score in the CT and he would get a good score in the DCT. In other words, the learner's map could be used to evaluate the learner's comprehension.

D. Questionnaire

After we had finished the last session's DCT, the learners answered the questionnaire to evaluate the learning method of using KB-mapping method and compare it with the SU strategy. Table VI shows the results of this questionnaire.

The questions (1-6&9) were multiple choice questions with 5 options that measure the participants agreement with the mentioned point by the question. The choices that used were "A. Strongly agree, B. Agree, C. Natural, D. Disagree and E. Strongly Disagree", and the questions (7&8) were multiple choice questions with 3 options, that compare between the two learning methods. The choices that used were "A. SU, B. Same, C. KB-mapping"; to normalize the results of this questionnaire, we tried to summarize all the results of our questionnaire and convert them to arithmetical form that means (1 Strongly agree, 0.5 Agree, 0 Natural, -0.5 %Disagree, and -1 strongly Disagree) and in the same time. It is means (1 Map building, 0 same and -1 underlining). As a summarization of the questionnaire evaluation, 0 means the normal, the positive means agreement and the negative means disagree and the value shows the strength of the agreement or the disagreement.

From questions (1&4&7)(2&5)(3&6),the learners found out that the KB-mapping method was useful to understand English text as the SU strategy and also useful to answer the CT just after the learning activity, but they thought that KB-mapping method was more useful to answer the DCT two weeks after, Also from questions (8,9) they thought that the KB-mapping method was more difficult to carry out, but they liked to use KB-mapping method in RC task but they mentioned that they need more time to do it.

TABLE VI. EVALUATION OF KB-MAPPING METHOD FOR EFL READING COMPREHENSION BY COMPARING WITH SU STRATEGY

Explanation	Average Agreement
Q1. Do you think that SU strategy was useful to understand English text?	0.1875
Q2. Do you think that SU strategy was useful to answer the test after reading?	0.1875
Q3. Do you think that SU strategy was useful to answer test two week later?	-0.375
Q4. Do you think that KB-mapping method was useful to understand English text?	0.5
Q5. Do you think that KB-mapping method was useful to answer the test after reading?	0.375
Q6. Do you think that KB-mapping method was useful to answer test two week later?	0.125
Q7. Do you think that KB-mapping method was more useful to understand English text?	0.625
Q8. Do you think that KB-mapping method was more difficult to carry out?	0.75
Q9. Did you like to use KB-mapping method to understand English text?	0.0625

VI. CONSIDERATION

We found that the KB-mapping method user could retain the knowledge as the underlined method users, but they are more effective in the recalling. This result is harmonized with principle of structured storing of knowledge in memory, which proposed by cognitive psychology research. It has shown that the knowledge in the memory is stored in a structured form that determines the ability to retain, recall and use it to solve problems[HYPERLINK \l "JMi01" 29]30][HYPERLINK \l "Ohn11" 31]. For KB-mapping method, the learner tried to build the KB-map, which is a structured form of the knowledge, by using the provided information from the original text.

In general, the use of KB-mapping method needs concentration in reading the text and needs to read with attention to distinguish the two concepts, which can be related,

and to find the corresponding relation, which can connect them together. In the same time, this process requires the learner to understand the information in the text deeply and required him to comprehend the text in whole. So we can explain our result by the required high load on the learner memory to comprehend deeply the whole text to complete the learner's map, this load forces the learner's memory to keep most of the information that he has already comprehended.

When constructing a KB-map, the focus is on the relationships among concepts, because the learner does not required to think about distinguishing the concepts that are already provided by the system, so his entire constraining is on the understanding of the whole text.

VII. Conclusion and Future Work

In this paper, we describe the effects of using KB-mapping method as a supportive tool for the RC of English texts as EFL reading. We conducted an experiment composed of six experimental RC sessions. Overall, from this experiment we can say that the use of KB-mapping as learning supportive tool for RC is good as SU strategy in the short term, but it is so better for the long term.

Our next step goal is to design a new experiment to compare our KB-mapping method with the Scratch Build Concept mapping method for RC supporting to investigate in deeply why KB-mapping method is more useful for recalling after a while.

References

[1] M. F. Graves, F. Michael, J. Connie, and B. B. Graves, Teaching reading in the 21st century. Order Processing, Des Moines, 1998.

[2] A. L. Brown, J. C. Campione, and J. D. Jeanne, "Learning to learn: on training students to learn from texts", Journal. of Educational researcher , pp. 14-21, 1981.

[3] K. Rayner, B. R. Foorman, C. A. Perfetti, D. Pesetsky and M. S. Seidenberg, "how psychological science informs the teaching of reading", Journal. of Psychological Science in the Public Interest, Vol. 2, No 2, pp. 31–74, 2001.

[4] N. J. Anderson, and X. Cheng, Exploring second language reading: Issues and strategies, MA: Heinle & Heinle, Boston 1999.

[5] M. Zoghi, Masoud, M. Ramlee, and T. N. R. B. T. Mohamad "getting to know L2 poor comprehenders", Journal. of English Language Teaching, vol. 4, no. 1, pp. 98, 2011.

[6] J. Pang, "Research on good and poor reader characteristics: implications for L2 reading research in china", Journal of Reading in a Foreign Language, vol. 20, no.1, pp. 1-18, 2008.

[7] M. Alkhateeb, Y. Hayashi, T. Rajab, and T. Hirashima, "Comparison between kit-build and scratch-build concept mapping methods in supporting EFL reading comprehension", Journal of Information and Systems in Education, vol. 14, no.1, pp. 13-27, 2015.

[8] P. Manoli, and M. Papadopoulou.: "Graphic organizers as a reading strategy: research findings and issues", Journal of Creative Education,vol. 3, no.03, pp. 348 -356, 2012.

[9] M. Kalhor, and G. Shakibaei, "Teaching reading comprehension through concept map", Journal of Life Science Journal, vol. 9, no. 4, pp. 725-731, 2012.

[10] A. D. Salehi, S. Jahandar, and M. Khodabandehlou, "The impact of concept mapping on EFL student's reading comprehension", Journal of Indian J. of Fundamental and Applied Life Sciences,vol. 3, pp. 241-250, 2013.

[11] D. P. Clark, "first, poor readers lack the ability to construct", Journal of Dissertation Abstracts International, vol. 46, no. 8, pp. 22-46, 1985.

[12] I. Lundberg, "The computer as a tool for remediation in the education of students with reading disabilities: A theory based approach", Journal of Learning Disabilities Quarterly, vol. 18, pp. 89-100, 1995.

[13] R. Wagner, and J. K. Torgesen, "The nature of phonological processing and its causal role in the acquisition of reading skill", Journal. of Psychological Bulletin, vol. 101, pp. 192-212, 1987.

[14] S. J. Zeng, "Research of low achievement schoolchild in their work memory, the phonetic handling ability and the speed of reading", Proc. of the diagnosis of schoolchild's reading difficulty, pp. 5-28, taiwan1999.

[15] Schnell, R. Thomas, and J. R. Daniel, "a comparison of underlying strategies for improving reading comprehension and retention", Journal of Reading Horizons, vol. 18, no 2, pp. 106-109, 1978.

[16] M. Alkhateeb, Y. Hayashi, and T. Hirashima, "the effects of using kit-build method to support reading comprehension of EFL" Proc. of Human Interface and the Management of Information. Information and Knowledge in Applications and Services, Springer International Publishing, pp. 3-11, 2014.

[17] F. I. Craik, and R. S. Lockhart, "Levels of processing: A framework for memory research", Journal of verbal learning and verbal behavior, vol. 11, no 6, pp. 671-684, 1972.

[18] S. G. Paris, and M. Myers, "Comprehension monitoring, memory, and study strategies of good and poor readers", Journal of Literacy Research, vol. 13, no 1, pp. 5-22, 1981.

[19] B. Laufer, "Vocabulary acquisition in a second language: Do learners really acquire most vocabulary by reading? Some empirical evidence", Journal of Canadian modern language review, vol. 59, no 4, pp. 567-587, 2003.

[20] B. J. Meyer, D. M. Brandt, and G. J. Bluth, "Use of top-level structure in text: key for reading comprehension of ninth-grade students", Journal of Reading research quarterly, pp. 72-103, 1980.

[21] A. Piolat., T. Olive. and T. Ronald "Cognitive effort during note taking", Journal of Applied Cognitive Psychology, vol. 19, no 3, pp. 291–312, 2005.

[22] J. A. Huber "A closer look at SQ3R", Journal of Reading Improvement vol. 4, no. 2, pp. 108-112, 2004.

[23] B. L. Ngovo, "Study strategies for narrative texts: PORPE and annotation", Journal of Developmental Education, vol. 23 no. 2, pp. 24-28, 1999.

[24] H. Funaoi, K. Ishida, and T. Hirashima, "comparison of kit-build and scratch-build concept mapping methods on memory retention", Proc. of ICCE 2012 , pp. 539-546, 2012.

[25] K. Yamasaki, H. Fukuda, T. Hirashima, and H. Funaoi, "Kit-Build concept map and its preliminary evaluation", Proc. of the 18th International Conference on Computers in Education , pp. 290-294, 2010.

[26] T. Hirashima, K. Yamasaki, and H. Fukuda, "Framework of kit-build concept map for automatic diagnosis and its preliminary use", Journal. of Research and Practice in Technology Enhanced Learning, vol. 10, no. 1, pp. 1-21, 2015.

[27] K. Sugihara, T. Osada, S. Nakata, H. Funaoi, and T. Hirashima, "Experimental evaluation of kit-build concept map for science classes in an elementary school", Proc. of ICCE 2012 , pp. 17-24, 2012.

[28] K. Yoshida, K. Sugihara, Y. Nino, M. Shida,and T. Hirashima, "Practical use of kit-build concept map system for formative assessment of learners", Proc. of ICCE 2013 , pp. 906-913, 2013.

[29] J. Michael, "In pursuit of meaningful learning", Journal of Advances in physiology education, pp. 145-158, 2001.

[30] R. C. Atkinson, and R. M. Shiffrin, "Human memory: A proposed system and its control processes", Journal of The psychology of learning and motivation, vol 2, pp. 89-195, 1968.

[31] T. Ohno, T. Hasegawa, T. Tsuruoka, K. Terabe, J. K. Gimzewski, and M. Aono, "Short-term plasticity and long-term potentiation mimicked in single inorganic synapses", Journal of Nature materials, vol.10, no. 8, pp. 591-595, 2011.

Regression Test-Selection Technique Using Component Model Based Modification: Code to Test Traceability

Ahmad A. Saifan
Department of Computer Information Systems
Yarmouk University
Irbid, Jordan

Iyad Alazzam
Department of Computer Information Systems
Yarmouk University
Irbid, Jordan

Mohammed Akour
Department of Computer Information Systems
Yarmouk University
Irbid, Jordan

Feras Hanandeh
The Hashemite University
Zarqa, Jordan

***Abstract*—Regression testing is a safeguarding procedure to validate and verify adapted software, and guarantee that no errors have emerged. However, regression testing is very costly when testers need to re-execute all the test cases against the modified software. This paper proposes a new approach in regression test selection domain. The approach is based on meta-models (test models and structured models) to decrease the number of test cases to be used in the regression testing process. The approach has been evaluated using three Java applications. To measure the effectiveness of the proposed approach, we compare the results using the re-test to all approaches. The results have shown that our approach reduces the size of test suite without negative impact on the effectiveness of the fault detection.**

Keywords—Regression Test; Regression Test selection technique; Meta-Model; Models Traceability

I. INTRODUCTION

The importance of software testing is increasingly driven by an extensive dependability on software systems. Software testing is one of the main techniques to enhance and increase the quality of software. Regression testing is a type of software testing that has a clear impact on the quality of software systems that evolve extremely over time in order to meet the needs for new requirements.

Regression testing is one of the methods used in increasing the quality of software. Regression testing is mostly used when new changes are made on the software and it aims to ensure that the introduced changes do not incur errors and change the intended behavior of the software. However, the cost of regression testing is very high since the tester needs to rerun all test cases of the previous test suites. Regression Test Selection (RTS) is an approach used in reducing the number of test cases to run on the modified software.

The main objective of regression testing is to uncover errors in the software after a new modification has been made. Moreover, it is to ensure that the new changes have not introduced more errors in the software. A quite good amount of research has been conducted in the area of regression testing including traceability of regression [1], test automation, test environment [2], reduction of the code size [3] [4] [5] [6] where several regression testing techniques and tools are used and compared. Regression testing uses the previous test suites to find if the new modification caused errors or not, as such, it would be very expensive to run all the test cases. Regression test selection is used to minimize the cost of regression testing by selecting sub-set of test cases in each test suite in the testing process. Many regression test selection techniques have been proposed [7] [8] [9] [10] [11] [12] [13].

Regression test selection techniques are directed to address the problem of reducing the regression test after software system modification [17, 18, 19, 20]. To the best of our knowledge, no work investigates the test suite reduction based on the specific reduction in software structure. Except for the research on change propagation [8], they provided a model-driven approach that maps structural adaptations in autonomic software, to update for its runtime test model.

The main contributions of our work are summarized as follows: we introduce an approach that employs Meta models in test case reduction; it is based on creating several models that are associated with different targets. We created two models that represent the test and component structure models of the software systems under study. We design and build the meta-models using Eclipse Modeling Framework [14]. Our approach synchronizes test models with their corresponding structure model. When any changes takes place in the component structure model of the system under test (reductive modification), component meta model will specify and transmit changes that should be taken to update the test model. In this work, we address the modification of software system when an existed component is removed. After removing test cases that belong to particular component, they were updated/ deleted according to the role of the targeted components. We performed several reductive modifications of three Java software systems that are placed in different domains. To measure the effectiveness of our proposed approach, automatic

inter-class mutants were inserted into the source code by MuJava tool [15]. MuJava tool is widely used to perform mutation analysis [16].

II. BACKGROUND

This section discusses the related work on regression testing, regression testing techniques, model synchronizations and naming convention techniques:

A. Regression Testing

Let *S* be a system or software, Let *S'* be a modified version of *S* and TC be a test case for S. The standard regression testing process is described as following:

Select *TC'* subset if *TC*, a set of test cases to execute on *S'*

Test S' with TC', ascertaining the correctness of S' through TC'

Construct TC'', a set of additional test cases for S'

Test S' with TC'', ascertaining the correctness of S'' through TC''

Construct TC''', a new test case for S' from TC, TC', and TC''.

B. Regression Test Selection Techniques

Regression test selection is the process of choosing a subset of suitable test cases from an original set of test cases to test and ensure that the changes introduced do not reveal errors. Regression test selection process involves two main steps: (1) discovering and highlighting the modified segments of the system , (2) test case selection which means selecting a subset of test case from the original set of test cases that can successfully test the unchanged segments of the software [17].

Many researchers have proposed approaches on techniques of regression test selection [18] [19] [20]: Following are some of the approaches in the literature:

- **Minimization technique:** this technique reduces the number of test cases through selecting a minimum set of test cases with the intention of getting coverage of changed or altered segments of the software. This technique depends on the finding and expressing the relations between basic blocks, test cases, and selecting set of test cases that ensure that each modified basic block is covered by at least one test case [21].
- **Dataflow techniques:** this technique uses the definition-use pair in reducing the number of test cases through selecting test cases that cover each changed definition-use pair.
- **Safe technique:** this technique differs from the above techniques in that the selected test cases have the ability to reveal errors in the modified and updated system. One technique in safe regression test selection is using control flow graph to represent the system under test;
- **Ad hoc (random technique):** this technique is used when the development team does not have enough time to execute all test cases and when the too is not available for test selection. Testers select number of test cases arbitrarily.
- **Retest all:** this technique uses all test cases and runs them against the modified software without excluding any of them which is very expensive computationally especially when there is a huge amount of test cases.

C. Model Synchronization and Naming Convention Strategy

Model Synchronization is the process of confirming the correspondence between two models as soon as one model is changed. Originally, the model synchronization is offered in the Model Driven Architecture (MDA) in order to obtain instrument for obtaining uniformity and modification traceability [22] [23]. Two approaches are used in model synchronization: explicit and implicit. In the implicit approach the relative among models are enclosed in higher order expressions. In the explicit approach the dependence relatives among models are enclosed and encoded directly.

Name convention is the process of concluding beneficial information from a set of harmony data. The strategy of naming convention is used to control the traceability correlation in the functional requirements that structure of the system which helps the engine of modification propagation to search for certain test according to the requirements [24] [25].

We utilize naming convention strategy towards managing the relationship of several components in structure model and their associated tests in the test model. Naming convention strategy allows automatic search for certain related-test items in the test model. In order to deal with the consistency between included models (i.e., models of test and component) our synchronization approach is founded on traceability relations among the interconnected models. Conventions involved use unique and distinctive identifiers for entire test cases and components, and using again component IDs inside test IDs for traceability.

III. METHODOLOGY

In this section, we describe our approach to regression test selection (i.e. reduction) using change traceability of software structure to its test model that occurs during software maintenance.

Figure 1 summarizes the main steps of our approach. The shaded boxes represent the major steps, and ovals represent inputs and outputs associated with each step. The approach consists of five main steps: dependency extraction, creation of test and software structure dynamic models, simulates reductive changes experiments, mutants' generation, and fault detection effectiveness measurement. The following subsections describe each step in detail.

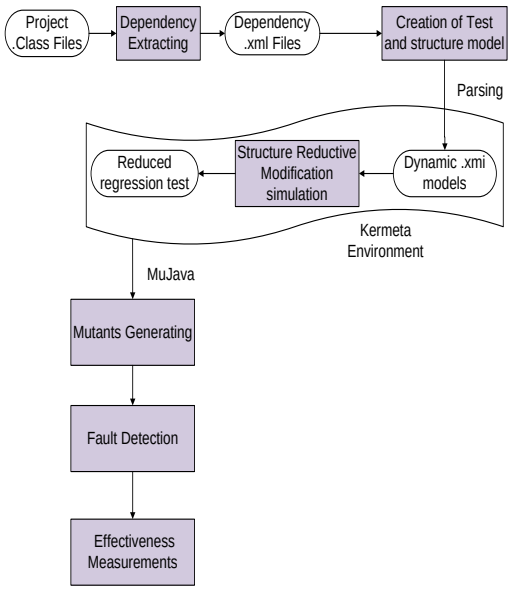


Fig. 1. Input, output and major steps of the proposed approach

A. Dependency Extraction

Dependency happens when one class in the system makes use of another in order to accomplish a specific task. For instance, this can occur when an object of one class is used in another class. These dependencies are helpful and valuable for both programmers and testers when making modifications on the system. The process of finding and discovering dependencies among classes within the system is called extraction. Dependency finder is an open source software tool which is available on [30], and it has been used in many research areas [26, 27, 28]. This tool discovers and reveals three different and specific dependencies: (1) feature to feature (2) feature to class and class to class. The feature means any part of class such as attribute name, method name and constructors. This tool extracts all the dependencies from any type of compiled Java such as Class files, ZIP files, and JAR files.

B. Dynamic test and structure models

We studied test cases dependencies in Java software system. In particular, we created a met model that can be utilized to help in reducing the regression tests after software system modification. In this paper, the meta-model plays a major role in propagating the changes from software structure to test model.

Figure 2 presents a meta model revealing the dependencies in a test model for a given Java software system.

In order to design and build a meta-model, we have chosen Eclipse Modeling Framework [14].EMF is a modeling framework and code generation facility for building tools and other applications based on a structured data model.

The EMF code generation facility is capable of generating everything needed to build a complete editor for an EMF model.

As shown in Figure 2, each test case in the model is composed mainly of two dependencies. (1) Test Hierarchical: for the current test to be run, other tests must be executed and pass (2) Internal: consisting of the Component Under Test (CUT), test drivers, and test stubs. Keeping information on the component under test allows maintaining the traceability links with associated elements in the test model such as scaffolding test.

In order to automatically create the dynamic test and structure models (.xmi), we created a Java based parser to catch the required information from the .xml files that were populated by Dependency finder tool. Dependency tool provides method to address the entire test suite and component structure dependency. We picked naming convention a strategy to manage the traceability.

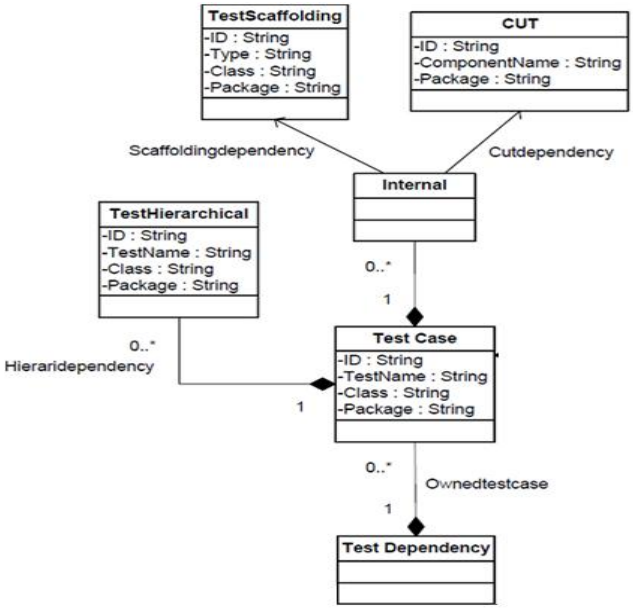


Fig. 2. A meta-model to support regression test reduction technique

C. Simulating Reductive Modification

Reductive modification occurs when the software maintenance or evolution is about to remove existing component interfaces and their implementations from the system. Removing component from the body of the software system is required to remove the unit tests that belong to that element from the test suite. As the unit test checks a single assumption about the behavior of the component.

In this paper, we address the cases when the targeted component for reductive simulation has a dependent and/or dependee. Where the dependees are components and/or tests that are called by the targeted component. In the case of the targeted component has dependees, the integration test that validated the interaction between the targeted components and dependees might be removed from the test suites.

Our assumption is that there is no cyclic dependency in the software projects (i.e. component A depends on component B means, removing A will not affect B as the dependency is unidirectional). The last case occurs when the targeted component has dependents where removing the component

will affect the work of its dependents. Necessary updates to the integration tests that validate the interaction between the targeted component and its dependents will be made.

Software structure and test models instantiation and propagation were achieved using Kermeta, which facilitates the programmatic manipulation of EMF models (.ecore files).

Project Name	# of Reductive Experiments	Percentage of Unit Test Case Reduction	Percentage of Integrated Test Case Reduction
BlackJack	**5**	**43%**	**18%**
PureMVC	**10**	**23%**	**77%**
RealState	**13**	**29%**	**43%**

Kermeta [29] is a meta-modeling language which allows describing both the structure and the behavior of models. Kermeta is intended to be used as the core language of a model

oriented platform. It has been designed to be a common basis to implement Metadata languages, action languages, constraint languages or transformation languages.

Kermeta therefore provided us with a programming environment with which we could set up our simulation.

Mutants generating, and fault detection effectiveness measurement will be described in details in E.1 and E.2 subsections.

D. Experimental Data

The projects that were used in this study are four open-source applications implemented in Java. Table I shows a summary of the selected applications. The selected applications are different in the development processes, features, goals, and the domain.

TABLE I. APPLICATIONS UNDER STUDY

Project	# Classes	# Methods	Source
RealState	57	336	http://realsearchgroup.com /rose/
PureMVC	22	129	http://puremvc.org/
BlackJack	18	102	https://code.google.com/p/ blackjack

The two game applications (RealState and BlackJack) represent applications where systems have to interact to satisfy the logic rules of the underlying strategy of the game logic. Since Java web-applications are widely used nowadays, we chose PureMVC which implements the famous web design pattern Model-View-Controller (MVC).

E. Result and Discussion

Eclipse Modeling Framework (EMF) provides a method to help us in loading, changing and saving software structure and test models by using Kermeta [29]. We used Kermeta language and environment to simulate a reductive change in the abovementioned Java systems. This was achieved by creating and applying a transformation mechanism to the projects under study. Our approach created and applied a set of transformative actions to update and reduce the test models. We arbitrarily simulated 14 reductive changes in RealState project, 10 reductive changes in PureMVC project, and 6 reductive changes in BlackJack project.

1) Test Suite Size Reduction

Table II shows the total number of reductive experiments in each system under test, along with the percentage that reveals the ability of our proposed approach is to reduce test suite size of the three systems. We measure the percentage of test cases reduction for each component that was targeted in the reductive simulation, and then we calculate the mean reduction percentage for the unit and integration test in each system.

TABLE II. TEST SUITE SIZE AFTER SELECTION

We performed 5, 10, and 13 reductive experiments in BlackJack, PureMVC, and RealState, respectively. Our approach was able to reduce 43% of total unit cases and 18% of the integrated test cases in BlackJack system. Interestingly, in PureMVC the proposed approach performed the best reduction in the integration test cases which is about 77%. Finally, the approach reduced 29% and 43% of the total unit and integration test suite respectively in RealState system.

1) Fault Detection

Test selection techniques and after system modification are targeted to reduce the cost of regression test by choosing a portion of an existing test suite, these techniques might lead to lower fault detection effectiveness by neglecting crucial test cases that detect an existing faults. The trade-off between selecting a subset of test cases in order to reduce test suite and fault detection effectiveness should be addressed when we run a specific regression test selection technique.

To measure the effectiveness of fault detection using the proposed approach, we compared it with retest-all approach. Retest-all technique simply reuses all existing test cases after system modification.

In order to measure the effectiveness of the proposed approach, inter-class mutants were seeded into the source code of the above mentioned systems automatically by MuJava tool [15]. We ran retest all technique to examine how many mutants could be killed by executing all the test cases associated with each project (both unit and integrated test cases). We then execute the same mutants against the reduced test suite that is produced by our approach. Finally, we compare the number of killed mutants using the two approaches, that is, retest all test cases approach and the suggested reductive test case approach. Table III shows the fault detection effectiveness of selected test suites using our approach in comparison with rerun-all test cases regression technique.

TABLE III. FAULT DETECTION EFFECTIVENESS AFTER SELECTION

Project Name	# of Mutants	Killed Mutants	
		Retest All	After Reductive Changes
BlackJack	**25**	**13**	**13**
PureMVC	**86**	**35**	**35**
RealState	**121**	**68**	**68**

Table III shows the total number of mutants generated by MuJava and the number of killed mutants using our and retest all approache. MuJava generates 25 mutants from the

BlackJack project where 13 of them have been killed after executing all test cases (without reduction). The results show that our approach and after selection a subset of existing test cases is able to achieve the same degree of effectiveness in uncovering mutants in comparison with retest all technique. After executing the subset test cases which were selected by our approach on the systems under study, retest all and our approach killed equal number of mutants 13 out of 25, 35 out of 86, and 68 out of 121. Although the selected test cases were not detecting all seeded mutants, yet they reduced the test suite and achieved the effectiveness of retest all technique.

IV. Conclusion and Future Work

RTS is an approach used in reducing the number of test cases to run on the modified software. We employ meta-models to support regression test reduction. Our approach facilitates tracing crucial items in test models and its corresponding item in structure model of a Java system, when any changes take place in the component structure model of the system under test (reductive modification), component meta model will specifies and transmits changes should be taken to update the test model(removing, updating, and adding test cases). The result of our experiments reveals how our approach reduced test suite effectively without influence the fault detection effectiveness in comparison with retest-all regression test selection technique.

In future, we intend to perform controlled experiments to compare our approach with other regression test selection techniques are existed in the literature. We intend to use big Java applications to measure the effectiveness of our approach in detecting errors.

References

[1] Leung, H. K. N., and White, L. 1989. Insignts into regression testing. Proceedings of the Conference on Software Maintenance—1989 October, 60–69.

[2] Brown, P. A., and Hoffman, D. 1990. The application of module regression testing at TRIUMF. Nuclear Instruments and Methods in Physics Research Section A, .A293(1–2): 377–381.

[3] Binkley, D. 1992. Using semantic differencing to reduce the cost of regression testing. Proceedings of the Conference on Software Maintenance—1992 November, 41–50.

[4] Leung, H. K. N., and White, L. 1990. A study of integration testing and software regression at the integration level. Proceedings of the Conference on Software Maintenance—1990 November, 290–300.

[5] Leung, H. K. N., and White, L. J. 1991. A cost model to compare regression test strategies. Proceedings of the Conference on Software Maintenance—1991 October, 201–208.

[6] Lewis, R., Beck, D. W., and Hartmann, J. 1989. Assay—a tool to support regression testing. ESEC '89. 2nd European Software Engineering Conference Proceedings September, 487–496.

[7] H. Agrawal, J. Horgan, E. Krauser, and S. London, "Incremental Regression Testing," Proc. Conf. Software Maintenance, pp. 348–357, Sept. 1993.

[8] R. Gupta, M.J. Harrold, and M.L. Soffa, "An Approach to Regression Testing Using Slicing," Proc. Conf. Software Maintenance, pp. 299–308, Nov. 1992.

[9] T. Ball, "On the Limit of Control Flow Analysis for Regression Test Selection," Proc. Int'l Symp. Software Testing and Analysis,ISSTA, Mar. 1998.

[10] S. Bates and S. Horwitz, "Incremental Program Testing Using Program Dependence Graphs," Proc. 20th ACM Symp. Principles of Programming Languages, Jan. 1993.

[11] P. Benedusi, A. Cimitile, and U. De Carlini, "Post-Maintenance Testing Based on Path Change Analysis," Proc. Conf. Software Maintenance, pp. 352–361, Oct. 1988.

[12] D. Binkely, "Semantics Guided Regression Test Cost Reduction," IEEE Trans. Software Eng., vol. 23, no. 8, Aug. 1997.

[13] D. Binkley, "Reducing the Cost of Regression Testing by Semantics Guided Test Case Selection," Proc. Conf. Software Maintenance, Oct. 1995.

[14] Eclipse Foundation, *Eclipse Modeling Framework*, August 2003, http://www.eclipse.org/modeling/emf/ (July 2013).

[15] Yu-Seung Ma, Jeff Offutt, and Yong Rae Kwon, Mujava: an automated class mutation system: Research articles, Softw. Test. Verif. Reliab. 15 (2005), 97–133.

[16] Ammar Masood, Rafae Bhatti, ArifGhafoor, and Aditya P. Mathur, Scalable and effective test generation for role-based access control systems, IEEE Trans. Softw.Eng. 35 (2009), 654–668.

[17] Swarnendu Biswas and RajibMall . Regression Test Selection Techniques: A Survey. Informatica 35 (2011) 289–321

[18] E. Engström, P. Runeson, and M. Skoglund. A systematic review on regression test selection techniques. Information and Software Technology, 52(1):14–30, January 2010.

[19] E. Engström, M. Skoglund, and P. Runeson. Empirical evaluations of regression test selection techniques: a systematic review. In Proceedings of the Second ACM-IEEE international symposium on Empirical software engineering and measurement, pages 22–31, 2008.

[20] J. Bible, G. Rothermel, and D. Rosenblum. A comparative study of coarse- and fine-grained safe regression test-selection techniques. ACM Transactions on Software Engineering and Methodology, 10(2):149–183, April 2001.

[21] FISCHER, K., RAJI, F., AND CHRUSCKICKI, A. 1981. A methodology for retesting modified software. In Proceedings of the National Tele. Conference B-6-3 (Nov.). 1–6.

[22] H. Larsson and K. Burbeck. Codex - an automatic model view controller engineering system. In Proceedings of the Workshop on Model Driven Architecture: Foundations and Applications, Enschede, The Netherlands, June 2003.

[23] Igor Ivkovic and Kostas Kontogiannis. Tracing Evolution Changes of Software Artifacts through Model Synchronization. Proceedings of the 20th IEEE International Conference on Software Maintenance (ICSM'04).

[24] Abdallah Qusef, Rocco Oliveto, and Andrea De Lucia, Recovering traceability links between unit tests and classes under test: An improved method, Proceedings of the 2010 IEEE International Conference on Software Maintenance (Washington, DC, USA), ICSM '10, IEEE Computer Society, 2010, pp. 1–10.

[25] Bart Van Rompaey and Serge Demeyer, Establishing traceability links between unit test cases and units under test., CSMR (Andreas Winter, Rudolf Ferenc, and JensKnodel, eds.), IEEE, 2009, pp. 209–218.

[26] McCartin J. Tempero E. Dietrich, J. and S. M. A. Shah, On the existence ofhigh-impact refactoring opportunities in programs, Australasian Computer ScienceConference (ACSC 2012) (Melbourne, Australia) (M. Reynolds and B Thomas, eds.),CRPIT, vol. 122, ACS, 2012, pp. 37–48.

[27] RdigerLincke, Jonas Lundberg, and WelfLwe, Comparing software metrics tools., ISSTA (Barbara G. Ryder and Andreas Zeller, eds.), ACM, 2008, pp. 131–142.

[28] Alexander Serebrenik, Serguei A. Roubtsov, and Mark van den Brand, Dn-basedarchitecture assessment of java open source software systems., ICPC, IEEE ComputerSociety, 2009, pp. 198–207.

[29] Triskell Team, *Kermeta - Breathe life into your metamodels*, October 2005,http://www.kermeta.org/ (July 2013).

[30] Tessier J., Dependency finder, 2008, http://depfind.sourceforge.net

A Novel Method to Design S-Boxes Based on Key-Dependent Permutation Schemes and its Quality Analysis

Kazys Kazlauskas
Institute of Mathematics and Informatics
Vilnius University
Vilnius, Lithuania

Robertas Smaliukas
Institute of Mathematics and Informatics
Vilnius University
Vilnius, Lithuania

Gytis Vaicekauskas
Institute of Mathematics and Informatics
Vilnius University
Vilnius, Lithuania

***Abstract*—S-boxes are used in block ciphers as the important nonlinear components. The nonlinearity provides important protection against linear and differential cryptanalysis. The S-boxes used in encryption process could be chosen to be key-dependent. In this paper, we have presented four simple algorithms for generation key-dependent S-boxes. For quality analysis of the key-dependent S-boxes, we have proposed eight distance metrics. We have assumed the *Matlab* function "*randperm*" as standard of permutation and compared it with permutation possibilities of the proposed algorithms. In the second section we describe four algorithms, which generate key-dependent S-boxes. In the third section we analyze eight normalized distance metrics which we have used for evaluation of the quality of the key-dependent generation algorithms. Afterwards, we experimentally investigate the quality of the generated key-dependent S-boxes. Comparison results show that the key-dependent S-boxes have good quality and may be applied in cipher systems.**

Keywords—data encryption; substitution boxes; generation algorithms; distance metrics; quality analysis

I. INTRODUCTION

Cryptographic objects are private key algorithms, public key algorithms and pseudorandom generators. Block ciphers usually transform the 128 or 256 bits string of the plaintext to a string of the same length cipher text under control of the secret key. The private key cryptography, such as DES [1], 3DES, and Advanced Encryption Standard (AES) [2], uses the same key for the sender and for the receiver to encrypt the plaintext and to decrypt the ciphertext. The private key cryptography is more suitable for the encryption of a large amount of data. The public key cryptography, such as the Rivest-Shamir-Adleman algorithm defined by the National Institute of Standards and Technology of the Unite (RSA) or Elliptic Curve algorithms, uses different keys for encryption and decryption. The AES has been accepted to replace DES. AES overpasses DES in an improved security because of larger key sizes. AES is suitable for 8 bit microprocessor platforms and 32 bit processors [3].

The essential part of every block cipher is an S-box. To secure the cipher against attacks, the nonlinearity of the S-box should have a maximum nonlinearity, and the difference propagation probability should be minimum. Substitution is a nonlinear transformation that performs confusion of bits. A nonlinear transformation is important for every encryption algorithm and it is proved to be a strong cryptographic method against the linear and differential cryptanalysis. Nonlinear transformations are implemented as lookup tables (S-boxes). An S-box with p input bits and q output bits is denoted as $p \rightarrow q$. The DES uses eight $6 \rightarrow 4$ S-boxes. S-boxes are designed for software implementation on 8-bit processors. The block ciphers with $8 \rightarrow 8$ S-boxes are SAFER, SHARK, and AES. For processors with 32-bit or 64-bit words, S-boxes with more output bits provide a high efficiency. The Snefru, Blowfish, CAST, and SQUARE use $8 \rightarrow 32$ S-boxes. The S-boxes can be selected at random as it is in Snefru, and can be computed using a chaotic map, or have some mathematical structure over a finite Galois field. Examples of the last approach are SAFER, SHARK, and AES. Key-dependent S-boxes are slower, but more secure than the key independent S-boxes [4]. The use of the key independent chaotic S-boxes are analyzed in [5], in which the S-box is constructed with a transformation $F((X+K) \bmod M)$, where K is the key [6].

There are two ways to fight against the linear and differential cryptanalysis. The first one is to create S-boxes with low linear and differential probabilities. The other is to design the round transformation so that only trails with many active S-boxes would occur. The round transformation must be designed in such a way that differential steps with few active S-boxes would be followed by differential steps with many active S-boxes [6].

Two general principles of block ciphers are confusion and diffusion. Confusion is transformation that changes the dependence of the statistics of the cipher text on the statistics of the plaintext. Diffusion is spreading of the influence of one plaintext bit to many cipher text bits with the purpose to hide the statistical structure of the plaintext. In most cipher systems the confusion and diffusion are achieved by means of round repetition. Repeating a single round contributes to the cipher's simplicity [6]. Modern block ciphers consist of four transformations: substitution, permutation, mixing, and key-adding [7], [8].

Block cipher systems depend on the S-boxes, which are fixed and have no relation with the secret key. So only a changeable parameter is the secret key. The only nonlinear

component of AES is S-boxes, so they are an important source of cryptographic strength. Research of the S-box design has focused on determination of S-box properties which yield cryptographically strong ciphers, with the aim of selecting a small number of good S-boxes for use in a block cipher DES and CAST [8]. Some results have demonstrated that a randomly chosen S-box of sufficient size will have several of these desirable properties with a high probability [9]. In [10] a dynamic AES-128 with a key-dependent S-box is designed and implemented. The paper of [11] presents a new AES-like design for key-dependent AES using the S-box rotation method. An approach for designing a key-dependent S-box defined over $GF(2^4)$ in AES is presented in [12]. A key-dependent S-box of AES algorithm using a variable mapping technique is analyzed in [13]. A key-dependent S-box generation algorithm in AES block cipher system is proposed in the paper [14]. Hamdy *et al.* [15] have proposed a customized version of the AES block cipher in which the key-dependent S-box generation algorithm is used. In the paper Hosseinkhany *et al*. the dynamic S-box is generated in the AES cipher system using the secret key [16]. Other systems, using key-dependent S-boxes were proposed in the past, the most well-known of which is Blowfish [7] and Khufu [17]. Each of these two systems uses the cryptosystem itself to generate the S-boxes. In [19] for generation S-boxes an algorithm based on chaotic map and composition method is used. In [20] a method for the construction of block ciphers with multi-chaotic systems is proposed. In the paper of D. Lambic the security analysis and improvement of a block cipher with dynamic S-boxes based on tent map is analyzed [21]. In the paper of Ozkaynak *et al*., is done analysis of a novel image fusion encryption algorithm based on DNA sequence operation and hyper-chaotic system [22].

This paper outlines the work of the authors' investigation into the design of a new pseudo-randomly generated key-dependent S-boxes. We have presented four simple algorithms for generation key-dependent S-boxes. For quality analysis of the key-dependent S-boxes, we have proposed to use eight distance metrics. Modeling results show, that the proposed algorithms have a good cryptographic strength, with an additional benefit that the algorithms are resistant to the linear and differential cryptanalysis, which require that the S-boxes be known. In the second section, we analyze four algorithms for generation of key-dependent S-boxes. In the third section we propose eight distance metrics for evaluation of the quality of key-dependent S-boxes. Afterwards, we discuss the experimental results and give conclusions.

II. Algorithms for Generation Key-Dependent S-Boxes

In this section we analyze four algorithms for generation of key-dependent S-boxes *Sboxm*. These algorithms use some key-dependent permutations of the elements of the initial substitution box *Sbox* to get key-dependent substitution box *Sboxm*. Algorithm 1 was proposed in the paper [18].

The initial substitution box *Sbox* may be the AES substitution box (table) or the ordered numbers 0,1,…,255, or these numbers mixed in any order. In all these cases the sender and the receiver must know these initial S-boxes. We assume that the S-boxes are rearranged according to the rows to the 256-size vectors, i.e., the initial substitution box *Sbox*(*i*), i = 0,1,…,255 and the key-dependent substitution box *Sboxm*(*i*),

i = 0,1,…,255 are 256-size vectors of the different integer numbers (bytes) from the interval [0, 255]. The indexes *i* of these vectors are also the integer numbers (bytes) from the interval [0, 255]. In the encryption process, the indexes *i* of the vector *Sbox* (or *Sboxm*) are replaced by the corresponding values *Sbox*(*i*), (or *Sboxm*(*i*)). In the decryption process, the values of the vector *Sbox*(*i*), (or *Sboxm*(*i*)) are replaced by the corresponding indexes of the vector *Sbox* (or *Sboxm*).

A. Algorithm 1 (A1)

Input:

The secret key *key*(*i*), $i = 1,\ldots, l$ is the vector of l integer numbers (bytes) from the interval [0, 255].

The initial substitution box *Sbox*(*i*), i = 0,1,…,255 is the vector of different integer numbers (bytes) from the interval [0,255].

Output:

The key-dependent substitution box *Sboxm*(*i*), i = 0,1,…,255 is the vector of the integer numbers (bytes) from the interval [0, 255].

The key-dependent inverse substitution box *invSboxm*(*i*), i = 0,1,…,255 is the vector of different integer numbers (bytes) from the interval [0, 255].

1. Compute the initial value *j*, which depends on all the secret key's values *key*(*i*), i = 1,2,…,l from the interval [0, 255]:

$$j \leftarrow \sum_{i=1}^{l} key(i) \bmod 256$$

for all $i = 0,1,\ldots,255$ *do*

2. Compute the index *j* which depends on the values of the initial substitution box *Sbox* and on the values of the secret key *key*:

$$k \leftarrow (Sbox(i) + Sbox(j)) \bmod l$$
$$j \leftarrow (j + key(k)) \bmod 256$$

3. Replace the values *Sbox*(*i*) by the values *Sbox*(*j*), and the values *Sbox*(*j*) by the values *Sbox*(*i*):

$$p \leftarrow Sbox(i)$$
$$Sbox(i) \leftarrow Sbox(j)$$
$$Sbox(j) \leftarrow p$$
end for

4. Write the key-dependent substitution box values to *Sboxm*:

$$Sboxm \leftarrow Sbox$$

5. Compute the key-dependent inverse substitution box values *invSboxm*:

for all $i = 0,1,\ldots,255$ *do*
$$invSboxm(Sboxm(i)) \leftarrow i$$
end for

B. Algorithm 2 (A2)

Input:

The secret key *key*(*i*), $i = 1,\ldots, l$ is the vector of l integer numbers (bytes) from the interval [0, 255].

The initial substitution box *Sbox* is (16x16)-size matrix of the different integer numbers (bytes) from the interval [0, 255].

Output:

The key-dependent substitution box *Sboxm*(*i*), $i = 0,1,\ldots,255$ is the vector of the different integer numbers (bytes) from the interval [0, 255].

The key-dependent inverse substitution box *invSboxm*(*i*), $i = 0,1,\ldots,255$ is the vector of different integer numbers (bytes) from the interval [0, 255].

1) 128 bits of the secret key key divide to the left (key_1) and right (key_2) equal parts.

2) The left part of the key key_1 divide into 16 equal parts $k_1(i)$, $i = 1,2,\ldots,16$. $k_1(i)$ are the integer numbers from the interval [0,15].

3) for all $i = 1,2,\ldots,16$ do

cyclically rotate bytes of the rows of the initial substitution box *Sbox* to the left according to $k_1(i)$

end for

The result is the intermediate substitution box $Sbox_1$.

4) The right part of the key key_2 divide to the 16 equal parts $k_2(j)$, $j = 1,2,\ldots,16$. $k_2(j)$ are the integer numbers from the interval [0,15].

5) for all $j = 1,2,\ldots,16$ do

cyclically rotate bytes of the columns of the intermediate substitution box $Sbox_1$ to the left according to $k_2(j)$.

end for

The result is the key dependent substitution box *Sboxm*.

6) Rearrange key-dependent substitution box ((16x16)-size matrix) *Sboxm* to the vector.

7) Compute the key-dependent inverse substitution box invSboxm:

for all $i = 0,1,\ldots,255$ *do*

$invSboxm(Sboxm(i)) \leftarrow i$

end for

C. Algorithm 3 (A3)

Algorithm 3 is a mixed version of Algorithm 1 and Algorithm 2. In that case, the initial substitution box *Sbox* is the input of the Algorithm 1. The output of the Algorithm 1 is the intermediate substitution box *Sboxm*1. The input of the Algorithm 2 is the intermediate substitution box *Sboxm*1. Finally, the output of the Algorithm 2 is the output of Algorithm 3, i.e., *Sboxm*.

D. Algorithm 4 (A4)

Algorithm 4 is a mixed version of Algorithm 2 and Algorithm 1. In that case, the initial substitution box *Sbox* is the input of the Algorithm 2. The output of the Algorithm 2 is the intermediate substitution box *Sboxm*2. The input of the Algorithm 1 is the intermediate substitution box *Sboxm*2. Finally, the output of the Algorithm 1 is the output of Algorithm 4, i.e., *Sboxm*.

III. Distance Metrics for Evaluation of the Quality of S-Boxes

We introduce several distance metrics that are able to calculate a distance between the given initial S-box *Sbox* and the key-dependent S-box *Sboxm*. We assume that the S-boxes are rearranged according to the rows to the *N*-size vectors. The key-dependent S-box *Sboxm* is key-dependent permutations without repetition of a given set of *N* integer elements of the initial box *Sbox*. The initial S-box *Sbox* and the key-dependent S-box *Sboxm* have the same length. For AES S-box $N = 256$. The *i*-th integer element of S-box is represented as *Sbox*(*i*). We have normalized all distances between the S-boxes. The smaller the value of the normalized distance, the more similar are the *Sbox* and *Sboxm*, and *vice versa*. For example, the normalized distance between *Sbox* and *Sbox* is equal to 0. For convenience, in this chapter we use indexes of S-boxes from 1 to *N* instead of from 0 until *N*–1.

1) The Hamming distance. The Hamming (H) distance between two S-boxes of equal length is defined as a number of not equal elements in the same positions in the initial S-box Sbox and in the key-dependent S-box Sboxm

$$d_H(Sbox, Sboxm) = \sum_{i=1}^{N} d_i, \text{ where } d_i = \begin{cases} 1, \text{ if } Sbox(i) \neq Sboxm(i) \\ 0, \text{ otherwise} \end{cases} \quad (1)$$

The mean of the *H* distance is equal to (*N*+1)/2. The maximal *H* distance is equal to *N*. The normalized *H* distance $\bar{d}_H$ is obtained by dividing the distance d_H by the maximal *H* distance

$$\bar{d}_H = \frac{1}{N} d_H. \quad (2)$$

2) Spearman's distance. Spearman's (S) distance is an absolute distance between two rank vectors

$$d_S(Sbox, Sboxm) = \sum_{i=1}^{N} \sum_{j=1, i \neq j}^{N} |i - j|,$$

for such (*i*, *j*) that *Sbox* (*i*) = *Sboxm*(*j*) (3)

S distance is the summation of the absolute differences between two ranks of all equal elements from *Sbox* and *Sboxm*. *S* distance is similar to the Manhattan distance that used for quantitative variables. The mean of the *S* distance is equal to $N^2/3$. The *S* distance is maximal between *Sbox* and inverted *Sbox*, and is equal to $N^2/2$ for even *N* and to $(N^2-1)/2$ for odd *N*. The normalized *S* distance for even *N* is defined as

$$\bar{d}_S = \frac{2}{N^2} d_S, \text{ for even } N. \quad (4)$$

3) Squared Spearman's distance. The squared Spearman's (SS) distance assigns a larger distance when deviations between equal elements of two S-boxes are larger. It is defined as follows:

$$d_{SS}(Sbox, Sboxm) = \sum_{i=1}^{N} \sum_{j=1, i \neq j}^{N} (i-j)^2,$$
$$for\ such\ (i, j)\ that\ Sbox(i) = Sboxm(j). \quad (5)$$

The mean of the *SS* distance is equal to $N^3/3$. The maximal *SS* distance is equal to $(N^3 - N)/3$. The normalized *SS* distance is defined as

$$\bar{d}_{SS} = \frac{3}{N^3 - N} d_{SS}. \quad (6)$$

The *SS* distance is similar to Spearman's rank correlation coefficient, a metric often used in statistics to calculate the correlation between two rankings.

4) The T distance. The T distance is the number of transpositions required to transform Sbox into *Sboxm*.

$$d_T(Sbox, Sboxm) = \sum_{i=1}^{N-1} \sum_{j=1}^{N-1} d_{ij},\ where\ d_{ij} =$$
$$= \begin{cases} 1\ if\ \ Sbox(i) = Sboxm(j)\ \&\ Sbox(i+1) = Sboxm(j+1) \\ 0\ otherwise \end{cases} \quad (7)$$

The maximal value of d_T is *N*–1. The normalized *T* distance is defined as

$$\bar{d}_T = 1 - \frac{1}{N-1} d_T. \quad (8)$$

5) Kendall distance. The Kendall (K) distance is given by

$$d_K(Sbox, Sboxm) = \sum_{i=1}^{N} \sum_{j=1, i<j}^{N} d_{ij},\ where\ d_{ij} =$$
$$= \begin{cases} 1\ \ if\ \ Sbox(i) < Sbox(j)\ \&\ Sboxm(j) < Sboxm(i) \\ 0\ \ otherwise \end{cases} \quad (9)$$

This distance is equal to the number of pair-wise adjacent permutations required to transform *Sbox* into *Sboxm*. The mean of the *K* distance is equal to $N^2/4$. The maximal value of *K* distance is $(N^2 - N)/2$. The normalized *K* distance is defined as

$$\bar{d}_K = \frac{2}{N^2 - N} d_K. \quad (10)$$

The normalized *K* distance lies in the interval [0,1]. For example, the normalized *K* distance 0.3 indicates that 30 % of the pairs of S-boxes elements differ in ordering between *Sbox* and *Sboxm*.

6) Correlation coefficient distance. We introduce the correlation (C) coefficient distance as follows: normalize Sboxm elements $x = \{x_1, ..., x_N\}$

$$y = \frac{x - mean(x)}{std(x)}, \quad (11)$$

and define the correlation coefficient of *Sboxm* elements as

$$d_C(Sboxm) = std(corr(\tau)), \quad (12)$$

where *mean* is the arithmetic mean, *std* is the standard deviation, *corr*(τ) is the correlation function of *y*. We assume, that *corr*(0) = 0. The maximal value of the correlation coefficient is *N*–1. We define the normalized correlation coefficient distance as follows:

$$\bar{d}_C = 1 - \frac{d_C}{N-1}. \quad (13)$$

7) Pearson's correlation coefficient distance. For the initial S-box Sbox with elements $\{x_1, ..., x_N\}$ and for the key-dependent S-box Sboxm with elements $\{y_1, ..., y_N\}$ the formula of the Pearson (P) correlation coefficient is [23]

$$r = \frac{N \sum x_i y_i - \sum x_i \sum y_i}{\sqrt{N \sum x_i^2 - (\sum x_i^2)} \sqrt{N \sum y_i^2 - (\sum y_i)^2}}, \quad (14)$$

Pearson's correlation coefficient distance between initial S-box *Sbox* and key-dependent S-box *Sboxm* can be rescaled to a distance measure of range [0 – 1] by:

$$\bar{d}_P = 1 - \text{abs}(r). \quad (15)$$

The Pearson's correlation coefficient distance between two S-boxes $\bar{d}_P$ = 1 if correlation coefficient *r* is equal to zero and $\bar{d}_P$ = 0 if correlation coefficient *r* is equal to ± 1.

8) The longest common subsequence distance. The length of the longest common subsequence (LCS) is a measure of the similarity between Sbox and Sboxm. The minimum length of the LCS is equal to one and the maximum is equal to N. We define the LCS distance d_{LCS} (Sbox, Sboxm) as N minus the length of the longest common subsequence. The LCS distance lies between 0 and N–1. The normalized longest common subsequence distance $\bar{d}_{LCS}$ is defined as

$$\bar{d}_{LCS} = \frac{d_{LCS}}{N-1}. \quad (16)$$

IV. Experimental Results and Discussion

A. Experiment 1

Consider the 128 bit length secret key in hexadecimal form:

key = {CA, 6A, C5, 21, 5B, 46, 50, 3D, 98, 19, F0, 72, 6D, 41, 43, C7}. (17)

We have generated the key-dependent S-box *Sboxm* using Algorithm 2 and key (17). The initial S-box *Sbox* is the AES S-box (Table I). The key-dependent S-box *Sboxm* is given in Table II. We have assumed that S-boxes are rearranged

according to the rows to the 256-size vectors, i.e., the initial substitution box *Sbox*(*i*), *i* = 0,1,…,255 and the key-dependent substitution box *Sboxm*(*i*), *i* = 0,1,…,255 are 256-size vectors of the different integer numbers from the interval [0, 255]. Thus, AES S-box (Table I) is the 256-size vector in hexadecimal form: {63, 7C,…,76; CA, 82,…,C0; … ; 8C, A1,…,16} and the permuted key-dependent S-box *Sboxm* (Table II) is the 256-size vector {56, D6,…, 6A; 45, DD,…, CC;…; C5, 52,…, 47}. Using eight metrics, we have calculated normalized distances between these two vectors. The normalized distances between initial AES S-box *Sbox* and the key-dependent S-boxes *Sboxm* are given in Table III. In the row „Algorithm2" of the Table III we have used algorithm A2 for key-dependent permutation of the AES S-box, while in the row „*randperm*" of the Table III we have used Matlab function „*randperm*" for permutation of the AES S-box. From Table III we can see that all distances between the *Sbox* and *Sboxm* for algorithm "*randperm"* and for algorithm A2 are similar. It follows that the proposed algorithm A2 permutes the bytes of the AES S-box no worse than the Matlab function „*randperm*". It confirms the good quality of the proposed key-dependent permutation algorithm A2.

B. Experiment 2

In order to evaluate the performance of our four algorithms, we have generated initial S-box *Sbox* – ordered integer numbers {0,1,…,255}. Then, using Matlab function "*randperm*" and Algorithms A1 – A4, we have calculated randomly permuted integer numbers (bytes) without repetition, i.e. key-dependent S-boxes *Sboxm.* After, we have evaluated eight normalized distances between initial *Sbox* and generated key-dependent S-boxes *Sboxm* for function "*randperm*" and for A1 – A4 algorithms. Such experiments we have repeated 1000 times with different randomly generated keys and have calculated the means and standard deviations of these normalized distances. We have used 128-bit length 1000 random keys, which we have generated using *Matlab* function "*randperm*". We have assumed the *Matlab* function "*randperm*" as standard of permutation of the integer numbers. Hence, we could evaluate the performance of our four algorithms comparing the averaged normalized distances $\bar{d}_i$ (*i* =1,…,8) of the function "*randperm*" with the averaged normalized distances $\bar{d}_i^{(j)}$ (*i* = 1,…,8; *j* = 1,…,4) of the proposed four algorithms using the measure (18)

$$\hat{d}^{(j)} = \frac{1}{8}\sum_{i=1}^{8}\frac{\left|\bar{d}_i - \bar{d}_i^{(j)}\right|}{\bar{d}_i}100\,\%, \quad j = 1,2,3,4 \qquad (18)$$

in which $\bar{d}_i$ is the normalized mean of *i*-th distance between initial *Sbox* and *Sboxm* in case we have used for permutation of the initial S-box "*randperm*" function; $\bar{d}_i^{(j)}$ is the normalized mean of *i*-th distance between initial *Sbox* and key-dependent *Sboxm* in case we have used for permutation of the initial S-box *j*-th algorithm.

TABLE I. AES S-BOX IN HEXADECIMAL FORM

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
0	63	7C	77	7B	F2	6B	6F	C5	30	01	67	2B	FE	D7	AB	76
1	CA	82	C9	7D	FA	59	47	F0	AD	D4	A2	AF	9C	A4	72	C0
2	B7	FD	93	26	36	3F	F7	CC	34	A5	E5	F1	71	D8	31	15
3	04	C7	23	C3	18	96	05	9A	07	12	80	E2	EB	27	B2	75
4	09	83	2C	1A	1B	6E	5A	A0	52	3B	D6	B3	29	E3	2F	84
5	53	D1	00	ED	20	FC	B1	5B	6A	CB	BE	39	4A	4C	58	CF
6	D0	EF	AA	FB	43	4D	33	85	45	F9	02	7F	50	3C	9F	A8
7	51	A3	40	8F	92	9D	38	F5	BC	B6	DA	21	10	FF	F3	D2
8	CD	0C	13	EC	5F	97	44	17	C4	A7	7E	3D	64	5D	19	73
9	60	81	4F	DC	22	2A	90	88	46	EE	B8	14	DE	5E	0B	DB
A	E0	32	3A	0A	49	06	24	5C	C2	D3	AC	62	91	95	E4	79
B	E7	C8	37	6D	8D	D5	4E	A9	6C	56	F4	EA	65	7A	AE	08
C	BA	78	25	2E	1C	A6	B4	C6	E8	DD	74	1F	4B	BD	8B	8A
D	70	3E	B5	66	48	03	F6	0E	61	35	57	B9	86	C1	1D	9E
E	E1	F8	98	11	69	D9	8E	94	9B	1E	87	E9	CE	55	28	DF
F	8C	A1	89	0D	BF	E6	42	68	41	99	2D	0F	B0	54	BB	16

TABLE II. KEY-DEPENDENT S-BOX *SBOXM*. GENERATION ALGORITHM IS A2. INITIAL S-BOX IS AES S-BOX. SECRET KEY IS AS IN (17)

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
0	56	D6	7F	4A	D7	B2	9E	BA	32	AA	0A	1C	FC	F7	F0	6A
1	45	DD	BE	21	50	A4	2F	DF	70	C8	40	6D	48	4D	05	CC
2	78	13	2E	69	9D	5A	9A	BC	35	02	3D	10	D8	58	16	E1
3	1E	DA	14	64	27	9F	76	8C	3E	4F	66	BF	97	B1	A0	C4
4	C0	63	F8	3A	11	F2	2A	33	5B	46	99	7E	62	DE	E3	F3
5	CA	A1	37	0D	FA	06	38	85	C2	01	B8	EA	91	4C	19	15
6	F5	6C	D4	AC	1F	65	3C	0B	75	B7	7C	25	7B	36	D5	44
7	B5	7D	18	A6	90	17	E8	A5	F4	B9	4B	FF	E4	84	04	82
8	24	88	61	12	74	E9	86	5D	AE	CF	09	FD	98	26	1B	03
9	5E	8B	A8	53	C7	89	C3	20	D9	4E	5C	9B	3B	57	0F	CE
A	43	E6	B4	A9	41	CB	87	2B	B0	95	1D	D2	D0	83	77	1A
B	D1	C9	ED	92	6B	F6	C6	30	F9	2D	AF	FE	7A	28	73	51
C	5F	59	8E	0E	AD	B6	67	F1	9C	BD	BB	DB	CD	EF	93	FB
D	8F	22	3F	42	94	34	A7	A2	E2	71	C1	AB	79	60	A3	23
E	EB	55	72	08	E0	0C	2C	EC	49	96	6F	68	07	EE	E5	B3
F	C5	52	D3	80	39	29	54	31	8A	E7	81	00	DC	8D	6E	47

TABLE III. NORMALIZED DISTANCES BETWEEN AES S-BOX AND KEY-DEPENDENT S-BOX *S-BOXM*. GENERATION ALGORITHM IS A2. SECRET KEY IS AS IN (17)

Distance	$\bar{d}_H$	$\bar{d}_S$	$\bar{d}_{SS}$	$\bar{d}_T$	$\bar{d}_K$	$\bar{d}_C$	$\bar{d}_P$	$\bar{d}_{LCS}$
Algorithm2	0.9922	0.6282	0.4770	0.9961	0.4811	0.9605	0.9447	0.8980
"*randperm*"	0.9960	0.6368	0.4805	0.9960	0.4908	0.9560	0.9496	0.8939

In equation (18) and Table IV $\bar{d}_1$, $\bar{d}_1^{(j)}$ are Hamming distances $\bar{d}_H$; $\bar{d}_2$, $\bar{d}_2^{(j)}$ are Spearman's distances $\bar{d}_S$; $\bar{d}_3$, $\bar{d}_3^{(j)}$ are squared Spearman's distances $\bar{d}_{SS}$ and so on.

The proposed normalized correlation distance $\bar{d}_C$ and Pearson correlation distance $\bar{d}_P$ are similar, but standard deviation of the proposed distance $\bar{d}_C$ is about eight times less as compared with $\bar{d}_P$. From Table IV, and according with introduced quality measure (18), we can see that the best is Algorithm 3 – 0.3650 %, after follows Algorithm 4 – 0.4400 %, Algorithm 2 – 2.2849 % and Algorithm 1 – 2.4240 %. From Table V, it follows that Algorithm 1 generates 1000 S-boxes during 0.0940 sec., Algorithm 2 – during 0.2340 sec. and, finally, Algorithm 3 and Algorithm 4 – during 0.3280 sec.

TABLE IV. MEANS AND STANDARD DEVIATIONS OF 8 NORMALIZED DISTANCES BETWEEN INITIAL S-BOX *SBOX* (ORDERED INTEGER NUMBERS {0,1,…,255}) AND KEY-DEPENDENT S-BOXES *SBOXM*

i	**Algorithm** / **Distance**	**„*randperm*“** $\bar{d}_i$	**A1** $\bar{d}_i^{(1)}$	**A2** $\bar{d}_i^{(2)}$	**A3** $\bar{d}_i^{(3)}$	**A4** $\bar{d}_i^{(4)}$
1	$\bar{d}_H$	0.9960 ± 0.0040	0.9936 ± 0.0326	0.9961 ± 0.0066	0.9960 ± 0.0039	0.9961 ± 0.0039
2	$\bar{d}_S$	0.6668 ± 0.0253	0.6381 ± 0.0347	0.6605 ± 0.0762	0.6586 ± 0.0274	0.6585 ± 0.0269
3	$\bar{d}_{SS}$	0.5005 ± 0.0308	0.4669 ± 0.0372	0.4969 ± 0.0562	0.4933 ± 0.0319	0.4899 ± 0.0314
4	$\bar{d}_T$	0.9961 ± 0.0038	0.9914 ± 0.0186	0.9393 ± 0.0553	0.9959 ± 0.0040	0.9959 ± 0.0480
5	$\bar{d}_K$	0.5008 ± 0.0204	0.4803 ± 0.0226	0.4927 ± 0.0376	0.4998 ± 0.0206	0.5008 ± 0.0211
6	$\bar{d}_C$	0.9560 ± 0.0031	0.9554 ± 0.0115	0.9495 ± 0.0078	0.9559 ± 0.0032	0.9559 ± 0.0031
7	$\bar{d}_P$	0.9496 ± 0.0368	0.9246 ± 0.0599	0.9125 ± 0.0660	0.9495 ± 0.0376	0.9487 ± 0.0382
8	$\bar{d}_{LCS}$	0.8939 ± 0.0081	0.8861 ± 0.0280	0.8519 ± 0.0313	0.8940 ± 0.0082	0.8949 ± 0.0080
	$\hat{d}^{(j)}$		2.4240 %	2.2849 %	0.3650 %	0.4400 %

TABLE V. GENERATION TIME OF 1000 KEY-DEPENDENT S-BOXES *S-BOXM* WITH COMPUTER AMD ATHLON-X2, 2.59 MHZ

Algorithm	A1	A2	A3	A4	“*randperm*”
Time, sec.	0.0940	0.2340	0.3280	0.3280	0.0359

V. CONCLUSIONS

This paper proposes a new simple algorithms to generate key-dependent S-boxes. The generated key-dependent S-boxes can be used in block ciphers such as AES cipher. We suggest for testing key-dependent S-boxes to apply eight distance metrics. The results of the experiments and tests show that the new generated S-boxes are truly random. Using our algorithms, we can get 256! different substitution values instead of 256 values as it is in AES S-box. It increases the encryption complexity and aggravate the cryptanalysis process. It was established that for any changing secret key, the structure of the key-dependent S-boxes are changing essentially. Also it was shown that this is achieved with negligible time delay. For example, if we use algorithm A1, 1000 S-boxes were generated during 0.0940 sec. These algorithms will increase the security of the cipher systems. As compared with algorithm in the paper [14], these algorithms generate S-boxes about eight times faster. For measure of the quality of the permuted S-boxes we have proposed to use eight distance metrics. The distances between initial S-boxes and key-dependent S-boxes of our algorithms we have compared with appropriate distances of the *Matlab* function “*randperm*”. We have assumed this function as standard of permutation of the integer numbers (bytes) of S-boxes. Also it was found that the proposed new correlation distance metric $\bar{d}_C$ as compared with Pearson distance metric $\bar{d}_P$ is about eight times more accurate.

REFERENCIES

[1] Data Encryption Standard (DES) National Bureau of Standards. FIPS Publication, 1977.

[2] Advanced Encryption Standard (AES). Federal Information Processing Standards Publication 197, 2001.

[3] P. Su, T. F. Lin, C. T. Huang, and C. W. Wu, “A high-throughput low-cost AES processor,” IEEE Communications Magazine, vol. 41, pp. 86-91, 2003.

[4] B. Schneier, ”Description of a new variable-length 64-bit block cipher,” Fast Software Encryption, pp.191-204, 1997.

[5] G. Jakimovski and L. Kocarev, “Chaos and cryptography: block encryption ciphers based on chaotic maps,” IEEE Transaction on Circuits and Systems, Part I, vol. 48, pp. 163-169, 2001.

[6] N. Masuda, G. Jakimovski, K. Aihara, and L. Kocarev, “Chaotic block ciphers: from theory to practical algorithms,” IEEE Trans. on Circuits and Systems – I: Regular Papers, vol. 53, pp. 1341-1352, 2006.

[7] B. Schneier, Applied Cryptography: Protocols, Algorithms, and Source Code in C. Wiley, 1996.

[8] J. Menezes, P. C. Van Oorschot, and S. A. Vanstone, Handbook of Applied Cryptography. CRC, 1997.

[9] L. Keliher, Linear cryptanalysis of substitution-permutation networks. PhD Thesis, Queen’s University, Kingston, Canada, 2003.

[10] M. Mahmoud, A. B. El Hafez, T. A. Elgarf, and A. Zekry, “Dynamic AES-128 with key-dependent S-box,” Int. J. of Engineering Research and Applications, vol.3, pp.1662-1670, 2013.

[11] J. Juremi, R. Mahmod, S. Sulaiman, and J. Ramli, “Enhancing Advanced Encryption Standard S-box generation based on round key,” Int. J. of Cyber-Security and Digital Forensics, vol.1, pp.183-188, 2012.

[12] M. El-Sheikh, O. A. El-Mohsen, and A. Zekry, “A new approach for designing key-dependent S-box defined over GF in AES,” Int. J. of Computer Theory and Engineering, vol. 4, pp.158-164, 2012.

[13] Y. Mohammad, A. E. Rohiem, and A. D. Elbayoumy, “A novel S-box of AES algorithm using variable mapping technique,” 13 Int. Conference on Aerospace Sciences & Aviation Technology, May 2009, Cairo, Egypt, 1/9-10/9, 2009.

[14] K. Kazlauskas and J. Kazlauskas, “Key-dependent S-box generation in AES block cipher system,” Informatica, vol. 20, pp. 23-34, 2009.

[15] N. Hamdy, K. Shehata, and H. Eldemerdash, “Design and implementation of encryption unit based on customized AES algorithm,” Int. J. of Video & Image Processing and Network Security, vol. 11, pp.33-40, 2011.

[16] R. Hosseinkhani and H. S. Javadi, “Using cipher key to generate dynamic S-box in AES cipher system,” Int. J. of Computer Science and Security, vol. 6, pp. 19-28, 2012.

[17] R. Merkle, "Fast software encryption functions," Advances in Cryptology: Proceedings of CRYPTO'90, Springer-Verlag, Berlin, pp. 476-501, 1991.

[18] K. Kazlauskas, G. Vaicekauskas, and R. Smaliukas, "An algorithm for key-dependent S-box generation in block cipher system," Informatica, vol. 26, pp. 51-65, 2015.

[19] D. Lambic, „A novel method of S-box design based on chaotic map and composition method"Chaos, Solitons & Fractals, vol. 58, pp.16-21, 2014.

[20] M. Khan, T. Shah, H. Mahmood, and M. A. Gondal, „An efficient method for the construction of block cipher with multi-chaotic systems," Nonlinear Dynamics, vol. 71, pp. 489-492, 2013.

[21] D. Lambic, „Security analysis and improvement of a block cipher with dynamic S-boxes based on tent map," Nonlinear Dynamics, vol. 79, pp. 2531-2539, 2015.

[22] F. Ozkaynak and S. Yavuz, „Analysis and improvement of a novel image fusion encryption algorithm based on DNA sequence operation and hyper-chaotic system," Nonlinear Dynamics, vol.78, pp.1311-1320, 2014.

[23] N. A. Rahman, A Course in Theoretical Statistics. Charles Griffin and Company, 1968.

Cultural Dimensions of Behaviors Towards E-Commerce in a Developing Country Context

Fahim Akhter
Department of Management Information Systems
College of Business Administration,
King Saud University, Riyadh, Saudi Arabia

***Abstract*—Customers prefer to shop online for various reasons such as saving time, better prices, convenience, selection, and availability of products and services. The accessibility and the ubiquitous nature of the Internet facilitate business beyond brick and mortar. The web-based businesses are required to understand the consumers' expectations, attitudes, and behavior across the globe and take into consideration of cultural effects. Saudi Arabia has become a highly potential lucrative market for web-based companies. However, the growing number of Saudi Internet users has not become leading online shoppers. It is important for web based companies to identify the barriers that are causing Saudi users to stay away from online shopping mainstream. This led to understanding Saudi culture, expectations, behavior, and decision-making process to promote e-commerce. The purpose of this study is to investigate the effects of Saudi Arabian culture on the diffusion process of e-commerce. The study addresses the cultural differences, risk perceptions, and attitude by investigating Saudi people about shopping online. An empirical study was conducted to collect the data from Saudi users.**

Keywords*—Electronic Commerce; Security; Culture; Online Shopping; Privacy; Saudi Arabia

I. INTRODUCTION

E-commerce is a globally accepted medium of conducting online business, and within a relatively short time, its services have risen to become a core element of the electronic business. The Census Bureau of the Department of Commerce declared that the estimate of U.S. retail e-commerce sales for the fourth quarter of 2015 was $89.1 billion, an increase of 2.1 percent from the third quarter of 2015 [1]. E-commerce sales in the U.S. are projected to reach $482 billion by 2018 [2], accounting for approximately 9% retail sales within the country. The number of digital buyers already reached 171 million in 2015 and continues to increase, with the total number of digital buyers projected to surpass 190 million by 2018. Credit and debit cards (73%) are the payment method of choice for U.S. online shoppers with digital payments (16%) increasing in popularity [2]. The growth of e-commerce is not in the U.S. alone, but the sign points towards continued growth globally.

There is now substantial evidence that the Internet has changed the way in which customers conduct online transactions in respect of their culture norms. This led us to assume that the success of the online business is subject to specific culture and norms. Due to it, web-based companies consider product localization to ensure business survival in the local market competition. The Chinese government and local online business do understand the importance of the benefits of e-commerce in terms of enhancing Chinese presence in the international business arena, strengthening business processes and channels, and forming better customer relationships [4]. Online trade in China skyrocketed by 120% [5] in the 2005 - 2015.

The researchers are agreed that a one of the main barriers facing the full deployment of e-commerce is the development of trust on the side of the consumer, particularly in developing countries [6]. The diverse characteristics of local environments and their cultures have created a significant level of variation in the acceptance of e-commerce. For example, a study [7] describes that Qatari people are vulnerable to e-mail phishing scheme. It has been found out that the country-specific factors, interests, beliefs, religion and personal characteristics are the main factor that causes Qatari citizens to become vulnerable to e-mail phishing attacks. The study has found that Qataris put too much trust in technology as compared to their own capabilities to detect email phishing.

Online vendors who target global customers could cultivate trustworthiness by acknowledging the norms and values of consumer culture. Nowadays, websites are accessed from any part of the world. Therefore, language, culture, and infrastructure issues top the list of online vendors [4]. It is beneficial for an online vendor to be accepted in a new culture being aware of the differences in language and customs that make up the culture in which they do business. It would be added value for vendors to understand the difficulties faced by customers to access their websites. Even though access to the Internet is very cheap in Europe and North America, some countries still charge heavy amounts for accessing the Internet. A report issued in 2001 by Human Rights Watch stated that many countries in the Middle East have been hesitant about allowing their citizens free access to the Internet. The report also notes that many countries in North Africa regularly monitor their citizen's access to the Internet and have taken steps to prevent the exchange of information outside their controls. In Pakistan and India, the respective governments use proxy servers to filter content. Some countries do not directly ban e-commerce, but do have strong local requirements that put extra pressure on vendors to compete in that segment. According to the report by U.S. Commercial Services on Buyusa.gov, the French government requires that an advertisement for products and services must be in French. Thus, a vendor who advertises and ships

products to France may have to offer a French version of his website. Vendors could show their respect and values to local consumers by adopting a local culture's norms in their business. Vendors could adopt local norms by observing local dialogs and their meanings, and symbols, with respect to their products and services. As mentioned on English-Zone.com, Pepsi's "come alive" advertising campaign did not achieve much success in China because its message came across as "Pepsi brings your ancestors back from the graves." Therefore, it is important for vendors to adjust the content of the website if the language of the website's content is different from the language of those who will be using it.

Different cultures respond differently to design, images, vocabulary, and color schemes. For example, in India, it is inappropriate to use the image of a cow in a cartoon or in a laughable setting. Indian culture has a religious status for cow. Muslim countries can be offended by an image that shows human arms or legs uncovered. Some consumers may not like a web page's color scheme if it clashes with their norms. For example, a web page that has large white elements can be offensive to Japanese consumers because the color white is symbolic of death in their culture. It is also inappropriate for websites with English content in Japan to use the word "four" because in Japanese this is "shi," which is also associated with the word for death. In China, the word "clock" is similar to the word for "death" and white, blue or black are associated with funerals. Websites in China could use red, yellow and pink because they see as pleasurable colors. Websites can attract local customers by avoiding conflicting images, vocabulary or color scheme because cultural norms and values affect consumers' motives and attitudes towards choices, intentions, and behaviors [4]. A rigorous assessment of aspects of online transactions is suggested as further work in view of the scarcity of empirical research in this area. This paper reports, research carried out in Saudi Arabia, which investigated consumer's attitudes towards using computers and the Internet and their Internet usage patterns

This paper is divided into six sections. After this introduction, Section II includes a brief literature review. Section III introduces the methodology adopted in this research. Section IV discusses and concludes the research.

II. Literature Review

Researchers [16] have suggested that there are several barriers besides security and privacy, which contribute to the adoption of B2C e-commerce: consumer resources, age, knowledge, lifestyle, educational level, attitude, motivation, marital status, personality, values, and cultural values. These factors could have different weights in influencing a consumer's buying decision, depending on the local culture and lifestyle of the consumer. For example, the attitude of consumers towards online shopping in developed countries such as the UK or the U.S. may be different from consumers from less developed countries such as Pakistan, Bangladesh, United Arab Emirates and Saudi Arabia.

Many studies on the impact of cross-cultural [7] have revealed major differences among cultures in respect of shopping online. Therefore, there may be a strong relationship between online shopping and culture which requires further study. It has been reported [8] by Saudi managers that lack implementation of management information systems (MIS) is one of the barriers for not adopting-commerce. Many of the managers of the local companies have admitted that if they were using the MIS systems, they would be indulging in e-commerce

III. Methodology

Since the nature of the information needed in this research is related to opinions, attitudes, and beliefs, the online questionnaire was best suited for collecting data. Therefore, a qualitative research method in the form of a web-based survey was used to explore and understand how the Saudi consumers perceive and evaluate the risks of e-commerce. The survey included one open-ended question at the end, which allowed the respondents to comment in their own words. Respondents were asked to express their opinions with a number of statements on their online shopping experience.

The online survey targeted at students consisted of 12 questions about their attitudes towards online shopping, problems they are facing and how the experience of online shopping and services could be improved. There were 1,491 participants who received via e-mail an invitation to take part in the online survey. In this message, there was a link to the questionnaire. The final sample consisted of 141 males and 104 females. The data were collected between January 11 and February 28, 2016. About 73.21 percent of the subjects were between 18 and 27 years of age. The majority of the respondents (61.23%) were studying at the undergraduate level and 98.12 percent had a computer and Internet access and were familiar with the online buying process. Web-based questionnaires were posted online along with a note that provided general information on the nature and the importance of the study and the significance of the contributions. Participants were assured of the confidentiality of their responses and promised a summary of the study results if they desired them. The data were collected through online surveys.

The instruments used in this research are adapted from the advice of fellow colleagues and experts in the research domain. The instruments used in this research were checked for validity, appropriateness, reliability, and accuracy. A number of steps were taken to meet this requirement. A panel of students, staff, and faculty at King Saud University were asked to answer the questionnaire and provide their comments on the wording and the content. Their feedback helped significantly in reforming the questionnaire. The analysis provided valuable suggestions such as shortening the questionnaire, changing some words, and using a common vocabulary. The researcher discussed the wording and the content of the questionnaire in detail with respondents.

IV. Findings

The collected data were analysed to extract Internet usage patterns. The composition of gender and its impact on Saudi Arabian society is rigorously analyzed. It is observed from the data that male respondents spend slightly more time (58%) on the Internet as compared to female respondents (42%). This means that men and women in almost equal numbers use the

Internet, with a marginally higher percentage of men going online. It shows that 56 percent of male respondents spend more than 11 hours a day on surfing the Internet as compared to 44 percent of female users. The data also reveal that male respondents hold 62 percent of full-time employment as compared to 38 percent females. It has been noted that respondents who spend more time on the Internet also hold a higher percentage of full-time jobs as compared to those who spend less time on the Internet and hold less percentage of full-time jobs

This study supports the interpretation that the gender that spends more time on the Internet in Saudi Arabia holds more full-time employment. This analysis indicates that users who spend more time on the Internet are likely to hold full-time jobs as compared to those who spend less time on the Internet. The other interpretation of this analysis might be that more users access the Internet from workplaces as compared to from home.

Figure 1 shows that online shopping is positively related to the time spent on the Internet. It has been observed that respondents who spend more time surfing on the Internet also tend to buy products from online vendors. This analysis tends to point that users who spend more time on online surfing probably have a higher tendency to become potential customers.

The majority of the respondents access the Internet from work. It has been observed that 62 percent of respondents access the Internet from work while 38 percent access from another location, such as college, hotspot, and home

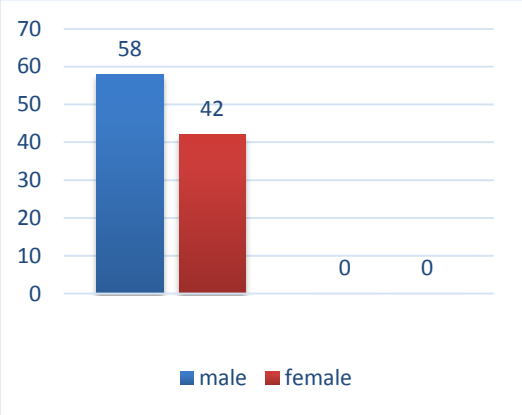


Fig. 1. Shopping Online

It is also noticed from the responses that users do not prefer to use Internet café facilities to access the Internet. This is a pretty interesting observation as Internet café is a popular option in other cultures such as in North America.

In has been observed that, Saudi Internet users are heavily using mobile devices to access social media platforms [Fig 2]. The most accessible social website, though mobile is Facebook (65.56%) and Twitter (33.73%). The statistics showed that users are inclined to explore online for entertainment, products, and services. There are many technical challenges needs to be addressed in Saudi Arabia such as enhancement and readiness of mobile services. Research [9] has argued that socioeconomic inequalities exist in Saudi Arabia with respect to the use of mobile government services, as more educated users are able to access to online services while less educated users could not utilize the online services.

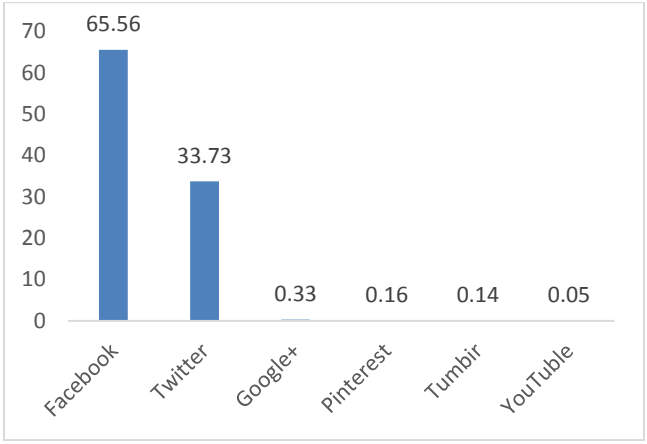


Fig. 2. Social Media access through Mobile (Jan 15 - Jan 16)

Respondents have chosen the Internet to complete diverse activities such as communication, education or being social [Fig 3]. It is important to address the elements that negatively influence the users to access the Internet such as lack of touching the product, customer service, security, trust, and negotiation [10]. The Saudi government has to address how to serve the needs of the educated users who access the Internet through mobile devices. This concern has been supported by the report from the World Bank (2012), that most Saudis are not able to avail the benefits of mobile government, as they are not aware of the complete usage of mobile devices. This challenge could be addressed through the awareness programs among the Saudis users about the benefits and usage of mobile commerce. The government may initiate steps to ensure the credibility of the system and provide assurance to the users that their data is safe and protected from adversaries. The government initiatives required a comprehensive legislative framework regarding cybercrimes, laws specifying the rights of citizens and responsibilities of data owner and data custodians.

The Saudi culture has a strong impact on users to shop traditionally such as touching the product to have a feel of it, visit the malls, spent time in the food court, negotiate the price and shop at the recommended businesses. These cultural events and values may indicate a lower rate of adoption of e-commerce by Saudi users compared to others culture. The cultural differences may also affect the rationale of Internet usage. The mindset of users from the specific cultures has been addressed by Hofstede [3] as “The collective programming of the mind which distinguishes the member of one human group from another. Culture, in this sense, includes a system of values. And values are among the building blocks of culture”. He further defines the value of culture by stating that it's a central component of cultural and a broad tendency to prefer certain states of affairs over others. The future study will measure the value of Saudi culture and compare with others by addressing five dimensions developed by Hofstede. These dimensions namely Power Distance, Uncertainty Avoidance, Individualism-Collectivism, Masculinity-Femininity and Long

versus Short-Term Orientation will reveal further factors influencing adoption of e-commerce in Saudi Arabia.

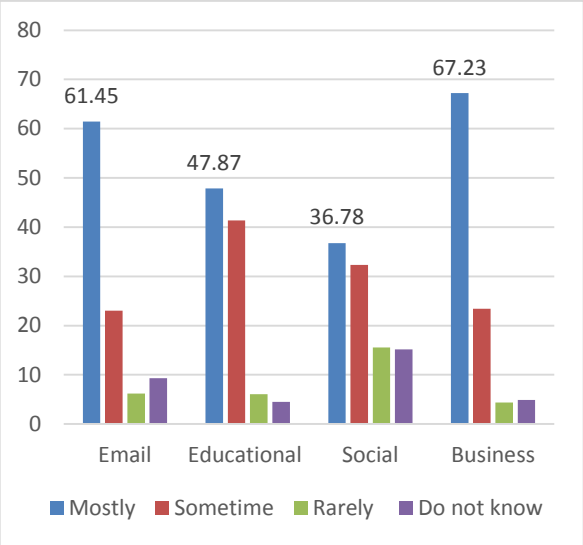


Fig. 3. Rationale for accessing the Internet

The issue which came to the fore was the lack of required IT infrastructure. Respondents have been noticed that the some of the barriers are not satisfied with the local service provider, such as Saudi Telecom & Communication, Mobily, and Zain. The users prefer continuous availability and faster bandwidth and stronger security in place to protect their personal data. The respondents have a concern about the cost to access the Internet. They have suggested that the government should intervene to minimize the cost of accessing the Internet especially tariff on mobile access. This will benefit users who mostly access the Internet through mobile devices.

V. Conclusion

This study explored the factors that affected the Saudi citizens with respect to online shopping. The Internet usage from the perspective of cultural anthropology was studied, focusing on its influence on Saudi society. There is a high level of Internet penetration among the Saudi Arabian population. The results of this study show that the participants who were using Internet services are displeased with the services. There is a lack of certainty amongst the respondents about the current security measures in place to protect their data and privacy. Furthermore, some respondents who thought that the current charges to access the Internet are expensive. Differences have been reported on attitudes toward Internet searching and usage patterns among women and men. Women and men have little differences in their general attitudes toward the perception of the Internet. Patterns of gender difference show that the Saudi male users are likely to have positive attitudes towards the Internet. They spent more time on the Internet and used the Internet more extensively. With respect to technology usage in the Saudi Arab, traditionally men are considered as 'high tech' and expected to take a lead role in managing technology related tasks. The barriers that influenced the acceptance of e-commerce are privacy, security, the cost of accessing the Internet, social values and incapability of physically examining the products.

Acknowledgment

The author extends his appreciation to the Deanship of Scientific Research at King Saud University, represented by the Research Centre at the College of Business Administration, for funding this research.

References

[1] U.S. Census Bureau News. Quarterly Retail E-Commerce Sales 4th Quarter 2015. Available at: https://www.census.gov/retail/mrts/www/data/pdf/ec_current.pdf (Accessed: 2 March 2016).

[2] PFS web. eCommerce Summary. Available at: http://www.pfsweb.com/pdf/global-ecommerce-book/2016-Global-eCommerce-Book-USA.pdf (Accessed: 5 March 2016).

[3] Hofstede, G.H. (1984). Culture's consequences, International Differences in Work-Related Values. Beverly Hills, CA: Sage Publications.

[4] Adel. A. Alyoubi, "E-commerce in Developing Countries and how to Develop them during the Introduction of Modern Systems" International Conference on Communications, management, and Information technology, Volume 65, 2015, Pages 479–483

[5] Hussain, Atiya. "E-Commerce and Beyond: Opportunities for developing country SMEs." International Trade Forum, no. 4. 2013.

[6] F. Meskaran and Z. Ismail, "Customers' trust in e-commerce: In collective culture setting," Information Retrieval & Knowledge Management, 2012 International Conference on, Kuala Lumpur, 2012, pp. 182-186.

[7] M. Al-Hamar, R. Dawson, and L. Guan, "A Culture of Trust Threatens Security and Privacy in Qatar," Computer and Information Technology, 2010 IEEE 10th International Conference on, Bradford, 2010, pp. 991-995.

[8] Moteb Ayesh Albugami, "E-Commerce and Economy: A Case Study of Saudi Arabia", International Journal of Information Technology, December 2015; Vol. 7 No. 2; ISSN 0973 – 5658 916

[9] A.Alssbaiheen & S. Love, "Exploring the Challenges of m-Government Adoption in Saudi Arabia" Electronic Journal of e-Government Volume 13 Issue 1 2015.

[10] Fahim Akhter, "Implement Fuzzy Logic to Optimize Electronic Business Success" International Journal of Advanced Research in Artificial Intelligence, 5(3), 2016

A New Network on Chip Design Dedicated to Multicast Service

Mohamed Fehmi Chatmen
Microelectronics Laboratory, Faculty of Sciences Monastir, 5000, Tunisia

Adel Baganne
Laboratory of Science and Technology Information, communication and knowledge, UBS University, BP 92116, 56321 Lorient, France

Rached Tourki
Microelectronics Laboratory, Faculty of Sciences Monastir, 5000, Tunisia

***Abstract*—The qualities of service presented in the network on chip are considered as a network performance criteria. However, the implementation of a quality of service, such as multicasting, shows difficulties, especially at the algorithmic level. Numerous studies have tried to implement networks that support the multicast service by adopting various algorithms to maintain the network average latency acceptable. To evaluate these algorithms, their performances are compared with the algorithms based on the multi-unicast. As expected, there is always a performances improvement. Regrettably, there is a possible degradation of latency introduced by such a service because of the large occupation of the network bandwidth for some period (which depends on packet size). In this paper, we propose an architectural solution aiming to avoid this possible degradation.**

***Keywords*—Network-on-Chip (NoC); adaptive routing; Quality of service; Multicast**

List of acronyms

TT: type of transmission
SN: sub-network
TSN: type of SN
N.B: number of blockings
TNB: threshold of NB
LRM: local router monitor
TP: type of packet
TF: type of FLIT
LNLRM: number of last addressed LRM
NLRM: the number of the LRM
P.length : the packet length
STA.PW: signal state for the WEST output port
STA.PE: signal state for the EAST output port
STA.PN: signal state for the North output port
STA.PS: signal state for the South output port
STA.PL: signal state for the Local output port

I. Introduction

Multicast service is a way to transmit the same packet from one source to multiple destinations; this implies an excessive use of network resources and a major occupation of bandwidth (occupancy of interconnections links). Several studies have attempted to reduce this use of resources through algorithmic solutions and sometimes through architectural solutions. The proposed algorithmic solutions have reduced the occupation. However, these solutions present complexities not only in implementation but also significant latencies at packet transmission due to the necessity of reading all the flits of the packet header.

The occupation of the bandwidth results in a rapid network saturation, which will degrade overall network performance. It will also be shown throughout this paper that the proposed architectural solutions consume too many resources.

In this paper, we present a new network architecture dedicated to multicast service, which aims to reduce the effect of rapid saturation of the network for an acceptable cost regarding additional resources. Our solution is based on a newly developed algorithm called “last addressed router". The latter avoids the fact that all routers forming the path must read the entire packet header.

II. Similar Works

Several studies have tried to propose algorithmic solutions to implement the multicast service. The purpose of these algorithms was to pass packets to multiple destinations using the minimum of branches (interconnections links) and to reduce the number of duplications of the transmitted packet to reduce the occupation of network bandwidth and the transmission cost. This will improve the average latency of the network, comparatively to the multi-unicast algorithm. These algorithms could be classified into two main types:

The first one is based on the principle of routing tree (tree based) [1]. This type of algorithm is suffering from an extra latency used to establish the tree before starting the sending of the packet, and the tree is found only if all its branches are free (the tree based path is reserved and ready for packet transmission). This algorithm demands that each router must implement a routing table (look-up table) which requires more resources to read and update the table. Finally, we can say that this type of algorithm is interesting in the use of the network bandwidth (fewer branches are used), but presents a considerable network average latency.

The second type is path based [2, 4, 5, 10, 11]; this type of algorithm is based on the Hamilton principle. The packet follows a path across the network with a low number of duplications. With this type of algorithm, we avoid both the live lock and the dead lock issues, but suffer from a significant latency especially for nodes that are distant from each other.

There are also other algorithms used to implement the multicast service that are based on the principle of RPM [6]. The principle of this latter is to divide the network into eight parts at the current router (the last router receiving the packet)

and to forward the packet to the various parts according to certain rules to optimize the number of duplications.

Some works have adopted architectural solutions for implementing multicast services. The author in [7] adopted a new architecture based on adding extra nodes called QAMC that aim to provide more interconnections links within the network to facilitate the multicast service implementation and other qualities of services namely BE and GS; the architecture is developed around a 2D mesh structure with the size of 4x4. It is noted that the architecture proposed is greedy for used resources: we have an extra of 40 unidirectional ports compared to the standard 4x4 dimension mesh 2D network Larger network would consume much more resources.

III. Proposed Network Architecture

A. Sub-networks

We propose to divide the network into a set of sub-networks (SN). For each sub-network, the multicast service is controlled by a node (router) called LRM (Local Router Monitor) which is the router with the largest possible number of connections (the most connected router in the SN), the structures of these sub-networks are defined as follows:

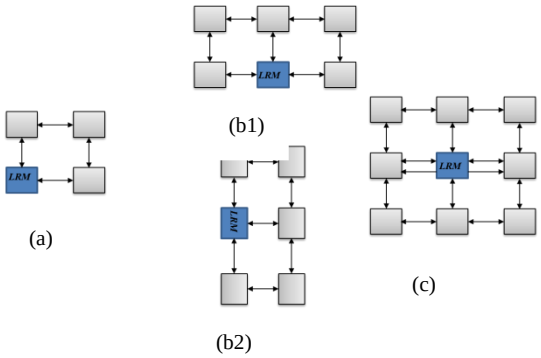


Fig. 1. different sub-networks

Any 2D-mesh network with any size can be constructed using the above structures.

Note the presence of additional ports at the structure of LRM router (c) and some neighboring routers for reasons that will be detailed later in this paper.

The router itself has three reference parameters, the first parameter indicates the router position in the global network, the second concerns its position in the sub-network and the third concerns the type of the sub-network, the number of ports associated with each router depends on these three parameters

B. Network structure

As is already mentioned, with the SN we can generate any 2D-mesh network. Figure 2 and Figure 3 show two examples of networks established using different SNs; they have respectively the dimensions of 5x5 and 6x6.

Note the presence of two additional ports at LRM (regardless to which SN it belongs), these ports are from neighboring LRM.

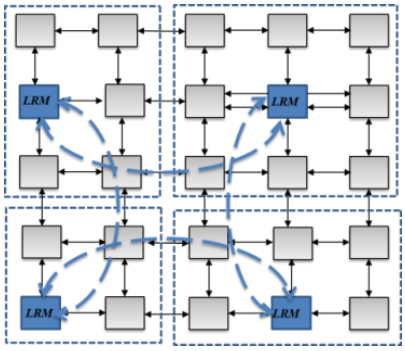


Fig. 2. the structures of networks with dimensions of 5x5 established using sets of the sub-networks (SN)

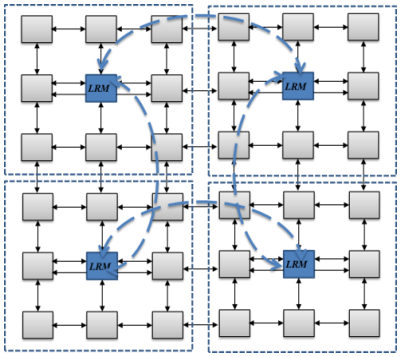


Fig. 3. the structures of networks with dimensions of 6x6 established using sets of the sub-networks (SN)

Actually, a packet must not cross all branches forming the ordinary path (following Hamilton algorithm) to reach its destinations, and then these ports can be considered as bridges between different (SN). These changes to the network structure are made for the multicast service. So a multicast packet can cross any branch of the SN while an ordinary packet can cross only the branches of the primitive network (2D-mesh network without considering the extra ports).

C. Setting up the router according to his SN

To identify each router forming the network, we have defined the following three parameters.

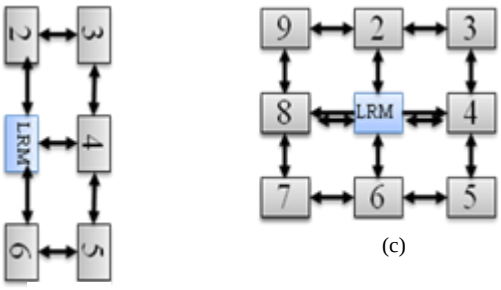
- its position coordinates (X, Y) within the network
- its position (SN Id) within the SN
- the type of SN (TSN) to which it belongs

1) Router position

(X, Y) Are the router coordinates within the global network . The routing unit considers only this parameter to transmit an ordinary packet.

2) Position of the router within its SN (SN Id)

Each router has an SN Id identifier (SN Id); it is a number indicating the position of the router within its SN. Indeed, when the SN Id is set to 1 then it is an LRM router; other routers are numbered from 2 to M (where M is the number of routers forming the SN). The SN Ids are assigned in clockwise order starting from the North as shown in Figure 4.



(b2) (c)

Fig. 4. The SN Id related to sub-network (b2) and (c)

Figure 4 shows the SN Id for routers belonging to the sub-networks (b2) and (c)

3) Type of sub-network

In addition to its SN Id, each router must be configured by its sub-network type (TSN). In fact, two routers having the same SN Id doesn't mean that they have the same number of ports unless they belong to the same type (TSN)

In fact, the TSN is a binary number defined by 3 bits, and each TSN has its unique TSN code, the table below shows the codes associated with the different types of (SN):

TABLE I. THE ASSOCIATED TSR CODE TO EACH SUB-NETWORK

Sub-network type	TSN Code
(a)	001
(b1)	010
(b2)	011
(c)	110

IV. ROUTING ALGORITHMS

There are four routing algorithms:

A. The routing of an ordinary packet

To route a regular packet, only the following bidirectional ports towards the directions (North, South, East, South, and Local) are used.

The algorithm routes the packet from the input port to the output port in one clock cycle based on the new instantaneous routing principle (described later in this paper).

The used algorithm is adaptive. Indeed, the packet can be delivered through any free port that minimizes as much as possible the number of HOPS. This may generate "live lock".

To avoid the live lock, we limited the number of packets passing by alternative paths (other than those specified by the XY algorithm) by a threshold called TNB. Indeed, for each transmission, a signal called NB is used to specify the number of times the packet has not followed the right path (according to XY) and in the case of equality between the NB and TNB, the routing algorithm is no more adaptive and the XY deterministic algorithm is applied.

1) Routing principle

Our router runs according to the handshake communication protocol. The switching mechanism adopted for our network is the Virtual Cut-Through to avoid deadlocks.

In the presence of a request, the routing unit reads the packet header, and it starts storing the packet in the memory unit (buffer). Meanwhile, the routing unit seeks whether the direction designated by the corresponding header is free (according to the XY algorithm) and that there are no other requests (with higher priority) designating the same output port, then it routes the packet through the multiplexing units (specifying the appropriate values for the input selection of the multiplexing unit) and reports that the port is unavailable to other possible requests, if not (the direction specified by the XY algorithm is unavailable), it seeks the availability of other ports, starting with the ports associated with the shortest path, to route the packet through.

In case all ports are unavailable, the packet already stored in the memory waits for the availability of the port designated by the XY algorithm. It will be favored by a higher priority to the recent requests (designating the same output port) (see Figure5)

If the sending is completed, the routing unit releases the data stored in the memory and turns the state of the associated output port to available. Note that all ports are initially in the available state.

Figure 5 shows the principle of the routing algorithm based on an example of a packet transmission from the local port to the east output port. Note that the NB signal is used to switch between the adaptive algorithm and the XY routing algorithm, which will be more detailed later in this paper.

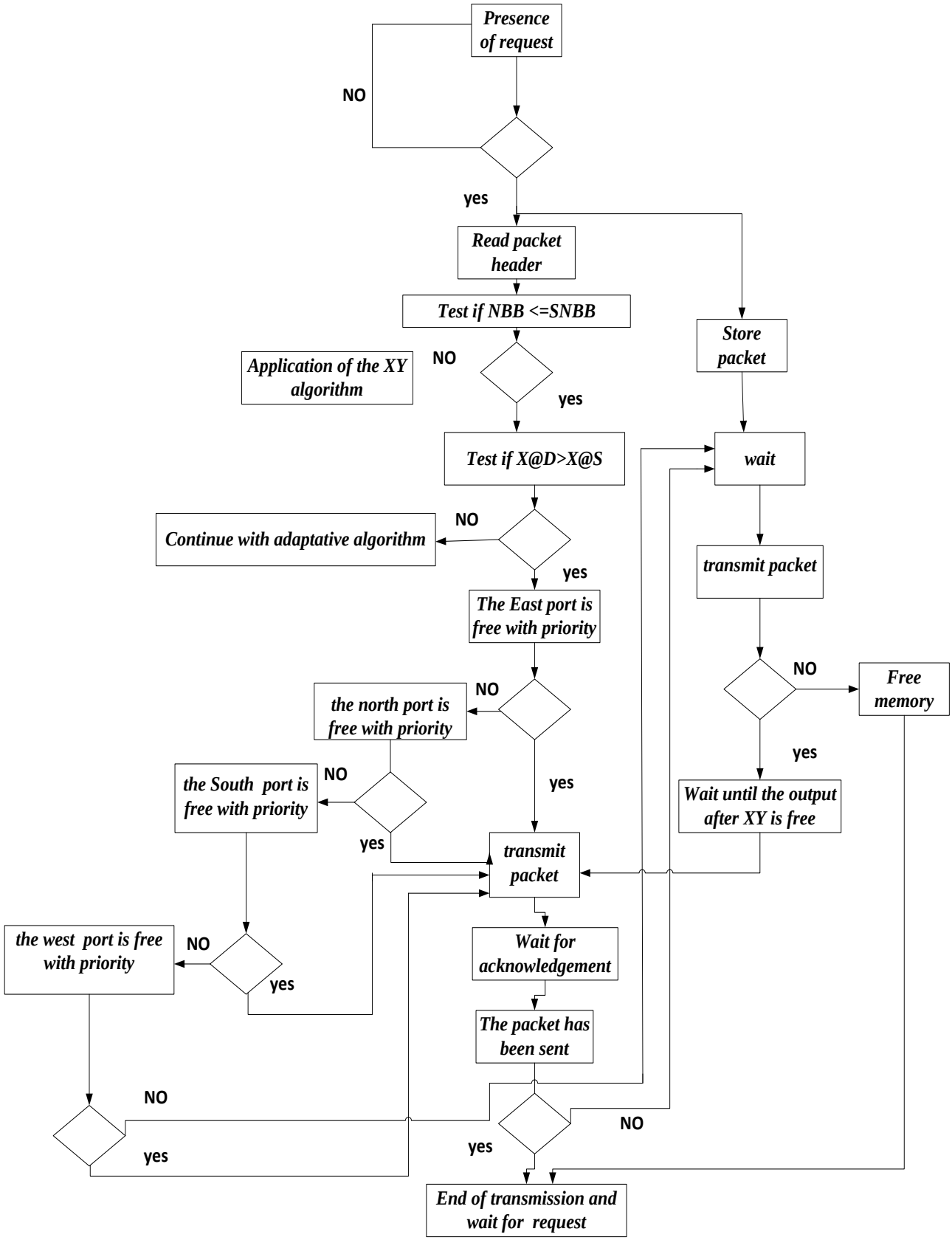


Fig. 5. Routing of an ordinary packet

The routing unit selects the output port instantly when reading the header. Indeed, a signal state (STA.P) is assigned to each output port that is set to 0 if the port is free and set to 1 if not. So we have five states signals, each of which is associated with an output port.

The states signals are STA.PW, STA.PE, STA.PN, STA.PS, and STA.PL, and are respectively associated with output ports following the directions WEST, EAST, NORTH, SOUTH, and LOCAL.

The signal state value is set to 0 when the port is free (when no packet is being transferred to the associated output port and no already stored packets are waiting for the port availability). Otherwise, the signal state value associated with the output port is set to 1.

The signal states are also used in the case of a multicast packet transmission.

TABLE II. SIGNAL STATES UNDER DIFFERENT TRANSMISSIONS SCENARIOS

STA.PW	STA.PS	STA.PN	STA.PE	X\{w}->w OR BX\{Bw}->w	X\{S}->S OR BX\{BS}->S	X\{N}->N OR BX\{BN}->N	X\{E}->E OR BX\{BE}->E
0	0	0	0+P	0	0	0	0
0	0	0+P	1	0	0	0	1
0	0	1	0+P	0	0	1	0
0	0+P	1	1	0	0	1	1
0	1	0	0+P	0	1	0	0
0	1	0+P	1	0	1	0	1
0	1	1	0+P	0	1	1	0
0+P	1	1	1	0	1	1	1
1	0	0	0+P	1	0	0	0
1	0+P (y ?)	0+P (y ?)	1	1	0	0	1
1	0	1	0+P	1	0	1	0
1	0+P	1	1	1	0	1	1
1	1	0	0+P	1	1	0	0
1	1	0+P	1	1	1	0	1
1	1	1	0+P	1	1	1	0
1	1	1	1	1	1	1	1

- 0: Port is in the available state
- 1: the port is unavailable
- 0 + P: the port is available and we favorite the transmission in his direction
- 0 + P (y?): Favors the transmission in the direction of the port if this minimizes the distance along the y-axis
- X \ {w}: set of input ports excluding WEST Input Port
- X \ {S}: set of input ports excluding SOUTH Input Port
- X \ {E}: set of input ports excluding the EAST Input port
- X \ {N}: set of input ports excluding the NORTH Input Port
- -> Ds : one of the entry ports is transferring data to the direction of the output port Ds
- BX: all buffers
- BX \ {BDe} -> Ds : one from the set of memories excluding the memory associated with the entry port BDe is transferring data to the output port Ds or waiting for the availability of the output port.

The table above shows the values associated with the different signal states by considering a packet transmission from the local input port and its packet header indicating a X@D coordinate greater than X@R (X@D and X@R are respectively the destination router and the current router coordinates following x-axis) in this case the routing unit facilitates the transmission of the packet to the EAST direction if its signal state is set to 0. If not, we send the packet according to the NORTH or to the SOUTH depending on their availability, but if both are available, the routing unit sends the packet to the direction that minimizes the distance between the destination IP and the current router according to the y-axis. In case both signals are set to 1, the packet will be sent to the WEST direction. If all signal states are set to 1, then the already stored packet waits for the availability of the EAST output port, this time with a higher priority compared to recent queries.

2) Priority of input ports and the elimination of live lock

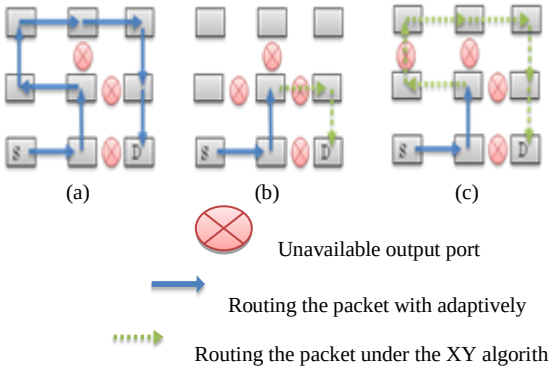
Our router uses the data stored in memory if there is a transmission error or all output ports are busy, otherwise, we can use any available output port to route the packets which

can generate live lock problems (the packet never reaches its destination). And to avoid this problem, a signal called NB is used to indicate to the addressed router the number of times the packet has not followed the path specified by XY algorithm in the already crossed routers. According to the value carried by the NB and compared to the threshold TNB. The routing unit decides which routing algorithm to use: either continues with the adaptive routing or just uses XY. The threshold depends on the network size (for example for the 3x3 network; the TNB is set to 3 and for the 4x4 network the TNB is set to 4).

The choice of the TNB was made after making some performances measurements that we won't consider in this paper, in fact, this algorithm which is a novel one, doesn't show better performances compared to the "look ahead" based ones.

We only adopted this algorithm because of its implementation simplicity and the performances improvement compared to the deterministic XY. In addition, the establishment of this algorithm is not considered as an aim in this paper.

Our network operates according to FIFO scheduling algorithm, the first input port addressing the output port will be served first. In some cases, multiple input ports address the same output port at the same time (called here instant requests), so the FIFO algorithm is not applicable, and the request having the highest NB number is the one who has the highest priority. In case where both requests have the some NB then an arbitrary priority assignment is considered. This priority is defined as follows: PE > PS > PW > PN > PL with Px is the priority associated with the input port x. Note that the local input port has the lowest priority and this for the simple reason that other packets coming from other directions are older.



(a) Behavior of the routing algorithm if the packet is carrying a number of two miss-routing compared to XY (NB = 2)
(b) Behavior of the routing algorithm if the packet is in a situation where all output ports are unavailable, in this case, the packet already stored in memory waits for the next port availability after the XY direction
(c) Behavior of the routing algorithm if the packet receives a number of three miss-routing(NB = 3) in this case, the NB is equal to the TNB then the XY algorithm will be applied even if there are better alternative paths.

Fig. 6. network behavior to avoid deadlock and live lock

3) Packet switching principle

Our router is actually operating in a virtual cut through packet switching type with one modification which is the storage strategy; in fact the router stores immediately the packet in the buffer regardless the state of the addressed port. This does not mean that we must save the entire packet before sending it, but it will be sent as soon as the destination port is available. The instant storage was necessary for two reasons:

The first one is to have two packets addressing the same output port at the same time (the routing unit sees that the output port is free and gives access to both input packets) this will result in the loss of data of one of two packets. In such case, the routing unit detects this error immediately (the routing unit always tests if it has given access to an output port to multiple input packets) and stops sending the one with lower priority without risk of losing data because they are already stored in memory. Those two actions operate as combinatorial functions.

The second reason is to have a sending error (receiving an error acknowledgment from the destination router) in this case the routing unit starts to resend the packet as soon as the output port is available.

The communication protocol adopted for our network is the handshake: In the presence of a request, the routing unit sends an acknowledgment (ack = 00) to indicate the receipt of the header and then sends (ack = 01) to indicate the beginning of the receipt of the rest of the packet. It sends an acknowledgment (ack = 11) to indicate the successful receipt of the entire packet. An acknowledgment (ack = 10) is sent in case of transmission error.

The treatment of these two communication obstacles was not specified by other on-chip network designers (it is imperative to store the packet for each communication to ensure that no packet will be lost).

B. Routing multicast packet

1) The last addressed router algorithm

The main objective of this paper is to introduce a new algorithmic solution to avoid the fact that all the routers forming the path for a multicast transmission have to read all the header flits. In fact, the router can forward the packet only after reading the packet's header which means a loss of a large number of cycles for redundant operations. That's why we propose a source routing algorithm that only, in our case, defines the last router address concerned by the multicast packet. So, all the routers forming the path must check only the first flit of the header (this flit has been established at the source node) before redirecting the packet to its destination. Thanks to the principle of the instantaneous routing principle presented in 4.1.1, one cycle is sufficient to transmit the multicast packet between two adjacent routers.

For a better understanding of the importance of this novel algorithm, we consider a simple path-based algorithm presented by [9], this algorithm specifies (at the source node) all the designated routers by the multicast packet in one direction (the multicast packet is instantaneously and horizontally transmitted in two directions). The author of [9] has used a method to avoid the redundant operation of reading the entire packet header but there is still redundancy in reading some of the header flits by the routers that form the path as shown in Figure 7.

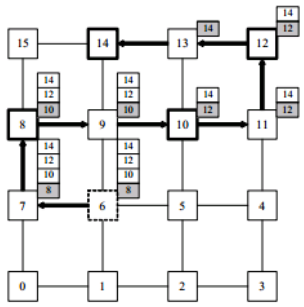


Fig. 7. Transmitting multicast packet header to establish a connection between the destinations nodes having labels greater than the source router (the label 6) using the Hamiltonian routing algorithm. [9]

The example studied by [9] shows that the routers with label (6, 7, 8) read the same packet header used to establish connections to the destinations nodes having labels greater than the source router. The packet header format is presented by the Figure 8.

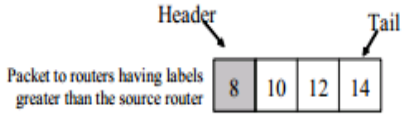


Fig. 8. Connection establishment packets format [9]

By adopting the novel algorithm proposed in this paper (without considering the proposed structure based on sub-networks); we consider, at each line of the network, a source node that will specify only the last router that received the packet and belonging to the some line level. The source nodes are the routers receiving the packet vertically (from the south or the north port), in this case, we have simply reduced the number of routers that should read all the packet header flits to only one at each network line as shown in Figure 9.

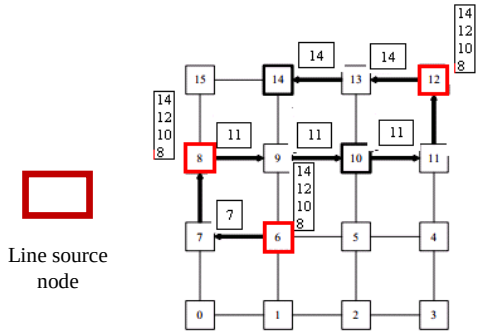


Fig. 9. Transmitting multicast packet header using the proposed algorithm in the case of the (Fig. 7) scenario. Note that source node at each line must read the entire packet header

Figure 9 shows that with the application of our proposed algorithm all the routers (including line source nodes) have to read a total of 18 header flits while 25 header flits have to be read adopting the method proposed by [9]

Reading the entire header packet at each line source node is an exhaustive operation that is why we have adopted the new network structure based on LRM routers (the number of LRM routers is, after all, lower than the number of lines presented in a 2D-mesh network)

2) Routing a multicast packet from a non-LRM router to neighboring routers

In this case, only the router that corresponds to the source node reads the entire packet's header and proceeds to the packet emission according to the most optimized path. If this latter is busy, it follows the path which is available. The source router specifies the last addressed router by the multicast packet and belonging to the same SN. It also specifies the router that will send the packet to the LRM in the case of a nine routers SN (see Figure 10).

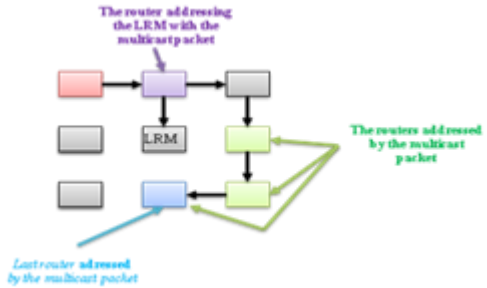
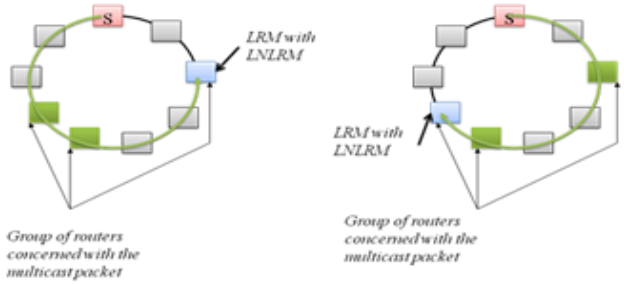


Fig. 10. Transmission of a multicast packet from a non-LRM router to neighboring routers (case TSN = 110)

For other types of SN, the source router specifies only the last router addressed by the multicast packet, assuring that the packet will be received by the LRM.

3) The routing of a multicast packet across different LRM routers

Actually, for this type of routing, a simple algorithm was adopted, this because the number of LRM is not defined in advance (it's not like in the case of SN). As the number of LRM depends on the number of SN, we have defined a parameter associated with each LRM called NLRM that defines a reference number of the LRM at the LRM network (The LRM network is defined in a ring topology to minimize the number of packet duplications). According to data carried by the FLITS forming the packet header, the LRM source defines the direction, and the NLRM associated with the last LRM addressed by the packet (which is called LNLRM);



(a) (b)

Fig. 11. Transmission of a multicast packet in LRM network (a) Transmission of multicast packet in a ring network formed by 9 LRM if the shortest route is busy (b) transmission of multicast packet in a ring network formed by 9 LRM if the shortest path is free

The information about the LNLRM is added to the header. Thus, LRM routers forming the path traversed by the packet don't read the entire packet header but only the first flit.

4) The routing of a multicast packet between an LRM and routers belonging to the same SN

This type of routing also depends on the type of SN, the different structures of SN are defined to transmit simultaneously two multicast packets in the network regardless which nodes are involved in the communication: for the 9 routers sub-network (Figure 1 (c)), the LRM has two paths for transmitting the packet, the first one is following the directions: from EAST to NORTH and from WEST to SOUTH, the second one is following the other ways (from EAST to SOUTH and from WEST to NORTH). That's for this reason that we added additional ports at the LRM of this type of SN. Such routing will only occur if the packet comes from the LRM network; we have two connections (bidirectional ports) between the LRM and the network of LRM, the first path is taken only if the packet comes from the first port (LRM NETWORK DATA IN PORT 1); the second path is taken if the packet is received from the second port (LRM NETWORK DATA IN PORT 2)

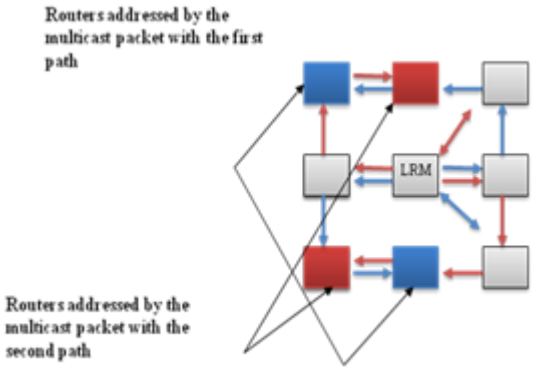


Fig. 12. Simultaneous transmission of a multicast packet between an LRM router and neighboring routers belonging to the same SN (TSN=110)

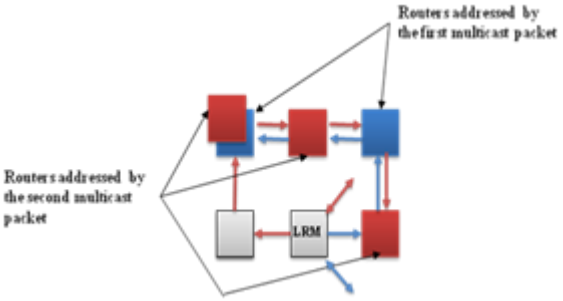
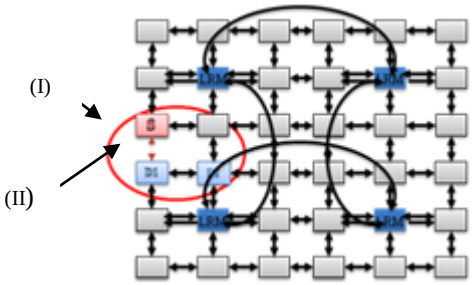


Fig. 13. simultaneous transmission of two multicast packets between a router and LRM neighboring routers belonging to the same SN (TSN=010)

For other types of SN we can transmit two multicast packets simultaneously without adding additional ports (see Figure 13); but the routing also depends on the port used for receiving the packet: if the port receiving the packet is the LRM NETWORK DATA IN PORT 1, the transmission direction of the packet is clockwise, if the port receiving the packet is the DATA IN NETWORK PORT LRM 2, routing is done in the other direction.
For both cases, the LRM always sets the latest routers addressed by the multicast packet; for SN (110), the signal L1R indicates the last router along the path to the north direction and the signal L2R indicates the last router along the path to the south direction, the two signals L1R and L2R are introduced at the FLIT level added by LRM in the packet header

5) Proposed structure's worst case

The worst case that our network may face is when a multicast transmission takes place only between neighboring routers belongings to different sub-networks (where short links can't be used).



(I): The source node R (0, 2) addresses the destinations nodes R (0, 3) and R(1,3) by a multicast packet, in this case, this direct link between R(0,2) and R(0,3) is useless
(II): Not used link for multicast transmission

Fig. 14. Worst case scenario of the proposed structure

In this case (Figure 14), the achieved latency is greater than most of other path-based proposed solutions. However, taking into consideration the ability for simultaneous transmission of multicast packets in the some sub-network and the ability of using the short links between the routers belonging to different sub-networks for other unicast transmission, this worst case situation is acceptable and doesn't prevent our proposed network to achieve better performances which is demonstrated later in this paper.

V. PACKET HEADER

The header of the packet differs depending on the packet type and the entering port. Indeed, there are two main types of packets

A. The case of a unicast packet

The header of this packet is defined by the following fields:

TT	X@S	Y@S	X@D	Y@D	P.length

TT: defined on 1 bit, it indicates the type of the transmission, in fact this bit is set to 0 if it is a unicast transmission and to 1 if it is a multicast transmission
P. length: this field is defined on 8 bits and indicates the packet size
X@S: set of 4 bits defines the coordinate along the X axis of the source router
Y@S: set of 4 bits defines the coordinate along the Y axis of the source router

X@D: defined on 4 bits defines the coordinate along the X axis of the destination router
Y@D: defined on 4 bits defines the coordinate along the Y axis of the destination router

B. Case of a multicast packet type

1) Packet coming from a local port

In this case, the packet header has not got any modification (no FLIT has been added)

TT	X@S	Y@S	P.length	Zone1	@{Ri}\1
T.F	Zone3	……	Zone N	@{Ri}\N	

With:
TF: defined on 1 bit, it specifies if the FLIT belongs or not to the packet header; this bit is set to 1 if the FLIT belongs and is set to 0 if not;
P.length: this field is defined on 8 bits it indicates the size of the packet without considering the added FLIT.
Zonei: This field is defined on 4 bits and indicates the SN addressed by the multicast packet.
@{R}\j: This field is defined on 9 bits and indicates the set of routers belonging to zone j and which are addressed by the multicast packet
The 9-bit of (@{Ri}\j) denote routers belonging to SN designated by the field j; each bit corresponds to a router and it is set to 1 if the router is addressed by the multicast packet and is set to 0 if not; this field is defined on 9 bits since we have at most 9 routers by SN.

2) Packet coming from a port other than the local port

a) Case of transmission of a multicast packet from an router other than LRM

In this case, we have a FLIT added to the packet header:

TT	T.P	LR	DR		
TT	X@S	Y@S	P.length	Zone1	@{Ri}
T.F	Zone3	……..	Zone N	@{Ri}	

With

TP: This field is defined on 2 bits, and it identifies the type of multicast transmission; in this case it is a transmission from a router other than LRM (TP is set to 01)
DR: sets of 4 bits and indicates which router drives the multicast packet to the LRM (this field is present only in the case that the SN is with a TSN equals to (110))
LR: Set of 4 bits and indicates the last router in the same SN as the source router which is addressed by the multicast packet

b) Case of transmission of a multicast packet from an LRM router to the LRM network

In this case, we have a FLIT added to the packet header:

TT	LNLRM				
TT	X@S	Y@S	P.length	Zone1	@{Ri}
T.F	Zone3	……..	Zone N	@{Ri}	

with

LNLRM: set of 4 bits that shows the last LRM addressed by the multicast packet

c) Case of transmission of a multicast packet from an LRM router to the sub-network (SN) to which it belongs

In this case, we have a FLIT added to the packet header:

TT	T.P	L1R	L2R		
TT	X@S	Y@S	P.length	Zone1	@{Ri}
T.F	Zone3	……..	Zone N	@{Ri}	

With

TP: this field is defined on 2 bits and identifies the type of multicast transmission; in this case it is a transmission from an LRM router to other routers belonging to the same SN (TP is set to 10)
L1R: set of 4 bits that shows the last router addressed by the packet by the north path
L2R: set of 4 bits that shows the last router addressed by the packet by the south path

VI. Example of a Multicast Transmission

Considering a Mesh network with size of 6x6 , the router with address (0,0) transmits a multicast packet to routers: {R (2,2), R (5,1), R (0.5), R (3,5)} and the router with address (5.5) transmits, after a few cycles, a multicast packet to routers: {R0,1), R (2,1), R (0, 4), R (1,3), R (5,2), R (4,3)}

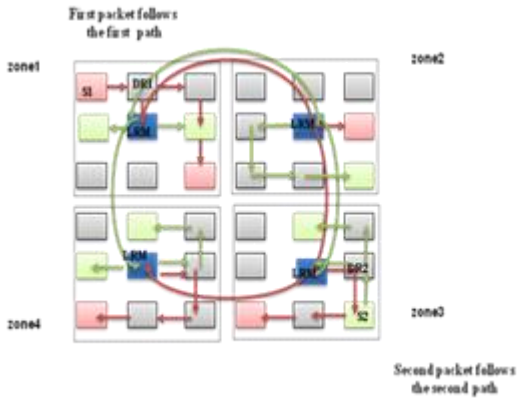


Fig. 15. Example of a transmission of two multicast packets at a network with size of 6x6

To understand how to forward two packets in the network, it is necessary to know the following information:

- The number of zones (sub-networks)

- The type of each sub-network*

the network with size of 6x6 is defined as a set of 4 sub-networks where each one is defined by a TSN = 110
Starting with the transmission of the first packet, the router R (0,0) transmits the first packet to the LR (last addressed router belonging to the same SN as R (0,0)) which is in this case, the router R(2.2), through the shortest path, and specifies the DR (the router that sends the multicast packet to the LRM which is the router R (1,0))
Once the packet is present in the LRM of the zone 1, the LRM transmits the packets into the LRM network according to the shortest path and marks the R(1,4) as the LNLRM (For our case there is no a shortest path because all the present SN are addressed by the packet).

The path adopted for the transmission of the packet is clockwise and as the LRM are placed in the network on the way that the LRM NETWORK DATA IN PORT 1 of each LRM is the entered port for such a path then the first path is taken at each SN receiving the first multicast packet. Indeed, each LRM must define the information of L1R and L2R of its own SN in the packet header before proceeding to transmission;

Indeed:

- For Zone 2: L1R = 0, L2R = 4
- For Zone 3: L1R = 0, L2R = 7
- For Zone 4 : L1R = 0, L2R = 7

For the second packet, the router R(5,5) transmits the packet to the LR, which is in this case, the router R(4.3) through the shortest path and specifies the DR which is the router R(5.4).

The LRM router associated with the zone 3 transmits the packet to other LRM and marks the router R(1.4) as LNLRM; this time, the transmission is counter-clockwise (the other direction is occupied by the first multicast packet), and the port LRM NETWORK DATA IN PORT 2 is the one that will introduce the packet at each LRM.

Finally, each LRM transmits the packet to its own SN after defining the information of the L1R and L2R. Indeed:

- For Zone 1: L1R = 4, L2R = 8.
- For Zone 2: L1R = 0, L2R = 5.
- For Zone 4: L1R = 1, L2R = 8.

VII. The Router

A. The router structure

The structure of our router is not too different from the classic one. In fact, we have the same components regarding buffers, routing unit, and port multiplexing units (called CROSSBAR).

The difference lies in the operating principle of these units; our router is designed to have a minimum latency. Routing the packet through the router is done by multiplexing units that operate according to combinatorial logic while the routing unit and memories operate at the clock edge. The routing unit allows instantaneously (at the front) to read the header of the packet and to deliver it to the destination port in the case of a unicast transmission.

The register present at each input port is used to synchronize between the establishment of the updated control signals used by the routing unit for the selection of the appropriate output port and the presence of input data at the entries of the CROSSBAR unit. In fact, these two operations are done at the same clock cycle, the CROSSBAR unit and the multiplexing units are combinatory components which explain the reduced latency of our router (1 cycle is sufficient to transmit data from the input to the output port) in the case of a unicast transmission.

In the case where it is a multicast transmission, the routing unit (associated with the source node router) stores the packet and reads the entire header before transmitting it to the output port; the difference between the transmission of a regular packet and the transmission of a multicast packet is the required number of cycles for reading the header. Also, for the transmission of a multicast packet, the routing unit modifies the packet header by adding a FLIT to facilitate the packet routing (see Figure 5). Indeed, the routing unit transmits the FLIT added in the first place, and then it proceeds to the transmission of the received packet.

In other works, the way of changing the packet header has not been specified; so our router not only allows packets transmission but modifies it if necessary.

Also, an ordinary packet can be routed directly from the input to the output port without using buffers, which is not the case for the multicast packets (we always speak of the router associated with the source node). This can be seen through the input port structure that is only reserved for the multicast service and the regular input port structure (reserved for both multicast and unicast services).

Except the source router of the multicast packet (the router that receives the packet from the outside of the network), the LRM leading the packet to the network of LRM, and the LRM that delivers the packet to different routers belonging to the same SN, all other routers intervening in the packet transmission, transmit the packet at the end of a single clock edge (it just read the FLIT added by the sources router or by the LRM routers).

For our architecture there are three types of routers:

B. Non-LRM router

This type of router has no connections with LRM network, so we don't have extra ports. We can say that the structure of this router is close to conventional router except that each output port is multiplexed from different input ports with an additional input coming from the routing unit itself (this one is used to add the extra FLIT to the packet header in case of a multicast packet) Figure 16 shows the router structure.

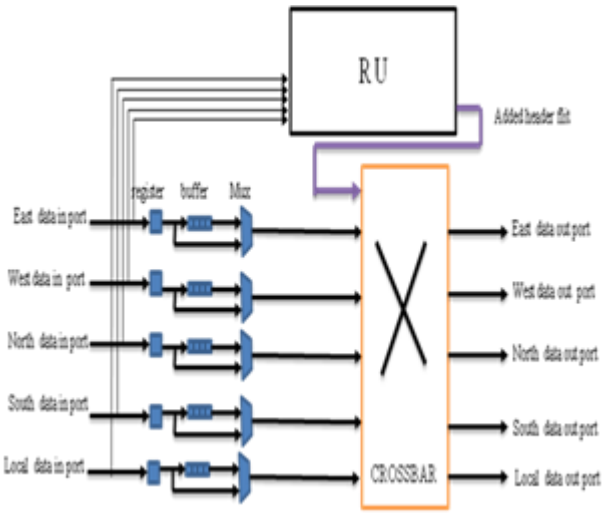


Fig. 16. Structure of router other than LRM

From Figure 16, we notice the presence of 5 inputs for each CROSSBAR instead of 4 (from 4 standard directions). The 5th port is the one used to introduce the additional FLIT.

C. LRM router with a TSN other than (110)

The difference between the structure of this type of router and the one previously presented is the presence of two additional ports. These ports provide communication between the sub-network to which the LRM router belongs and the network of LRM.

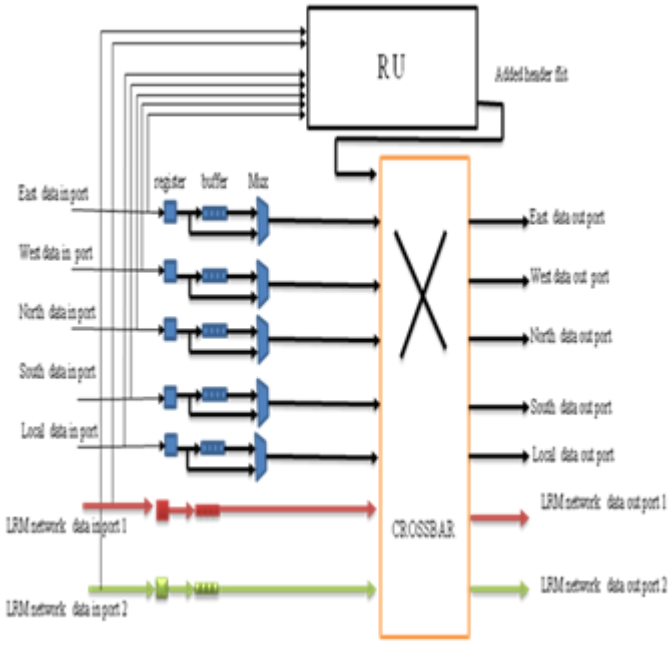


Fig. 17. structure of LRM router that belongs to an SN other than (110)

1) The structure of an input port reserved only to the multicast service

We associate to each input port used for multicast service:

•A register to synchronize between the arrival of the packet at the input port of the FIFO and the instruction of storing the packet in the buffer (if accepted) coming from the routing unit,
•A FIFO memory to store the packet, in fact the size of this memory is equal to the maximum size of the packet.

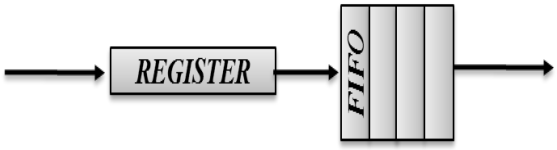


Fig. 18. Structure of input port reserved only to multicast Service

2) The structure of an ordinary input port:

we associate to each input port used for both multicast and unicast service:

•A register to synchronize between the arrival of the packet at both the input port of the FIFO and the input of the multiplexer unit, and the storing or the routing through the MUX instructions coming from the routing unit,
•A FIFO memory to store the packet,
•A multiplexing unit for selecting between the input coming directly from the input port or from the FIFO output; in the first case the packet is sent as it becomes available on the input port and in the second case we transmit an already stored packet.

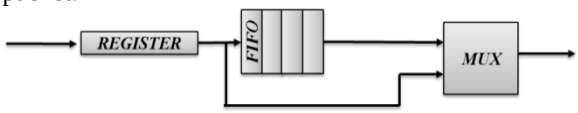


Fig. 19. the structure of an ordinary input port

D. LRM router with TSN (110)

This type of router is different from other LRM because of the presence of two additional ports following the EAST and WEST directions; these ports are unidirectional and are used to implement the proposed algorithms applied to the multicast packet.

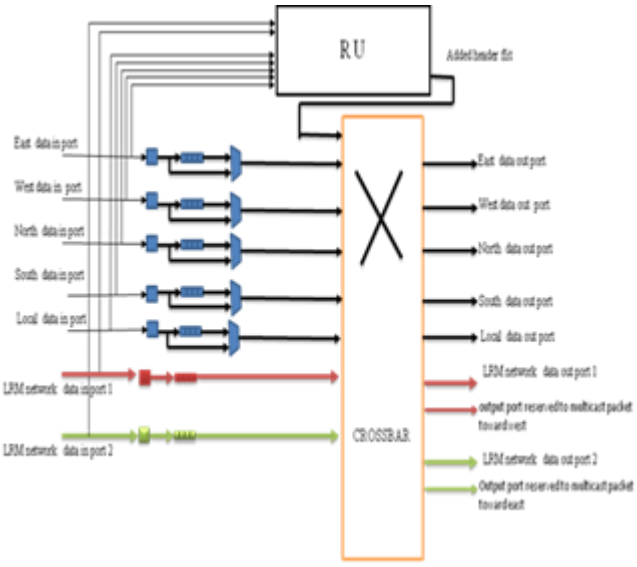


Fig. 20. LRM router with a TSN (110)

This router is greedy regarding employed resources. In fact, it has a total of 7 bidirectional ports and a couple of unidirectional ports. This router ensures the communication between the 9 routers of the SN with the LRM network.

VIII. Performance Analysis

To identify the performance of the proposed network regarding latency, we have measured the average latency of the network according to different injection rates and with a number of 10 destinations uniformly distributed at each SN.

We adopted for this, three type of traffic: uniform traffic, the bit complementary traffic, and transpose traffic. The percentage of traffic reserved for the multicast service is fixed to 20%. Figures 21 and 22 presented below show respectively the different results obtained by considering an 8x8 and 6x6 sized networks; these results were obtained by considering network simulations using modelsim6.5.

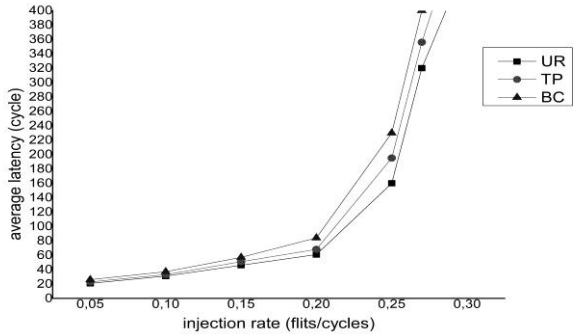


Fig. 21. The average latency under (UR, BC, TP) traffic for 6*6 sized network

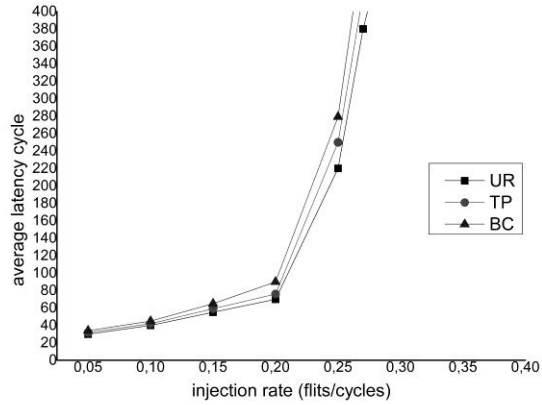


Fig. 22. The average latency under (UR, BC, TP) traffic for 8*8 sized network

The results obtained show considerable performance regarding latency, and these results are better compared to [9]; although the comparison can't be certain because there are several differences in network parameters, namely the router buffer size (we used 16 Flits sized buffer while they used 8 Flits sized buffer) and the routing algorithm adopted for the transmission of unicast packet.

Our network structure shows exceptional speed (this transmission is done in a single cycle through the principle of last addressed router) for the transmission of multicast packets by different routers (other than LRM and source router that should read the entire packet header). Nevertheless, this structure has two major weaknesses: the first one is the fact of sharing LRM network between routers; indeed, considering SN with a TSN (110) there are only two links to the LRM network which are shared between 9 routers. The second is that the LRM network allows only the simultaneously transmission of two multicast packet. These weaknesses have prevented our network to reach promising performances. The presence of these two weaknesses doesn't mean that our network structure is not suitable for multicasting service. We think that by adopting the path-based algorithm for the multicast service, our network, is by far, a respectable solution compared to most other works presented in the state of the art.

IX. Implementation on an FPGA Board

To get an idea about the consumption regarding the surface of the different network routers and the maximum frequency, we implemented these routers on an FPGA board VIRTEX 6 XC6VLX760 using the Xilinx ISE 14.1 tool. The obtained results are presented in the table below:

TABLE III. Area and Frequency of Our Different Routers

Type of router	LRM (010) (011)	LRM (001)	LRM (110)	R3	R4	R5
Number of Ports	6	5	9	3	4	5
frequency (MHZ)	33	203	15	340	32	19
area	14970	5332	37112	3963	8030	28419

With:

R3, R4 and R5 are the routers other than the LRM where R3 presents the routers having respectively 3, 4 and 5 input-output ports.

The LRM routers have been implemented by considering that it is a network with 4 LRM routers and with a size of 5x5. In fact, it is the same structure shown in Figure 2. The results showed a significant surface consumption which is explained by the presence of two adaptive routing algorithms for the different routers other than LRM; one for routing a regular packet and the other for routing a multicast packet. Higher consumption is obtained in case of a LRM router which depends on the type of LRM and the number of associated ports. Also, note that for LRM router, we have the presence of three routing algorithms: the first one for routing a unicast packet, the second used for transmitting the multicast packet to the LRM network and the last one for the transmitting the multicast packet in the associated SN.

X. Conclusion

The difficulty presented by the multicast service implementation is the need to optimize performance regarding number of packet duplication, the average network latency, bandwidth, and employed resources;

In this paper we have proposed a new network structure, based on the Mesh topology, to ensure acceptable performances with suitably added resources. We have also proposed a new routing algorithm called 'last addressed router' for preventing reading the entire packet header by all routers forming the path (here we talk about the Hamiltonian path) which has improved the network performance. Actually, this algorithm is suitable and simple to use for any network that operates with the "path-based" algorithm for the implementation of the multicast service, we have also shown the structure of our router allowing the implementation of this algorithm, and we have also explained how to change the packet header which has not been indicated by other similar works.

We tried in this new network structure to reduce the average latency by adding a few extra ports to create a network called LRM network, the results were not as expected (but it is still well performing compared to most other works). We succeeded to allow the multicast packet to cross the smallest possible number of HOPs to reach its destination. Nevertheless sharing LRM network between different routers of each SN did not allow us to have better performance. Our next work has for the main objective to implement this new structure using a new network topology (inspired from Mesh); with this new topology, we expect to get promising results.

REFERENCES

[1] N. Enright-Jerger, L.-S. Peh, and M. Lipasti, Virtual circuit tree multicasting: A case for on chip hardware multicast support, in Proc. Int. Symp. Comput. Architecture, Jun. 2008, pp. 229– 240.

[2] M. Daneshtalab, M. Ebrahimi, S. Mohammadi, A. Afzali-Kusha, Low distance path-based multicast routing algorithm for network-on-chips, IET Computer & Digital Techniques, vol. 3, no.5, pp. 430–442, 2009.

[3] L.Wang;P. Kumar,.; R.Boyapati; K.Hwan Yum; E. Jung Kim Efficient lookahead routing and header compression for multicasting in networks-on-chip in :Architectures for Networking and Communications Systems (ANCS), 2010 ACM/IEEE Symposium on

[4] X. Lin and L. Ni. Multicast communication in multicomputer networks. In : IEEE Trans. Parallel Distributed Systems, 4(10):1105–1117, 1993.

[5] R. V. Boppana, S. Chalasani, and C. S. Raghavendra, Resource deadlocks and performance of wormhole multicast routing algorithms, in: IEEE Transactions on Parallel and Distributed Systems, vol. 9, no. 6, pp. 535–549, Jun. 1998.

[6] L. Wang, Y. Jin, H. Kim, and E.J. Kim, Recursive partitioning multicast: a bandwidth efficient routing for Networks-on-Chip, in: Proceedings of the 2009 3rd ACM/IEEE International Symposium on Networks-on-Chip, IEEE Computer Society, 2009, pp. 64-73.

[7] S. Nambiar, K. Swaminathan, G. Lakshminarayanan and S. Ko, QaMC - QoS Aware Multicast router for NoC fabric, 2014 in : IEEE 27th Canadian Conference on Electrical and Computer Engineering (CCECE), 2014.

[8] M. Daneshtalab, M. Ebrahimi, S. Mohammadi, and A. Afzali-kusha. Low-distance path-based multicast routing algorithm for network-on chips. IET computers & digital techniques, 3(5):430–442, 2009.

[9] Carara, E.; Moraes, F. Deadlock-Free Multicast Routing Algorithm for Wormhole-Switched Networks-on-Chip. In: ISVLSI, 2008, pp. 341-346.

[10] M. Ebrahimi, M. Daneshtalab, P. Liljeberg, and H. Tenhunen, "HAMUM - A novel routing protocol for unicast and multicast traffic in MPSoCs," in Proc. Euromicro Int. Conf. on PDP, 2010, pp. 525–532.

[11] X. Lin, P. K. McKinley, and L. M. Ni, "Deadlock-free multicast wormhole routing in 2-D mesh multicomputers," IEEE Trans. Parallel Distrib. Syst., vol. 5, no. 8, pp. 793–804, 1994.

AUTHORS' PROFILE

Mohamed fehmiChatmen received his M.S. degree in intelligent and communicating system from the engineering school of Sousse, Tunisia, in 2011. Currently, he is a PhD student. His research interests include Network-on-Chip concept and design methodology, multimedia application

Adel Baganne received his M.S.degree and the Ph.D degree from Rennes University (France) in 1994 and 1997, respectively. He is presently an Associate Professor at the UBS University since 1998 and a member of the LESTER-CNRS Lab. He has published a number of research papers in the area of computer architecture and SoC design. His research interests include NOC design, communication synthesis, High level synthesis and CAD tools.

Rached Tourki received the B.S. degree in Physics (Electronics option) from Tunis University, in 1970; the M.S. and the Doctorat de 3eme cycle in Electronics from Institut d'Electronique d'Orsay, Paris south University in 1971 and 1973 respectively. From 1973 to 1974 he served as microelectronics engineer in Thomson CSF. He received the Doctorat d'état in Physics from Nice University in 1979. Since this date he has been professor in Microelectronics and Microprocessors with the physics department, Faculté des Sciences de Monastir. His current research interests include: Digital signal processing and hardware software codesign for rapid prototyping in telecommunications

New Artificial Immune System Approach Based on Monoclonal Principle for Job Recommendation

Shaha Al-Otaibi
College of Computer and Information Sciences
Princess Nora bint Abdulrahman University
Riyadh, Saudi Arabia

Mourad Ykhlef
College of Computer and Information Sciences
King Saud University
Riyadh, Saudi Arabia

***Abstract*—Finding the best solution for an optimization problem is a tedious task, specifically in the presence of enormously represented features. When we handle a problem such as job recommendations that have a diversity of their features, we should rely to metaheuristics. For example, the Artificial Immune System which is a novel computational intelligence paradigm achieving diversification and exploration of the search space as well as exploitation of the good solutions were reached in reasonable time. Unfortunately, in problems with diversity nature such job recommendation, it produces a huge number of antibodies that causes a large number of matching processes affect the system efficiency. To leverage this issue, we present a new intelligence algorithm inspired by immunology based on monoclonal antibodies production principle that, up to our knowledge, has never applied in science and engineering problems. The proposed algorithm recommends ranked list of best applicants for a certain job. We discussed the design issues, as well as the immune system processes that should be applied to the problem. Finally, the experiments are conducted that shown an excellence of our approach.**

Keywords—content-based filtering; computational intelligence; artificial immune system; clonal selection; monoclonal antibodies

I. INTRODUCTION

The Job Recommender System (JRS) has emerged in e-business online services in recent years. While companies post their jobs on online portals, an applicant uses them to establish his/her profile. A huge volume of job descriptions and applicant profiles are becoming available online. The need increases for applying the recommender system technologies that can support recruiters to handle the huge online information efficiently [1], [2].

Additionally, we notice the diverse nature of job specification that should be considered in candidates/job matching. Certain job requirements should be modeled in many forms to meet a diverse set of candidates that satisfy job requirements. We need to determine the job's requirements by a set of vectors of different features that can meet most possible appropriate applicants, and then rank applicants depending on the matching degree between the applicant and job requirements. We proposed an algorithm for JRS that used the traditional Artificial Immune System (AIS) paradigm [3].

The antibodies production in AIS algorithm used the natural response of the adaptive immune system that similar to the production of Polyclonal Antibodies (PAbs) in laboratories where the population of antibodies produced from multiple B-cells that have been activated by the immune response. The immune response to an antigen generally involves the activation of multiple B-cells all of which target a specific antigen. Consequently, a huge number of antibodies are generated with different affinities for that antigen. The antibodies and antigens in AIS algorithm represented by a vector composed of set of features that represents the job and applicant profiles.

Unfortunately, when we handle a problem such as job recommendations with a diverse nature in their features, the traditional AIS produced a huge number of antibodies that affects the system efficiency. Hence, The AIS applied the mutation strategy to perform the diversity topic in the previous algorithm. To leverage this issue, we will apply the AIS with Monoclonal Antibodies (MAbs) production principle that produces a population of antibodies from a single B-cell that recognize the antigen. However, the discovery of monoclonal antibodies has encouraged a revolution in medicine that is probably only second to the discovery of vaccination [4].

The algorithm that we will propose in this article is named Monoclonal Artificial Immune System for Job Recommender System (MCAIS-JRS). It recommends ranked list of best applicants for a certain job. We will represent each job's feature by a single antibody that tries to recognize the epitope of the antigen (applicant's feature) instead of using single antibody to represent a set of features. Hence, certain antigen has many epitopes surfaces that the antibody tries to recognize a single epitope at a time. Additionally, we propose a diversity operator to be used instead of mutation, and maintain a diversity pool of alternative solutions. Although, the proposed algorithm covers a wider range of problems that have numerous features with different alternatives such our considered problem.

This article is organized as follows: Section II demonstrates the related work. Section III presents the artificial immune system concepts used as basic technique in our proposed algorithm. Section IV illustrates the different issues related to modeling job recommendation in AIS using monoclonal antibodies principle. For example, the representation of both antibody and antigen, the applying of the diversity topic, as well as the using of similarity measures. Section V details the description of the proposed algorithm. Section VI shows the experimental results and the scalability study. Finally, we conclude our findings of this article in Section IVI.

II. Related Work

In recent years, much research has been conducted to discuss different issues related to the applying of recommender system technologies in job problem [5]. A survey of job recommender system was presented [6]. It covered the job requirements such as user profiling and similarity measures. Additionally, we presented a comprehensive survey of job recommendation and listed the advantages and disadvantages of technical approaches in different job recommender systems [5]. Moreover, author of [7] determined that we must consider unary attributes such as individual skills, mental abilities and personality that control the fit between the individual and the tasks to be accomplished. Several recommender system techniques have been applied in job/candidates matching problem, started by the personnel selection approach [1] that developed a probabilistic hybrid recommendation approach for job/candidates matching. Then, their model utilized and extended by [7], [8], and [9]. The model that tries to use the suitable personality traits and key specialized skills through information statistics and analytic hierarchy process was presented [10], [11].

Fazel-Zarandi and Fox combined different matchmaking strategies in a hybrid approach for matching job seekers and jobs using logic-based and similarity-based matching [12]. Additionally, the PROSPECT, which is a decision support tool assisting recruiters to shortlist candidate resumes list. It mines resumes to extract features of candidate profiles such as skills, education, and experience. It used IR techniques to rank applicants for a given job position [13]. The recommendation problem treated as a supervised machine learning problem [14]. This system recommends jobs to applicants based on their past job histories, in order to facilitate the process of choosing a new job. It trains a machine learning model using a large amount of job transitions extracted from person profiles available in the web.

Moreover, the TalentMatch system used at LinkedIn, in which the recommender system activated by a job posted on the site, scans the entire database to find the best candidates for the job. In this system, the semantic model computes the probability that the feature vector representing the candidate and the feature vector representing the job are a good match [15]. The hybrid recommender system that uses the job and user profiles, as well as the activities undertaken by users, in order to produce personalized recommendations of candidates and jobs. The generated data is modeled using a directed, weighted, and multi-relational graph, and the ranking algorithm [16]. Finally, candidate's profile and candidate's job preferences have been used to predict job recommendations. First, rules predicting the general preferences of the different user groups are mined. Then, the job recommendations for target candidate are made based on content-based filtering as well as candidate preferences, which are conserved either in the form of mined rules or obtained by candidate's own applied jobs history [17].

III. Artificial Immune System

The immune system has started to the emergence of AIS as a novel computational intelligence paradigm in the 1990s. A number of AIS models were existent, and they are used in pattern recognition, optimization, computer security, fault detection, and many other applications [18]. The AIS can be described as metaphorical computational systems developed using ideas, theories, components, and process derived from the immune system. It has two parts, the innate and adaptive immune systems [19]. The *innate immune system* is a stable mechanism that perceives and destroys specific invading organisms, whereas the *adaptive immune system* responds to anonymous foreign invader and provides a response that can persevere in the body over a long period of time. The *adaptive immune system* is comprised of a collection of different cells accomplishing different functions that spread over the body. The B-cells are the primary cells that responsible for the generation and secretion of antibodies using specific proteins which binding the antigen. Each B-cell can only provide one particular antibody. The antigen is located on the surface of the invading organism, and the binding of an antibody to the antigen is a signal to kill the invading cell [19], [20].

In general, there are many theories for the immune system, we will focus in this article on the clonal selection and mutation strategy that are primarily used in AIS models. The *clonal selection* is an algorithm used to define the basic features of an immune response to an antigenic stimulus. It is defined the idea that only those cells that recognize the antigen are proliferate. The mutation process is the most commonly used in AIS models; it means random genetic changes to the genes of the cloned cells. These changes caused proliferation and variation of high affinities antibodies. This concept was used as the basis for mutation in clonal selection algorithm, where the mechanism was the affinity between the antibody and the antigen. Additionally, the mechanism is often enhanced by the somatic hypermutation that is noticed as a process for optimizing the binding affinity of antibodies [21].

Moreover, there are two ways to produce antibodies in laboratories. Hence, antibodies include those produced by a combination of various B-cells, termed polyclonal antibodies and those secreted by a single cell of B-cell, termed monoclonal antibodies; both types have become crucial instruments in fundamental immunological research, diagnostic testing, immunohistochemistry, and vaccine quality control. PAbs and MAbs can be used for these purposes, although the production of these antibodies requires the use of a large number of animals. *PAbs* represent a population of antibodies collected from multiple B-cells that have been activated by the immune response of an immunized animal.

The immune response to an antigen includes the activation of multiple B-cells all of which target a specific epitope on that antigen. As a result, a huge number of antibodies are generated with different epitope affinities. However, *MAbs* represent a population of antibodies that recognizes a single epitope within an antigen, and produced by a single B-cell.

Authors of [22] discovered that this cell can be immortalized by fusion with myeloma cells, resulting in set of cells that are able to produce unlimited quantities of monoclonal antibodies. Since the Nobel prize winning work of these researchers, MAbs have become essential treatment in basic research as well as in diagnostic testing and medical

treatments. The PAbs and MAbs production processes are illustrated in Fig.1 that adapted from Biotech Resources[1].

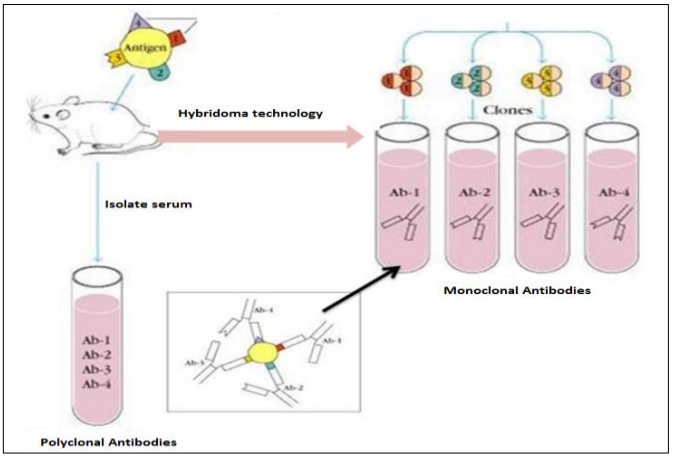


Fig. 1. The procedure of PAbs and MAbs production

Using the MAbs production principle that produces a population of antibodies from a single B-cell to recognize the antigen will leverage the previously mentioned issue in the traditional AIS [3]. However, the antibodies and antigens in traditional AIS algorithm represented by a vector composed of set of features that represents the job and applicant profiles. Unfortunately, when we handle a problem such as job recommendations that has too many features with a diversity in their nature, the traditional AIS produced a huge number of antibodies that affects the system efficiency. Hence, the AIS applied the mutation strategy to perform the diversity topic in the previous algorithm.

IV. Modeling Job Recommendation in AIS Using Monoclonal Antibodies Production Principle

A. Antibody and Antigen Representations

Choosing a suitable representation is very important for the algorithm's success. In JRS, there are many features that should be represented to model the job's requirements as well as the applicant's resume. For example, a set of qualifications and their information, skills, languages, experience, etc. The antibodies will represent certain job's profile. While, the antigens represent all applicants' profiles in the system. In MCAIS-JRS, an antibody indicates a specific feature that tries to recognize single epitope of the antigen. Hence, the antigen has many epitopes surfaces that the antibody tries to recognize a single epitope at a time. A possible mapping of the antibody's representation to vector space is

(<AB-CODE>, <AB-FEATURE >, <AB-STRENGTH >, < AB-FEATURETYPE>)

where <AB-CODE> gives a certain code for the current antibody, <AB-FEATURE> is value of a certain job's feature, <AB-STRENGTH> represents the antibody strength, it counts how many certain job's feature recognized. Hence, each recognized feature would increase the strength operator for current antibody. Finally, the <AB-FEATURETYPE> uses to distinguish a certain feature from other features when comparing it with antigen's features.

Example 1 Let us consider the following antibody's vector: (3, 2, 23, 7).

where,

3 is the <AB-CODE>,

2 is the <AB-FEATURE> that refers to the *Electrical Engineering* specialty,

23 is the <AB-STRENGTH>, it means the *Electrical Engineering* specialty found in 23 applicants' profiles, and

7 is the < AB-FEATURETYPE>, where 7 represents the <MSC SPECIALTY>, that means this antibody will be compared only with MSC specialty feature for the antigens.

On the other hand, the antigen represented by a vector space that contains a set of features. Each feature treated as an epitope that the antibody tries to recognize. A possible antigen's representation is:

(<APPLICANT-ID>,
(<PHD DEGREE> <PHD GPA> <PHD SPECIALTY>),
(<MSC DEGREE> <MSC GPA> <MSC SPECIALTY>),
(<BSC DEGREE> <BSC GPA> <BSC SPECIALTY>),
(<DIPLOMA DEGREE> <DIPLOMAGPA> <DIPLOMA SPECIALTY>),
(<LANGUAGE> <LANGUAGE LEVEL>),
<JOB CATRGORY>, <TYPE OF SKILL>, <EXPERIENCE>, <AFFINITY>).

Example 2 Let us consider the following applicant's vector,

(1,(0,0,0),(1,4.1,2),(1,4.02,2),(0,0,0),(1,3),1,5,0).
where applicant id = 1,
(0,0,0): no PhD degree,
(1,4.1,2): MSc degree with 4.1 GPA and spatiality number 2 (2 refers to *Electrical Engineering*),
(1,4.02,2): BSc degree with 4.02 GPA in *Electrical Engineering*,
(0,0,0): no diploma degree,
(1, 3): 1 refers to *English* language and 3 refers to excellent level,

1 refers to the OO programming skills,

5 refers to Experience years,

0 refers to the Affinity that represents the total distance between the current job (represented by a set of antibodies Ab_i) and any applicant (Ag_i). We will discuss how to calculate the total similarity in the following subsection.

B. Similarities Measures

The similarity or affinity is an important issue in building any AIS algorithm. The affinity is taken as the distance between a given antigen and the antibody. In MCAIS-JRS, we used mixed representations. Subsequently, the affinity is calculated using a combination of different types of similarity measures depending on the features types. Additionally, some of the parameters needs multi-level checking. It needs to determine if it is required by the recruiter for the recent job or

[1] http://www.bio.davidson.edu/molecular/MolStudents/01rakarnik/mab.html

not. For example, the certain degree that the applicant has may be it is not required for a certain job; in such case, this degree should not be scored. Then, the parameters distance is taken as a Boolean match 0, no match 1 as in

$$D_{K_j,i} = \begin{cases} 1, & \text{if } K_j \neq K_i \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

where $D_{K_j,i}$ represents the similarity between the antigen and antibody for specific parameter *K*, *j* represents the current antigen (Ag_i) and *i* represents the current antibody (Ab_i). For real attributes such as the GPA for certain degree, the similarity is calculated as a normalized distance between a given antigen and antibody.

$$D_{K_j,i} = \frac{-a(K_j - K_i)}{K_{MaxValue}} \quad (2)$$

Where $K_{MaxValue}$ is the maximum value for current attribute. For example, if the maximum GPA is 5 then the $K_{MaxValue}$ is 5. For example, if the applicant GPA is more than the job's required GPA it will be decreased the similarity (use *minus*). For integer attributes such as a certain specialty or skill, the similarity is also taken as (1). The total affinity for a certain antigen is calculated as

$$AFFINITY_j = \frac{\left(\sum_{k=0}^{L-1} D_k\right)}{L} \quad (3)$$

where *j* represents the current antigen (Ag_i), *K* is the current parameter and *L* is total number of parameters. The $AFFINITY_i$ takes values between [-1,1], where the similarity between the antibody and the antigen decreased from negative values to positive values. This means the $AFFINITY_{i,i} = 0$ is better than $AFFINITY_{j,i} = 0.99$.

C. Diversity Operator

The mutation process used in most AIS algorithms as main operator such as our previous algorithms in [3]. Although the exploration of search space achieved by using mutation process to reach higher affinity matches the invading antigen. It performs random changes to the cloned cells that are involved in controlling the antigen receptor. These changes caused proliferation and variation of high affinities antibodies. Therefore, a huge number of antibodies are generated with different affinities for that antigen. When, we handle a problem such as job recommendation that has enormous features with multiple alternatives, the mutation process will produce a huge number of antibodies that affects the system efficiency. To handle this issue, we apply the MAbs production in the AIS algorithm as well as the using of *Diversity Operator* instead of mutation to explore the search space.

The novelty here is that we propose a *Diversity Operator* to be used instead of mutation, and we maintain a diversity pool of alternative solutions. This works on each feature alone. Although, the purpose of the proposed algorithm is to cover a wider range of problems that have many features with more alternatives such as job recommendation problem. For example, some features such as specialties and skills have many possible choices that can be considered as possible valid solutions, and they will be added to the diversity pool. Consequently, a single antibody will be generated for each feature alternative and will be added to the antibodies pool. Figure 2 displays the antibodies' production in MCAIS-JRS for a certain job's profile and illustrates the *Diversity Operator* concept.

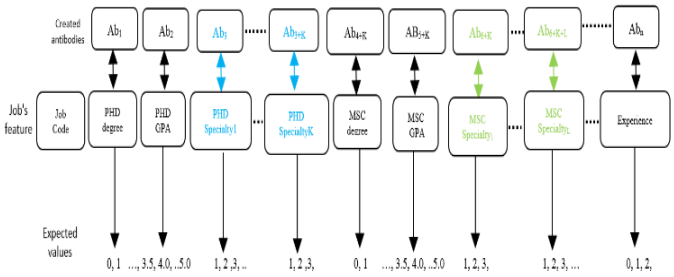


Fig. 2. Antibodies' production for a certain job profile

V. Monoclonal Algorithm for Job Recommender System

This section presents the design details of the proposed algorithm. We start by describing the metaphors and the parameters that used in MCAIS-JRS algorithm, where the antibody is a certain job's feature and the antigen is a vector of features for an applicant that considered as a target to be checked.

F	features list.
AB	antibodies pool.
Ab_i	current antibody.
n	number of antibodies.
f_k	certain feature.
f_i	current alternative for certain f_k.
AG	antigens pool.
Ag_j	current antigen.
Affinity	total distance between a given Ag_j (applicant) and antibody (Ab_i).
N	maximum size of antibodies pool.
Threshed	acceptance level of antigen.
Aga	set of accepted antigens.

The MCAIS-JRS starts by initializing some parameters such as the size of the antibodies pool (N) and the acceptance level of the antigen (Threshold). All antigens in the system will be exposed to the MCAIS-JRS algorithm steps in the same way. The algorithm performs one generation once all available antigens have been exposed to the antibodies cells, and all the MCAIS-JRS algorithm steps have been performed for each antigen. Figure 3 displays a diagrammatic representation of the MCAIS-JRS algorithm steps and Table 1 presents the pseudo code of the algorithm.

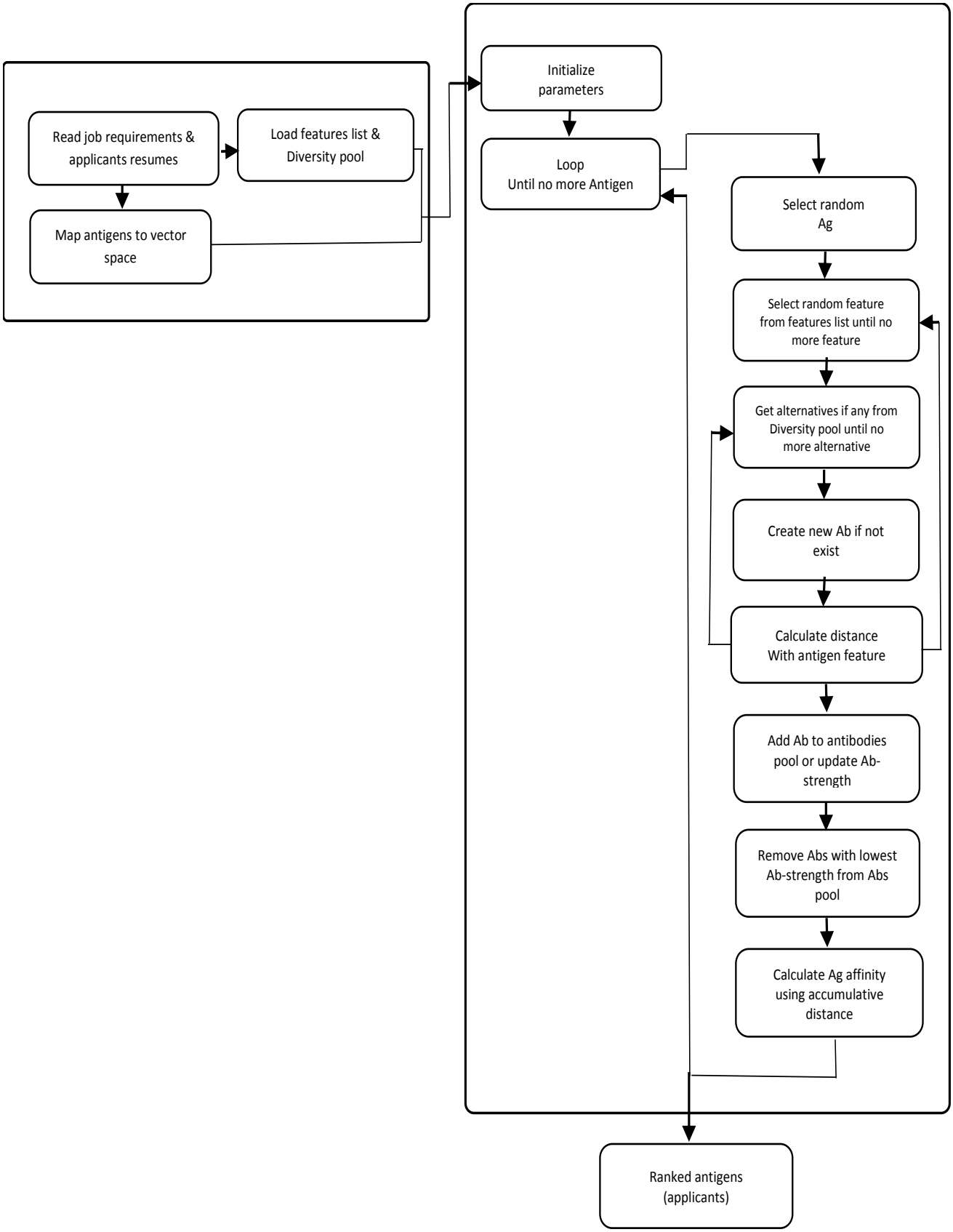


Fig. 3. A diagrammatic representation of the MCAIS-JRS

The following steps illustrate the MCAIS-JRS:

- The initial population is generated using the applicants' profiles (antigens set) and the job profiles (antibodies set).
- While the stopping condition is not met (no more antigens), the algorithm proceeds by performing a number of iterations to expose all antigens in the system. We use diversity operator to perform the job's diversification (e.g. specialties and skills).

1) The system selects random antigen (applicant), and the distance will be calculated for all antibodies (job's features) against the selected antigen.

2) Select a new feature from the job's features list (F).

3) Check the diversity pool that associated to the current job's profile and take one alternative for the selected feature if any.

4) Create antibody for the selected feature's alternative (Ab_i).

5) Each antibody represents a certain job's feature that match the same feature of the antigen (applicant). If match, then increase the < Ab_i-STRENGTH>. Repeat 3-5 until no more alternatives for current feature.

TABLE I. THE PSEUDO CODE FOR MCAIS-JRS ALGORITHM

Input

F: Features list of the current job,
Divf: Diversity pool for all features' alternatives, and
AG: List of antigens (set of all applicants).

Output

Aga: Ranked list of antigens (applicants).

Algorithm

```
1: Initialize parameters N, Threshold
2: For each Agj ϵ AG
3:     For each fk ϵ F                          // fk is a Job's feature
4:         fk ← OperateDiversity(Divf, fk)   // get alternatives
5:         For i=1  to Count(fk )
6:             If  Notfound(fi,AB)    //check current feature
alternative
7:                 CreateNewAntibody(Abi)
8:              If  Match (Agj, Abi)   using (1) &(2)
9:                 Abi-STRENGTH ++
10:                AB ←Update(Abi, AB)
11:            AB ← SortBySTRENGTH (AB)
12:            If  n > N
13:               Remove antibodies with lowest strength
14:                                  // until reach the antibodies pool size
15: For each Agj ϵ AG
16:        CalculateAffinity(Agj)   using  (3)
17:        If Affinity (Agj) ≤ Threshold
18:            Aga ← Add(Agj, AG)
19:  Aga ← SortByAffinity(Aga)
```

1) Repeat 2-5 until no more features.

2) If the total number of antibodies is more than the maximum size of the antibody's pool (N), the antibodies with lowest strength will be removed.

3) The antigen's affinity will be calculated as accumulative value from the distance with all antibodies. Then, the antigen's affinity is checked against a certain threshold that specified by the job's recruiter to determine the acceptance of an antigen.

4) The antibodies pool then taken to expose next antigen.

- Finally, the set of antigens (applicants) will be ranked depending on their affinities to produce Aga list.

VI. EXPERIMENTAL RESULTS

The MCAIS-JRS was constructed from the algorithmic outlines given in the previous sections. Therefore, the proposed recommender system integrated the monoclonal antibodies production principle and AIS processes, as well as, a diversity operator that used instead of the mutation strategy. This recommender system focus on the find good items task, where the job's recruiter is provided with a ranked list of qualified applicants who match the job's requirements. We collected data from resumes spread on the internet and generated the missing data. On the other hand, the jobs descriptions are realistic data gathered from LinkedIn and Monster websites.

A. An Illustrative Example

This example uses a simple job's profile with set of features. Minimal features were included in this example to simplify the concepts. Additionally, we use a diversity pool for alternatives specialties.

A design goal of MCAIS-JRS is to recommend the recognized antigens that represent the qualified applicants to the current job position. We will take the following job's profile: Job code = 2, MSC degree with 4.0 GPA and spatiality in Electrical Engineering, BSC degree with 4.0 GPA in Electrical Engineering, English language with excellent level, OO programming skills, and 8 experience years. Figure 4 presents the diversity pool of the specialties alternatives for this job.

Sp-Code	Name
1	Computer Science
2	Electrical Engineering
3	Telecommunication
4	Business Administration

Fig. 4. The diversity pool of specialties alternatives for the current job

The algorithm will produce the following antibodies for all features with their alternatives of the current job's profile.

Ab_i code	Feature value	Feature strength	Feature type
1	1	8	< MSC DEGREE>
2	4.0	5	< MSC GPA>
3	2	5	< MSC SPECIALTY >
4	1	10	<BSC DEGREE>
5	4.0	7	<BSC GPA>
6	2	8	< BSC SPECIALTY >
7	1	10	<LANGUAGE>
8	3	3	<LANGUAGE LEVEL>
9	3	1	<TYPE OF SKILL>
10	8	7	<EXPERIENCE>
11	1	0	Specialty alternative 1
12	3	3	Specialty alternative 2
13	4	2	Specialty alternative 3

Fig. 5. Antibodies pool of the current job,

Then, the algorithm proceeds to compute the distance between the antibodies and the current antigen. The following ranked list of applicants (antigens) with their affinities for the current job will be produced as Fig. 6.

Note that the parameters in Fig. 6, A_1, A_4, A_7, and A_{10} denote the PhD, MSc, BSC, and Diploma degrees respectively. They have *1* if the applicant has the degree or *0* if not. A_2, A_5, A_8, and A_{11}, represent the GPA for the different degrees. A_3, A_6, A_9, and A_{12}, denote the specialty of the different degrees and the values 1, 2, 3, and 4 in these columns represent *Computer Science, Electrical Engineering, Telecommunication, and Business Administration*. A_{13} and A_{14} used for the language and its level. In A_{13}, the value 1 represents the *English* language while 1, 2, and 3 in A_{14} column represent the *low*, *medium*, *and excellent* level respectively. Finally, A_{15} and A_{16} denote the skill and the experience. The different values in A_{15} represent the different skill types.

Ag_i	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9	A_{10}	A_{11}	A_{12}	A_{13}	A_{14}	A_{15}	A_{16}	Test	Affinity
0	0	0	0	1	4.1	2	1	4.0	2	0	0	0	1	2	3	9	☑	0.022381
8	1	4.0	2	1	4.0	2	1	4.0	2	0	0	0	1	3	5	12	☑	0.0714286
2	1	4.5	2	1	4.3	3	1	4.9	2	0	0	0	1	2	6	11	☑	0.0780952
4	0	0	0	1	4.0	2	1	3.5	2	1	3.0	3	1	3	6	12	☑	0.0785714
7	1	4.5	1	1	4.0	3	1	4.9	2	0	0	0	1	1	9	7	☑	0.115119
3	1	4.0	2	1	4.0	2	1	4.0	2	0	0	0	1	1	5	10	☑	0.1190476
5	0	0	0	1	3.8	4	1	4.0	2	0	0	0	1	3	6	3	☑	0.1193571
9	1	4.0	2	1	4.0	2	1	4.0	2	1	3.8	2	1	1	7	7	☑	0.1279762
6	0	0	0	0	0	0	1	3.5	3	1	3.1	1	1	2	8	12	☐	0.3166667
1	0	0	0	0	0	0	1	3.0	4	1	3.0	2	1	1	5	19	☐	0.347619

Fig. 6. The ranked list of applicants (antigens) with their affinities for the current job

B. Experimental Evaluation

For algorithm's evaluation, we applied the accuracy metrics that empirically measures how well a system can predict the most appropriate applicants for a specific job. For each posted job, there is a specific category that the active job related to it. Then, the algorithm generates recommendations using the similarity measures between job's features and applicants' profiles. If the algorithm able correctly to predict applicants who have the same job's category of the active job, then the algorithm is considered to perform well. However, for each job a ranked list of applicants is produced and evaluated such that whether those selected applicants match the same category for the active job. The accuracy metrics will be used to measure the frequency with which a recommender system makes correct or incorrect decisions about whether the selected applicant is appropriate.

In the experiment, we produce the recommended list of applicants that selected by MCAIS-JRS, then the result is evaluated and compared with our previous AIS algorithm [3]. In both algorithms, we check whether the applicants in the recommendation list from the same category of the active job. We applied the precision, recall, and F1 measures as details in the following paragraphs. The *precision* defined as the ratio of relevant items selected to number of items selected [23]. In our algorithms, it represents the ratio of relevant applicants selected to number of applicants selected.

$$P = \frac{N_{rs}}{N_s} \quad (4)$$

where N_{rs} refers to the number of relevant selected applicants and N_s refers to number of selected applicants. The *recall* defined as the ratio of relevant items selected to the total number of relevant items available [23]. In our algorithms, it represents the ratio of relevant applicants selected to the total number of relevant applicants. It denotes the probability that a relevant applicant will be selected.

$$R = \frac{N_{rs}}{N_r} \quad (5)$$

where N_r refers to the number of relevant applicants. Several approaches have been taken to combine *precision* and *recall* into a single metric. The *F1* metric that combines *precision* and *recall* into a single number has been used to evaluate recommender systems in [24].

$$R = \frac{N_{rs}}{N_r} \quad (6)$$

We perform the experiments according to different levels of similarity thresholds. The threshold levels determine which applicants will be included in the selected set of applicants depending on the affinity between the job's profile and the applicant. The experiments show that the MCAIS-JRS accuracy is high and more likely the accuracy of the traditional AIS. They show high accuracy when the threshold decreased. See Figures 7, 8,and .9, for precision, recall, and F1 respectively.

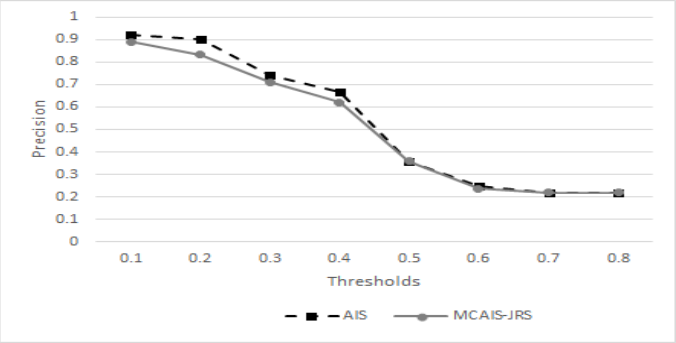


Fig. 7. The Precision rate of MCAIS-JRS and traditional AIS

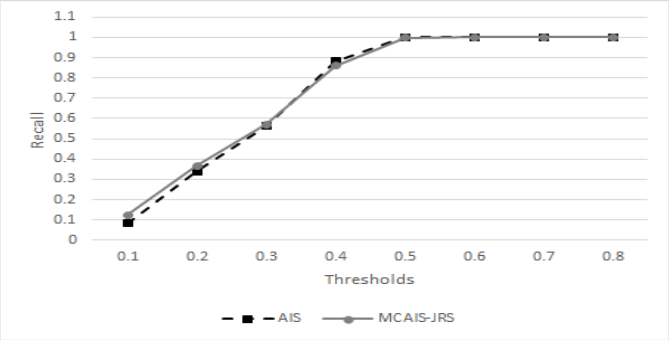


Fig. 8. The Recall rate of MCAIS-JRS and traditional AIS

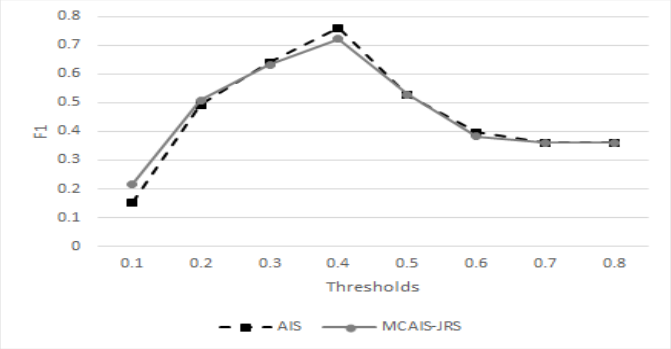


Fig. 9. The F1 rate of MCAIS-JRS and traditional AIS

C. Scalability Test

Scalability is an extremely important issue of recommender system efficiency, especially when they are used in larger datasets. We evaluate the scalability of the proposed algorithm by varying the datasets sizes. However, the execution time of the MCAIS-JRS and traditional AIS algorithms are compared using different datasets sizes. As seen in the previous section, both algorithms give relatively the likewise accuracy rates using different accuracy measures but we need to examine the system's efficiency when the datasets size grow. Figure 10 shows a comparison between MCAIS-JRS and the traditional AIS based on the execution time with different datasets sizes. When the number of applicants increased in the system, the execution time increases in a linear fashion, indicating that the traditional AIS efficiency is highly reduced while the execution time is stable in MCAIS-JRS and the system efficiency remains high in different datasets sizes. The maximum execution time for traditional AIS is 718 seconds with 7000 applicants in the system, while the MCAIS-JRS takes between 84 and 163 seconds in different datasets sizes. Specifically, the MCAIS-JRS achieves high efficiency with reasonable performance.

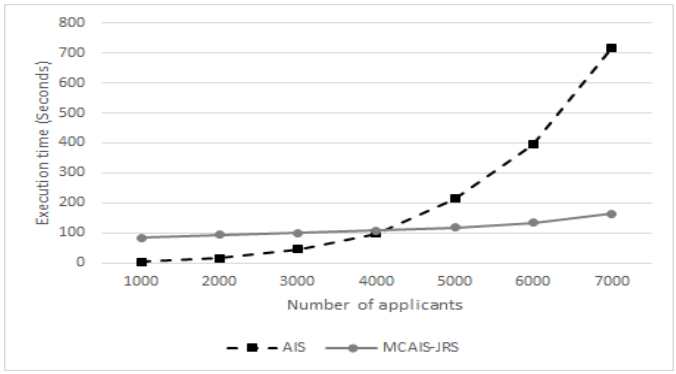


Fig. 10. Execution time of MCAIS-JRS and traditional AIS with different datasets sizes

VII. Conclusion

The discovery of monoclonal antibodies has encouraged a revolution in medicine that is probably only second to the discovery of vaccination. In this technology, the antibodies were produced in laboratories from single antibody cell. Our proposed model represents an optimization technique based on the abstractions of the monoclonal antibodies production principle and the AIS processes, and they are utilized towards enhancing the job recommendation. Hence, the job's requirements and applicant's resume exploited and used in the implementation of MCAIS-JRS that recommends the most qualified applicants to a specific job. In this algorithm, we produce antibodies using the monoclonal antibodies production principle by treating each job's feature as a single antibody. Additionally, we perform the diversity of job's requirements using diversity operator instead of mutation strategy that produces antibodies limited by the number of job's features and their alternatives. Moreover, as presented in experimental results section, our new proposed algorithm produces a good recommendation result and enhances the system efficiency compared with our previous AIS model.

Acknowledgment

This work was supported by Deanship of Scientific Research and Research Center of College of Computer and Information Sciences, King Saud University. The authors are grateful for this support.

References

[1] F. Färber, T. Weitzel and T. Keim, "An automated recommendation approach to personnel selection," in Americas Conference on Information Systems, Tampa, Florida, USA, 2003.

[2] X. Yi, J. Allan and W. B. Croft, "Matching resumes and jobs based on relevance models," in Proceedings of SIGIR, New York, NY, USA, 2007.

[3] S. T. Al-Otaibi and M. Ykhlef, "An artificial immune system for job recommendation," in proceedings of 3rd IEEE International Conference and Workshop on Bioinspired Intelligence (IWOBI 2014), Universidad Nacional, Liberia,Costa Rica, 2014.

[4] M. Leenaars and C. F. M. Hendriksen, "Critical steps in the production of polyclonal and monoclonal antibodies: evaluation and recommendations," ILAR Journal, pp. Vol. 46(3) ,pp. 269-279, 2005.

[5] S. T. Al-Otaibi and M. Ykhlef, "Survey of job recommender systems," International Journal of the Physical Sciences, pp. Vol. 7(29), pp. 5127-5142, 2012.

[6] S. T. Zheng, W. X. Hong, N. Zhang and F. Yang, "Job recommender systems: a survey," in Proceedings of the 7th International Conference on Computer Science & Education (ICCSE 2012) , Melbourne, Australia, 2012.

[7] J. Malinowski, T. Weitzel and T. Keim, "Decision support for team staffing: an automated relational recommendation approach," Decision Support Systems, pp. [45], 3, 429-447, 2008.

[8] J. Malinowski, T. Keim, O. Wendt and T. Weitzel, "Matching people and jobs: a bilateral recommendation approach," in Proceedings of the 39th Annual Hawaii International Conference on System Sciences, Hawaii, USA, 2006.

[9] T. Keim, "Extending the applicability of recommender systems:a multilayer framework for matching human resources," in Proceedings of the 40th Annual Hawaii International Conference on System Sciences (HICSS'07), Hawaii, USA, 2007.

[10] P.-C. Chen, "A Fuzzy multiple criteria decision making model in employee recruitment," IJCSNS International Journal of Computer Science and Network Security, pp. [9], 7, 113-117, 2009.

[11] K. Kowsari, M. Yammahi, N. Bari, R. Vichr, F. Alsaby and S. Y. Berkovich, "Construction of FuzzyFind Dictionary using Golay Coding Transformation for Searching Applications" International Journal of Advanced Computer Science and Applications(IJACSA), 6(3), 2015.

[12] M. Fazel-Zarandi and M. S. Fox, "Semantic matchmaking for job recruitment: an ontolgy based hybrid approach.," in In Proceedings of the 3rd International Workshop on Service Matchmaking and Resource Retrieval in the Semantic Web at the 8th International Semantic Web Conference, Washington D.C., USA, 2010.

[13] A. Singh, R. Catherine and K. Visweswariah, "PROSPECT: A System for screening candidates for recruitment," in Proceedings of 19th ACM International Conference on Information and Knowledge Management (CIKM'10), Toronto, Ontario, Canada, 2010.

[14] I. Paparrizos, B. B. Cambazoglu and A. Gionis, "Machine learned job recommendation," in proceedings of the fifth ACM conference on Recommender systems,RecSys'11, Chicago, Illinois, USA, 2011.

[15] M. Rodriguez, C. Posse and E. Zhang, "Multiple objective optimization in recommender systems," in Proceedings of the sixth ACM conference on Recommender systems, New York, NY, USA, 2012.

[16] Y. Lu, S. E. Helou and D. Gillet, "A recommender system for job seeking and recruiting websites," in International World Wide Web Conference Committee (IW3C2), Rio de Janeiro, Brazil., 2013.

[17] A. Gupta and D. Garg, "Applying data mining techniques in job recommender system for considering candidate job preferences," in Proceedings of the International Conference on Advances in Computing,Communications and Informatics (ICACCI), Delhi, India, 2014.

[18] D. Dasgupta, "Advances in artificial immune systems," IEEE Computational Intelligence Magazine, pp. 40-49, 2006.

[19] L. d. Castro and J. Timmis, "Artificial immune systems: a novel paradigm for pattern recognition," Artificial Neural Networks in Pattern Recognition , Alonso L, Corchado J,Fyfe C (eds), p. pp. 67–84. University of Paisley, 2002.

[20] R. A. Goldsby, T. J. Kindt and B. A. Osborne, Kuby Immunology, Fifth Edition, W. H. Freeman, 2002.

[21] U. Aickelin and D. Dasgupta, "Artificial Immune Systems," in Search Methodologies: Introductory Tutorials in optimization and decision support techniques, Springer, 2003, pp. 375-399.

[22] G. Kohler and C. Milstein, "Continuous cultures of fused cells secreting antibody of predefined specificity," Nature, pp. Vol. 256, pp. 495 - 497, 1975.

[23] J. A. Herlocker, J. A. Konstan, L. G. Terveeni and J. T. Riedl, "Evaluating collaborative filtering," ACM Transaction, 2004.

[24] B. Sarwar, G. Karypis, J. Konstan and J. Riedl, "Incremental SVD-Based algorithms for highly scaleable recommender systems," in Proceedings of the 5th International Conference on Computer and Information Technology (ICCIT '02), 2002.

A New Method for Text Hiding in the Image by Using LSB

Reza tavoli
Department of computer, Islamic Azad University, chalus Branch, chalus, Iran

Maryam bakhshi
Department of computer, poyandegan danesh, chalus, Iran

Fatemeh salehian
Department of electronic, hadaf university,sari, Iran

***Abstract*—an important topic in the exchange of confidential messages over the internet is the security of information conveyance. For instance, the producers and consumers of digital products are keen to know that their products are authentic and can be differentiated from those that are invalid .The science of encryption is the art of embedding data in audio files, images, videos or content in a way that would meet the above security needs. Steganography is a branch of data-hiding science which aims to reach a desirable level of security in the exchange of private military and commercial data which is not clear. These approaches can be used as complementary methods of encryption in the exchange of private data.**

Keywords—Hiding text inside an image; image processing; Steganography; image compression; LSB

I. INTRODUCTION

From the time that humans became able to communicate, developing a secret connection was one of the main demands. In past, despite having minute means, people had tried to hide data to not be discovered easily. These information that often their security mattered, usually were associated with war or military information and details of the countries' borders, which were hided in various frames according to the level of their importance.

In ancient Rome for instance, they used to shave the herald's head and tattoo the desired text on his skull. The herald was in quarantine until his hair grows, then he moved to the destination, that again by shaving his head they would read the hidden content.In addition, Italians in the medieval era used a sort of ink that could penetrate the egg shell and give color to the egg white. Thus by peeling the egg, they could easily read the data. In ancient Persia also, they wrote crucial data with the use of onion juice on the paper, hence when it dried, there was no sign of the content. Then with slightly heating the paper, the letters would become clear and the information could be readable [1].

On the other hand, these days the considerable progress of internet and the rapid growth of its use have propelled human to the digital world and communicating by the use of digital data. Meanwhile, the communication security is a critical need and is felt more and more every day. Today the modern techniques of steganography have found many users. In the terrorist operation of 11 of September also steganography was used for information conveyance of this operation. Furthermore, other quite useful applications for steganography that are in this area are in public TV posts, network, controlling the products copyright, search engines, image, and bank cards. Even nowadays medical science and DNA use steganography [2].

Generally there are three approaches for hiding text in an image. First method is encryption, in which information is encrypted in a way that is not understandable for the third party; however the receiver and transmitter can decrypt the data with a common key [3],[4]. The encryption or decryption operations are performed by programming algorithms in the digital domain and occasionally one can realize confidential data, depending on the level of the algorithm's security. The second method is steganography in which not only the information remains secret but also the existence of confidential connection is hidden. In fact steganography is the art and science of hiding communication and its purpose is to hide the existence of any connections between the receiver and transmitter. Often it is thought that the connection is secured by coding the exchanged message; however sometimes practically coding is not enough. Accordingly several methods were proposed for hiding data instead of coding. The third method is watermarking which will be further explained bellow. Assume that a legal owner of a photo embeds a series of messages in an image. Whenever such an image is stolen and put in a website, its legal owner can provide this confidential message to the court as a proof of his ownership. This type of hiding is called watermarking [5].

Prior to the explanation of the LSB method, we should clarify the main difference between encryption and steganography. In fact, the difference between these two terms is the purpose of encryption which is the concealment of the message content and generally not the existence of the message. Whereas steganography aims to hide any sign of the existence of the message [5]. In cases that the exchange of encrypted information is problematic, the existence of the communication should be hidden. For instance if one accesses encrypted content in any way, he will know that this content contains encrypted messages. However in steganography the third person does not obtain any information about the existence of the hidden message at all. The steganography methods were developed for protecting the property rights of multimedia products. In other words this technique was designed to protect the media itself [6].

Prior to further explanation, a brief discussion concerning the LSB method is necessary to simplify next topics. Most steganography approaches that embed the data within the pixel space take advantage of the LSB method. When a file is made, usually some of its bytes are not usable or are worthless [7]. These bytes can be changed without harming the file

considerably. This allows us to write some information in these bytes without anybody being aware that the process has taken place. As it was mentioned before, each video file is merely a binary file that contains colors and light intensity of each pixel according to the binary number [8].

Images normally use an 8-bit or 24-bit format. In the 8-bit format we solely can use 256 color for each pixel (In these 8 bits, each bit is one of the values of 0 or 1 which totally provides 2 ^ 8 or 256 different colors). In 24-bit format also every pixel have the capacity of 2 raised to the 24 power. In this format each pixel uses of 3 bytes of 8 bits. Each byte shows the light intensity of three main colors of red, blue and green. For instance, colors in format html3.0 are according to the 24-bit. Each color in this format has a code based on 16 which comprises of 6 characters. The first two characters are associated with the color red; also the second and third characters are respectfully associated with the colors blue and green. For example for creating the color orange, the intensity values of the colors red, green and blue are respectively 100%, 50% and 0 which is definable with #FF7FOO in html. Furthermore, the size of an image depends on the number of pixels of that image. For instance for an image with the resolution of 640*480 that uses the dynamic range of 8-bit, the image size should be 640*480**3 = 900 KB. Suppose that three neighbor pixels are coded as below [9]:

	Red, Blue, Green
pixel1	10010101/00001101/11001001
Pixel2	10010110/00001111/11001010
Pixel3	10011111/00010000/11001011

Now assume that we want to embed the 9 bit of these data 101101101into these pixels (these 9 bits encrypted data are supposed to be a message)

Now if we use the LSB method, these 9 bits are put into the least significant bits of these three pixel's bytes, then we have the below chart:

	Red, Blue, Green
pixel1	10010101/00001100/11001001
Pixel2	10010111/00001110/11001011
Pixel3	10011111/00010000/11001011

It is seen that only four bits have been changed and this would not harm the image greatly. For instance, a change in the blue color bit from 11111111 to 11111110 is never detectable for the eyes.

Now we may want to hide a text in an image. In this case every character, takes up one byte (8 bites). Since we should put these bits into these image pixels, thus we need to divide these eight bits to a 1-bit packages (or larger packages), and each bit are placed in the least significant bits of one of the main three colors of pixels. This way, words of all languages that are compatible with ASCII or UTF-8 (or any other coding), can be embedded within an image [10].

The LSB method with taking advantage of the random factors and secret key enhances the necessary security for hiding the data. However, by investigating the researchers' studies, one can simply show that this method can be broken (decoded). Although the least significant bits of pixels are seemed random, practically they do not have the real random. In general, the type of these bits arrangement in an image represents some features of that image [10], [11].

II. A Review of Previous Studies

Studies on image compression and steganography have been an active area of research from the beginning of the digital image processing. The use of preprocessing methods for improving compression rate and elevating the level of encryption has interested many researchers. Here we briefly explain some articles.

In a research done by $Shatnavy$in 2012, he used a method of embedding in consecutive pixels. According to his technique, the message with the hidden data is saved in the difference between the values of the consecutive pixels' gray levels. Here the gray level range is within 0 to 255. Selecting this range according to the sensitivity of the human visual system leads to the color change. After that the image is divided to anon-overlapping two pixels blocks. Then the difference between the gray levels of the consecutive pair of pixels d is calculated. If d is in r_k range, then r_k indicates the number of hidden bits in these pair pixels (In fact the difference in these pixels).Thus in parts of the image that the difference between the consecutive pixels is high, the sensitivity of the human visual system is low and therefore more information are saved. Then this number of bit is chosen from the bit stream of the hidden message and is summed with the lowest value of the r_k in decimal format. So a new value such as d is eventuated for the difference of the gray levels of the pixels [12].

In an article by Reddy and others in 2004, he offered a steganography method according to singular value decomposition and discrete wavelet transform. In this type of steganography which is driven from the composition and decomposition of the singular value and discrete wavelet transform, two domains of spatial and frequency steganography were compounded. In this method, discrete wavelet transform is applied on both image and $stegano$ image (the image that we want to hide). As we know in discrete wavelet, the image is divided to four frequency areas which are $cA, cH, CV, cD. cA$ isthe approximation signal (the above left side picture), $cD$is the detail signal (the bottom right side picture), cH the horizontal detail signal (above right) and cV the vertical detail signal (bottom left). In this approach, it applies discrete wavelet transform to both images. Then again it applies discrete wavelet transform to cA region. The obtainedcA region is used for continuing the steganography procedure. In continue, the obtained CA's are converted to three matrixes by singular value decomposition. Then the yield singular value is multiplied in a number less than one hundredth and summed with the singular value of the coverage image. After that by multiplying the singular value of these three matrixes and applying the inverse discrete wavelet transform, it converts to an image in which the $stegano$image is hidden in it [13].

In a paper submitted in 2012 by d.rajedi and others, they proposed a simple method of hiding information. This method

includes the involvement of different secret keys in various stages with the implementation of various matrixes and summing a series of handwritten codes. The proposed method in this paper can be thought as a ladder, in which the normal and encrypted texts are embedded upon the first and final steps. Furthermore in this paper, d.rajedi applied his method on the three-dimensional image. First the typical text simply and without any changes enters to this model and then is followed by a series of transformations, operations of changing information and handwritten codes. Finally it converts to an object with the name of RNS coded object. We can use the produced RNS object as the background of images. This approach is implemented on the images with the use of the alpha factor (the alpha character is connected to the clarity character of the images). Ultimately we have a clear image in the background of the main image. This scheme consists of three main parts which are the simple text encryption, the method of decryption of the encrypted text and the RNS model [14],[15].

III. The Introduction of the Proposed Approach in Hiding Data in an Image

The method that is suggested in this paper uses a stage of the textual data compression and then coding it prior to steganography. In other words, first it applies a preprocessing technique on the desired text, and then puts the text into the image. The proposed method encodes the compressed text and then with the use of a 4*4 mask performs snake scan ordering. After that it loads the eventuated compressed and coded text on image pixels. The below block diagram depicts the stages and procedures. The following parts of this section explain the encryption and decryption stages of the image.

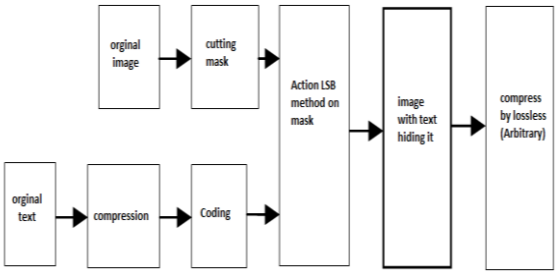


Fig. 1. The final block diagram of the proposed technique for steganography

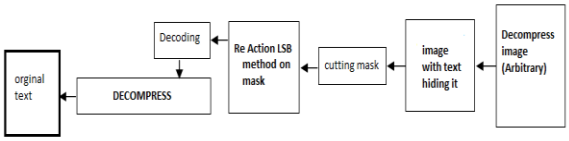


Fig. 2. The final block diagram of the proposed technique for steganography (Text mining)

A. Algorithms (the encryption steps for an image)

- Initially it decreases the volume of the raw image with the use of one the compression methods (Differential method for instance). In fact, this compression not only increases the capacity of number of saved characters in the image, but also it is thought as a type of coding.
- Then it divides text to an arbitrary length of segments. For example each 5 characters of the below array consists of the raw image and the presentation of the ASCCI cods of the desired image.

  Text = $gdcba$
  Text $_{(ASCII)}$ = 103,100,99,98,97
- The next step is the formation of the differential array from the previous scans in which it writes the first array element, and for remained numbers, it subtracts each array from its previous one.

  Array $_{(sub)}$ = 103, 3,1,1,1
- The maximum number from index one to the next ones (not the index zero) are specified for determining the required bit space for their storage. In previous step the maximum array number is three. The obtained maximum number requires two bits for storage. Now in decoding it should be specified that how many bits should be allotted to the data byte (which is two bits here). The storing format of data is as below:

Byte Sequence = (the first byte)(the length of data bytes)(the data bytes sequence)

- Data bytes are accompanied by a sign bit, which represents the sign of the difference from subtraction along an array, for which the succeeding number in the array is larger than the previous; then the bit sign is one, otherwise it is zero.
- For the further reduction to the number of bytes in a sequence, it employs the following code associated with sequence length:

$$if\ (MaxByte <= 1)\ SizeBye = 1, Else\ if\ (MaxByte < = 3)\ SizeBye = 2, Else\ if\ (MaxByte < = 7)\ SizeBye = 3\ \ldots$$
$$if\ (SizeBye = 1)\ insert\ 000\ Else\ if\ (SizeBye = 2)\ insert\ 001\ Else\ if\ (SizeBye = 3)\ insert\ 010\ \ldots$$

The below table values depicts the “data bytes length”

TABLE I. Binary Codes of Byte Size

Size Byte	code
1	**000**
2	**001**
3	**010**
4	**011**
5	**100**
6	**101**
7	**110**
8	**111**

- The output text of the first step is coded by an arbitrary algorithm encryption. An arbitrary key is used to XOR the encoded algorithm with the output text of the first stage and delivered to the next part.
- The XOR operation operates in a way that shows the difference between bits. In other words, if both bits are zero or one then the output is one, otherwise it is zero. For instance the two below characters is supposed:

Clear Text =ab

Text $_{(ASCII)}$ =97, 98 = (01100001, 01100010) $_{Base\ 2}$

- We suppose x as a key: Key $_{(ASCII)}$ = 120 = (01111000) $_{Base\ 2}$
- After applying XOR operation, the text is coded as below:

TABLE II. XOR TABLE FOR HIDING TEXT IN THE IMAGE

	a								b							
XOR	0	1	1	0	0	0	0	1	0	1	1	0	0	0	1	0
	0	1	1	1	1	0	0	0	0	1	1	1	1	0	0	0
Coded	0	0	0	1	1	0	0	1	0	0	0	1	1	0	1	0

- For converting the coded text to the normal mode (decoding), the key is XOR-ed with the data codes.

TABLE III. XOR TABLE FOR HIDING DECODES

XOR	0	0	0	1	1	0	0	1	0	0	0	1	1	0	1	0
	0	1	1	1	1	0	0	0	0	1	1	1	1	0	0	0
Decoded	0	1	1	0	0	0	0	1	0	1	1	0	0	0	1	0
	a								b							

- Obviously it is necessary for the receiver to have the key which the transmitter used for data coding.
- The output text from the second step is scanned with the LSB method and 4*4 masks (the snake scan is applied for more security).

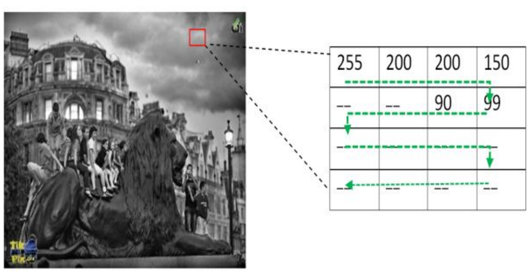


Fig. 3. snake ordering in mask

- If required, this step can compress the output image with lossless methods (such as LZW or differential).

B. Algorithm (the decoding stages of an image)

The decoding is the exact opposite of the above steps, thus preventing obvious explanations.

C. Evaluation

In order to evaluate our algorithm, we used the following arbitrary text and images:

An embedded text which includes 3681 characters and 4 images size 300*300 pixels.

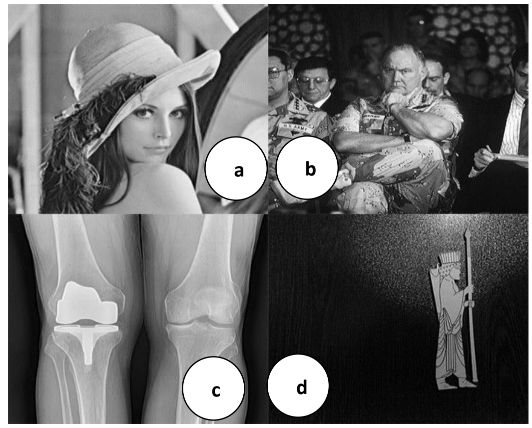


Fig. 4. Four used pictures in *steganography* (a) Lena's picture (b) USA's war minister picture (c) leg's scan picture (d) *Achaemenian* man picture

- Test results validation:

In this section, we validate our proposed method. Before encrypting a message, it has to transform into the binary form. This transformation is done based on 16 bites codes of UNICODE. For instance, each Farsi word in this transformation changes into the binary code of 16 bites. Thus, for obtaining the binary equivalent of a message, we have to put all 16 bites codes of characters in together. According to the proposed method, the equivalent binary message is encrypted by permutation technique, and then is placed in the image. The advantage of bit permutation over character permutation is that while changing encrypted bits into characters, difference characters are displayed rather than the main message. In this method the length of key is dynamic. In other word, it is arbitrary. As mentioned, each UNICODE is 16 bites. Therefore, if the length of a key is not a factor of 16, bites of difference characters are displaced while encrypting, and this considerably increases encryption power and causes that decryption becomes difficult. When encrypted bits are displaced in the image, recognizing whether any information is embedded into the image would become difficult for one who controls the connection in an unauthorized way.

As an instance, for validating, we encrypt the phrase of "this is a secret massage" in above image with the following key:

Key={17,5,14,20,1,18,3,10,4,16,0,2,6,11,13,7,12,8,9,15,19}

After that, it places it in the image and then extracts it by key and without key. Table below presents extracted values for both forms.

TABLE IV. COMPARE EXTRACTED TEXT WITH KEY AND WITHOUT KEY

	Information extracted
With key	This is a private message
Without key	炪╛ **G** 焐䛇'祲▯घ⋈ 堤□⁄鄔⒤譼•鄔≥絨**â**□□

As observed from the table, if an unauthorized person who controls the connection, suspected to the sending image, cannot recognize any information from it. Because while extracting LSB of an image without any information confronts with ambiguous data like those data in third row of table.

- Test results of peak signal to noise ratio:

Peak signal to noise ratio or PSNR is an engineering term for the ratio between maximum power of a signal and the maximum power of noise that affects the correctness of displaying an image. Or more simply, the less PSNR, the more noise which is due to *stagnography* in the image. For calculating PSNR, first we should obtain medium square error or MSE between main image and *stagnography* image. We use following expression for calculating MSE[16]:

$$\text{MSE} = \frac{1}{\text{n}}\sum_{\text{i=0}}^{\text{i=n}}(\widehat{Y}\text{i} - Yi)^2 \quad (1)$$

In which $\hat{Y}i$ is the main image and Yi is the *stagnography* image, respectively. Moreover, I is the length and width of both images. N is the number of image pixels. After MSE calculation, now we can calculate PSNR. The formula is as follow [16]:

$$\text{PSNR} = 10 * \log10\left(\frac{\text{M0}}{\text{M}}\right) \quad (2)$$

Where M_0 is thage has 32 bits, maximum value of a pixel is 2^{32}. Above expression can simply be presented as [17], [18]:

$$\text{PSNR}=20*\text{Log10}\left(\frac{\max \text{c}}{\sqrt{\text{MSE}}}\right)=20*\text{Log10}\left(\frac{2^{32}}{\sqrt{\text{MSE}}}\right) \quad (3)$$

Test results of selected images are given in below table:

TABLE V. MSE RESULT OF IMAGES WITH THE USE OF TRADITIONAL LSB METHOD IN EACH R,G,B CHANNEL

images	MSE B	MSE G	MSE R
a	37.49	37.77	37.64
b	112.91	113.17	113.02
c	52.47	52.72	52.57
d	50.42	50.65	50.56

TABLE VI. MSE RESULT OF IMAGES WITH THE USE OF PROPOSED LSB METHOD IN EACH R,G,B CHANNEL

images	MSE B	MSE G	MSE R
a	0.1075	0.4314	0.2803
b	0.1082	0.4304	0.2764
c	0.1049	0.4063	0.2592
d	0.1070	0.4326	0.2829

As the MSE criteria approaches zero, the less frequently the output image changes from the primary image, which is good. Accordingly, above tables show the fidelity of both methods.

TABLE VII. PSNR CALCULATION RESULTS OF IMAGES WITH THE USE OF SIMPLE LSB BY SEPARATING EACH R,G,B CHANNEL

images	PSNR B	PSNR G	PSNR R
a	32.39	32.35	32.37
b	27.60	27.59	27.59
c	30.93	30.91	30.92
d	31.10	31.08	31.09

TABLE VIII. PSNR CALCULATION RESULTS OF IMAGES WITH THE USE OF THE PROPOSED METHOD BY SEPARATING EACH R,G,B CHANNEL

images	PSNR B	PSNR G	PSNR R
a	57.81	51.78	53.65
b	57.78	51.79	53.71
c	57.92	52.04	53.99
d	57.83	51.76	53.61

For PSNR criteria, it is better that it approaches 100. The acceptable range in stagnography are within 50 to 100. In a simple LSB method, the criteria range is within 0 to 30; however table 4, simply presents that how these criteria were improved.

- Image Histogram:

Another way of recognizing a message in an image is comparing histogram of the main image with the *stagnography* image. In traditional LSB method, *stagnography* is done on sequential pixels, thus abnormality is created in the image histogram.

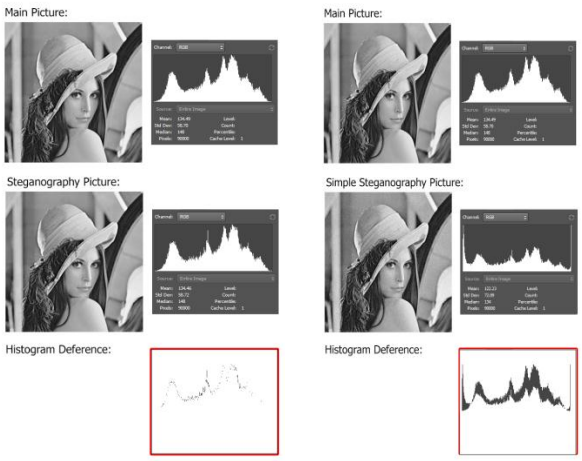


Fig. 5. Comparison between images and their histograms along with their difference histograms (right image is by simple traditional LSB and left image is by proposed method)

As it can be seen from Lena's picture, the histogram of a simple LSB image has a considerable difference with the main image, but in proposed method this difference is trivial due to recompressing words and using fewer image bits.

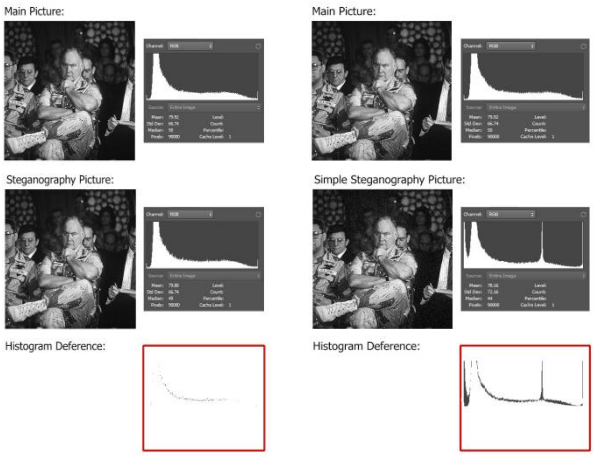


Fig. 6. Comparison between images and their histograms along with their difference histograms (right image is by simple traditional LSB and left image is by proposed method)

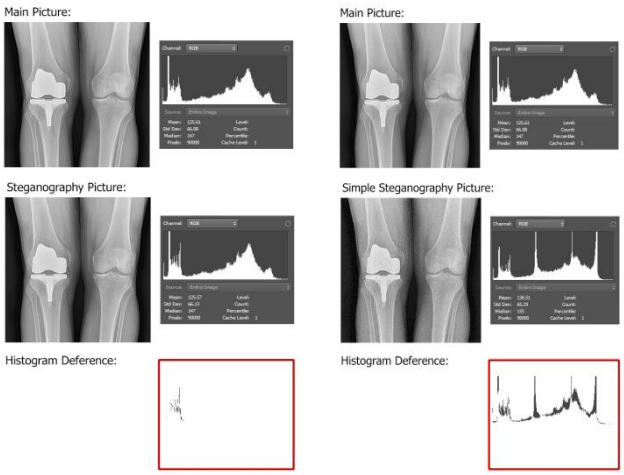


Fig. 7. Comparison between images and their histograms along with their difference histograms (right image is by simple traditional LSB and left image is by proposed method)

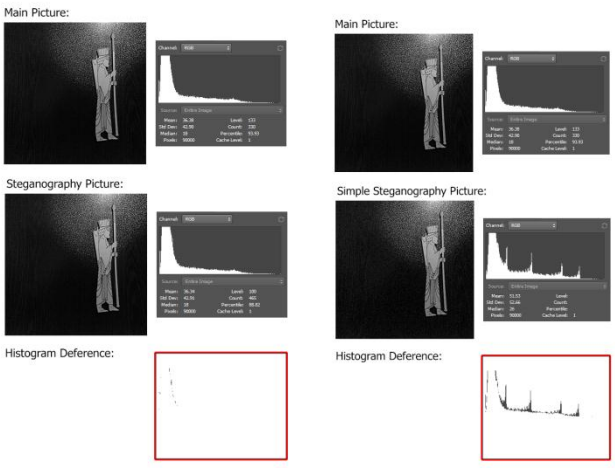


Fig. 8. Comparison between images and their histograms along with their difference histograms (right image is by simple traditional LSB and left image is by proposed method)

IV. Implementation

Regarding the primary desired aims of this paper that the most important one is the capability of commercialization, hence the implementation of the above steps was carried out in c# and under .NET FRAME WORK 4. Therefore it can be utilized in operating systems in a widespread range.

V. Conclusion

This paper initially investigates the multiple approaches of steganography in an image . It has shown that the space pixel provides more capacity than the frequency domain. In addition, the type of an image is effective in achieving the desired result in steganography.In comparison to today's proposed methods in frequency domain; our algorithm has the ability of storing a larger amount of information.

Compression before hiding step is more appropriate in invisibility of the steganography. Furthermore it increases the image capacity for data inclusion. Mixing the use of an appropriate mask with the application of a particular scan of an image, and moreover adding a step of encryption with each, allows for multiple separate phases of information security. By combining the mentioned stages with LSB approach, a desirable percentage of steganography was yielded. Therefore these steps decrease the odds of discovering the hidden data. In fact, by putting together a number of methods and designing an efficient algorithm, we have achieved an innovation and a relative improvement in LSB method. although performing all of the steps above successfully obtains a higher level of security using the LSB method, it also contributes to the problem of increased load to the processing system. For this shortcoming a solution should be devised which does not fit within the scope of this article. The second proposal is to have a random place for masks in the image. In detail, it chooses a fixed position for the first mask, but for the position of the next mask, it selects randomly. Like the structure of a linked list, the address of the next mask is saved in its previous mask. Hence we will reach to a higher level of security.

Reference

[1] Abas Chedad,Joan Condell,Kevin Curran,Paul McKevitt (2010), Digital image steganography: Survey and analysis of current methods, Signal Processing, 727–752

[2] Udda Lavanya,YangalaSmruthi,Srinivasa Rao Elisala, Data hiding in audio by using image steganography technique, Volume 2, Issue 6, November – December 2013

[3] Weiqi Luo,Jiwu Huang,Fangjun Huang (2010), "Edge Adaptive Image Steganography Based on LSB Matching Revisited", IEEE Transactions on Information Forensics and Security, vol. 5, pp. 201-214

[4] HadVitthal S.,BhosaleRajkumar S.,PanhalkarArchana R)2012(, A Novel Security for Secret Data using Cryptography and Steganography" International Journal Computer Network and Information Security, vol. 2, pp. 36-42

[5] Vanya,2 YangalaSmruthi,3 Srinivasa Rao Elisala(2013), "Data hiding in audio by using image steganography technique", International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), VOL 2.

[6] R. C. Gonzalez and R. E. Woods, Digital Image Processing, Addison Publishing Co., 1993.

[7] P.Wayner,Disaapearing Cryptography, 2nd Ed., Elsevier Science: US, 2002.

[8] R.J.Anderson and F.A.P.Petitcolas, "On the limits of steganography," IEEE J. of Selected Areas in Communication, vol. 16, no. 4, pp. 474-481, May 1998.

[9] Al-Shatnawi A.M., "A New Method in Image Steganography with Improved sequential bits", Applied Mathematical Sciences, vol. 6, no. 79, pp. 3907 – 3915, 2012.

[10] Reddy A.Adhipathi,and B.N.Chaterji 2004, "A new wavelet based logo-watermarking scheme", Pattern Recognition Letters, September 2004.

[11] Rajdeep Chowdhury,Nilanjan Dey,Saikat Ghosh" Design and Implementation of RNS Model Based Steganographic Technique for Secured Transmission", International Journal of Advanced Research in Computer Science and Software Engineering, Volume2, Issue 3, March 2012

[12] Phad Vitthal S.,Bhosale Rajkumar S.,Panhalkar Archana R) 2012(A Novel Security for Secret Data using Cryptography and Steganography" International Journal Computer Network and Information Security, vol. 2, pp. 36-42

[13] Nagham Hamid,Abid Yahya,R.Badlishah Ahmad & Osamah M.Al-Qershi,"Image Steganography Techniques: An Overview", International Journal of Computer Science and Security (IJCSS), Volume (6): Issue (3):2012

[14] Weiqi Luo, Jiwu Huang, Fangjun Huang,(2010) "Edge Adaptive Image Steganography Based on LSB Matching Revisited", IEEE Transactions on Information Forensics and Security, vol. 5, pp 201-214

[15] Zhijie Shi and Ruby B. Lee, "Bit Permutation Instructions for Accelerating Software Cryptography", Proceedings of the IEEE International Conference on Application-Specific Systems, Architectures and Processors, pp- 138-148, July 2000.

[16] Rohankar, Jayant (Nov 2013)- "SURVEY ON VARIOUS NOISES AND TECHNIQUES FOR DENOISING THE COLOR IMAGE" (PDF)- International Journal of Application or Innovation in Engineering & Management 2 (11).Retrieved 15 May 2015.

[17] Shamim Ahmed Laskar ,Kattamanchi Hemachandran(2012), High Capacity data hiding using LSB Steganography and Encryption ,(IJDMS) International Journal of Database Management Systems, Vol.4, No.6

[18] Strang, Gilbert (July 19, 2005), Linear Algebra and Its Applications (4th ed.), Brooks Cole, ISBN 978-0-03-010567-8

A New CAD System for Breast Microcalcifications Diagnosis

H. Boulehmi, H. Mahersia and K. Hamrouni
National Engineering School of Tunis, LR-SITI
ElManar University, BP-37, Le Belvédère 1002
Tunis-Tunisia

***Abstract*—Breast cancer is one of the most deadly cancers in the world, especially among women. With no identified causes and absence of effective treatment, early detection remains necessary to limit the damages and provide possible cure. Submitting women with family antecedent to mammography periodically can provide an early diagnosis of breast tumors. Computer Aided Diagnosis (CAD) is a powerful tool that can help radiologists improving their diagnostic accuracy at earlier stages. Several works have been developed in order to analyze digital mammographies, detect possible lesions (especially masses and microcalcifications) and evaluate their malignancy.**

In this paper a new approach of breast microcalcifications diagnosis on digital mammograms is introduced. The proposed approach begins with a preprocessing procedure aiming artifacts and pectoral muscle removal based on morphologic operators and contrast enhancement based on galactophorous tree interpolation.

The second step of the proposed CAD system consists on segmenting microcalcifications clusters, using Generalized Gaussian Density (GGD) estimation and a Bayesian back-propagation neural network.

The last step is microcalcifications characterization using morphologic features which are used to feed a neuro-fuzzy system to classify the detected breast microcalcifications into benign and malignant classes.

Keywords*—*Artifacts and pectoral muscle removal; Bayesian back-propagation neural network; Breast microcalcifications; CAD system; Digital mammograms; Galactophorous tree interpolation; GGD estimation; Morphologic features; Neuro-fuzzy system

I. Introduction

Breast cancer is the first cause of death among women worldwide. Studies have shown that detection of breast lesions at an early stage would increase the chances of survival and reduce the risk of sequels and obviously mortality. This detection can be achieved by submitting menopausal women and those with a family history of breast cancer on a mammogram every two years. However, analyzing mammograms by radiologists is not a trivial task: breast density is an important factor that can increase the risk of misinterpretation.

Computer-assisted diagnosis (CAD) offers radiologists a reliable aid to breast cancer screening. In this context, a new mammographic images enhancement approach is proposed, beginning with the application of the top hat to extract the breast area and eliminate artifacts. A wavelet contrast enhancement step is then carried out followed by a detection and suppression of pectoral muscle. Finally, oriented version of top hat is exploited for detection and interpolation of the galactophoric tree.

Then, a new technique is proposed for microcalcifications (Mcc) segmentation, based on the measurement of the generalized Gaussian density (GGD) and the use of a supervised classifier (a neural network with Bayesian back-propagation).

A classification approach of the detected lesions is finally introduced. It is to operate three morphological descriptors and a supervised classifier (neuro-fuzzy system) to distinguish between benign abnormalities and those malignant.

II. Related Works

Microcalcifications are tiny flecks of calcium, like grains of salt, in the soft tissue of the breast that can sometimes be an early indicator of breast cancer.

There is a variety of microcalcifications shapes as shown by Fig. 1: annular, round, linear, vascular... [38]

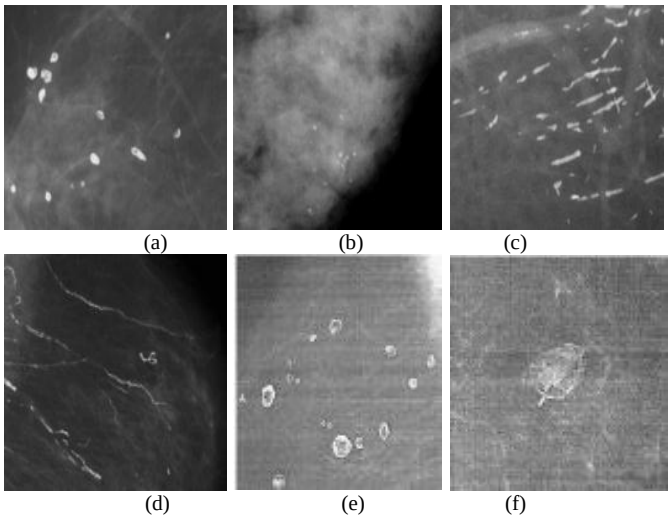


Fig. 1. Examples of microcalcifications (a) Annular, (b) Regular full round, (c) Linear, (d) Vascular, (e) With clear center and (f) Egg shell

Several works have been carried out in order to enhance microcalcifications clustering on digital mammograms, segment them, characterize them and classify them into benign and malignant classes [4, 7].

A. Enhancement techniques

The major defect that could oppose a better mammograms analysis is a low contrast. Therefore, a contrast enhancement is necessary to improve the quality of mammographic images and facilitate their exploitation.

There are three different families of contrast enhancement techniques [14]: conventional techniques, regions-based techniques and features-based techniques.

1) Conventional techniques: include global techniques (histogram stretching, histogram equalization [43] and convolution mask [54]) and techniques with sliding window (of fixed size [31] or adaptive size [22]). This category includes also adaptive enhancement techniques such as Sobel operators [47].

2) Regions-based techniques: mainly consist on performing a region growing algorithm from a well-chosen pixel called "a seed" [56].

3) Features-based techniques: consider the processed mammographic image characteristics and include morphological operations, wavelet transform and fractal approach [9, 47].

Although the performance of all these enhancement techniques depends on mammograms resolution [20], it has been proved that hybrid enhancement techniques are usually the best, since they allow strengthening advantages and filling disadvantages from several techniques.

B. Segmentation techniques

The main target of these techniques is to identify possible regions of interest (ROI). In [62], authors illustrated a detailed study of mammograms segmentation techniques that can be done either by using a unique view of the breast, or by considering multiple views.

1) Single view lesions detection: it consists of using a single mammogram to detect possible lesions.

This category includes regions based approaches, such as regions growing algorithm [33, 53, 71], watershed algorithm [32] and Split and Merge algorithm [13].

Some other approaches are based on the edge detection of mammogram components [5, 10, 21, 25, 28, 34, 37]

We also find, in this category, clustering based approaches. They consist of detecting clusters which may represent an eventual tumor [12, 48]. These techniques are well suitable to microcalcifications' clusters detection.

Another type of single view techniques is models based approaches based on comparing the patient mammograms to known images of healthy and pathological cases [17, 30].

2) Multiple views lesions detection: the lesions detection is done by comparing two mammographic images that can come from right breast and left breast. In this case, radiologists compare right and left mammograms to seek for abnormalities in both images [15, 55, 65].

Two different views of the same breast could be used as well; mostly one mediolateral oblique (MLO) view and one cranio-caudal (CC) view of the same breast [74].

The two views can also come from two mammograms of the same breast taken at different moments: the main purpose is detecting a possible lesion evolution [75].

The efficiency of all segmentation techniques has been widely proven in literature. However, each technique still presents some disadvantages [62]. For example, region-based approaches depend on the seed selection and the algorithm ending conditions. Some techniques (mainly fractal model technique) are known as time-consuming [28].

C. Characterization techniques

The main goal of these techniques is to extract several primitives to characterize the ROI selected during the segmentation step, in order to classify the lesions into benign and malignant classes. Several primitives have been exploited in the literature. In [14], Cheng et al. have summarized the different primitives used for lesions characterization.

There are characteristics related to microcalcifications clusters: description of the weight distribution, the area and the number of microcalcifications [18, 19, 20, 66].

Primitives extracted from co-occurrence matrix such as energy, entropy and contrast were used in [25].

Few works have used surround region dependence matrix – SRDM, gray level run length matrix – GLRL and gray level difference matrix – GLD [42].

Wavelet decomposition provides many primitives characterizing gray-levels frequencies from different orientations and has been widely used in breast cancer context [16, 23, 27, 41].

There are other techniques that have been exploited in breast lesion characterization such as Gabor filter bank [23, 57], Gaussian Laplacian filter [59] and fractal dimension [60].

The use of all the primitives described above can offer almost perfect results of classification [20]. However, characterizing techniques could not be evaluated separately, but rather in association with the classification approach.

D. Classification techniques

Several classifiers were used to distinguish malignant lesions from benign ones [14, 60], but the most commonly used are Neural networks [23, 35, 39], K nearest neighbors [44], Bayesian classifier [39, 57, 61], Quadratic classifier [52], Linear classifier [20], Expert system [18], Binary decision tree [71], Genetic algorithms [11], SVM [25, 58] and Adaptive thresholding [67].

Cheng et al. have evaluated the accuracy of the different classifiers in malignancy analysis as follows: from 87% to 90% with neural networks classifiers, from 71.08% to 83.13% with the k-nearest neighbors' technique and from 94% to 97.3% with decision tree [14].

However, this accuracy is highly sensitive to the primitives' selection during the characterization step.

III. Proposed Microcalcifications Cad System

The proposed CAD system chain contains four essential steps: mammograms enhancement by interpolating the galactophorous tree, microcalcifications detection using GGD estimation, morphologic characterization of detected clusters and neuro-fuzzy classification. This approach is described by Fig. 2.

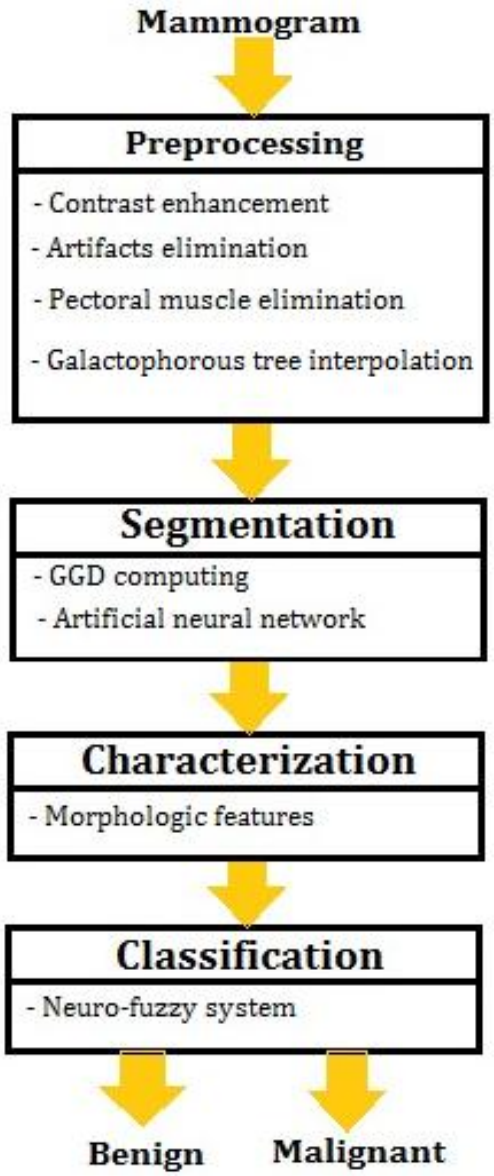


Fig. 2. Proposed microcalcifications CAD system

A. Proposed enhancement approach

Low contrast is the major problem encountered during mammograms analysis. Therefore, several contrast enhancement techniques have been developed, in order to solve this issue and facilitate lesions detection [9, 14, 22, 31,43, 47, 54, 56] as described in the previous section.

The proposed enhancement method, published last year in [6], operates in four steps aiming to delimitate breast area from a digital mammogram and increase contrast between normal tissue and possible microcalcifications' clusters.

The first step consists on removing all unnecessary details (radiologists' labels, scanning artifacts and film boundaries). The second step is to increase the image contrast and denoise it using wavelet transform. The third step consists of detecting and removing pectoral muscle and the fourth step aims to detect, then interpolate, galactophorious tree from the mammogram. These four steps help to prepare the breast image to further treatments by delimiting the breast region and enhancing the suspicious regions.

1) Artifacts and film boundaries removal: In order to extract the breast region, black and/or white vertical bands, corresponding to the film boundaries, are first removed. These columns are eliminated using a simple algorithm that detects the first four not-black corners in the image.

Then, morphological erosion with a square structuring element of size 13 pixels is applied, followed by a thresholding operation. A well-chosen threshold reveals two related areas of very different sizes: the big one corresponds to the mammary gland, and the other one to artifacts. A simple opening with a structuring element of a greater size helps deleting artifacts and keeping only the breast area (Fig. 3).

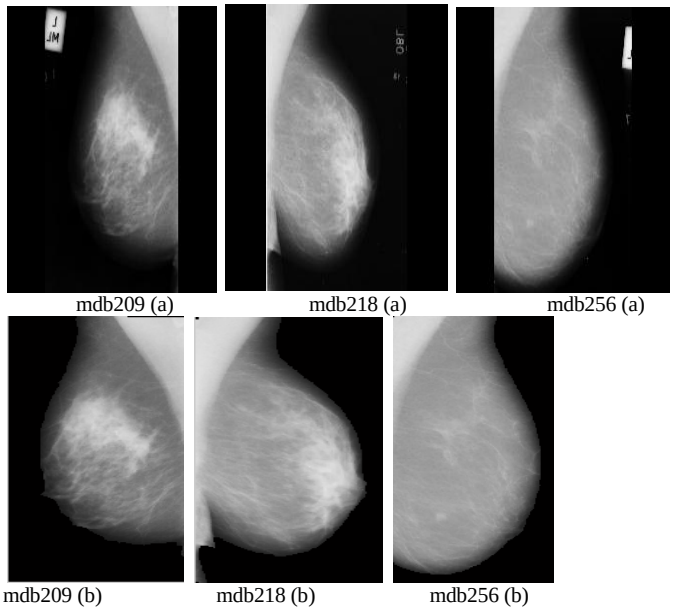


Fig. 3. Artifacts and film boundaries removal (a) Original images and (b) Resulting images

2) Contrast enhancement and wavelet denoising: Contrast enhancement step is very important to ensure better segmentation results. Classic techniques have shown limits in medical image processing. Several enhancement approaches have been proposed in literature [64, 68, 69, 70] aiming to improve mammographic images contrast. One of them has given better results with medical images: it is wavelet transform.

This second step has been inspired by Kumar et al. who have developed an algorithm based on wavelet multi-resolution theory and Wiener filtering in [46].

First, low frequency component L(x, y) is extracted from the preprocessed image I(x, y), using a Gaussian low-pass filter, in order to separate the useful information contained in the lower part of the image, from the noisy information contained in the higher parts. Then, a white top-hat (ToHb) and a black top-hat (ToHn) transforms are separately applied to L(x, y). These transformations are given by (1) and (2).

$$ToHb = L - (L \circ S) \quad (1)$$

$$ToHn = (L \bullet S) - L \quad (2)$$

White top hat is defined as the difference between the original image and its opening by the structuring element S and black top-hat transformation is defined by the difference between the image and its closing. The resulting image E(x,y) is given by (3).

$$E(x, y) = L(x, y) + ToHb - ToHn \quad (3)$$

A soft denoising algorithm is finally applied, based on three steps: 2-level Daubechies wavelet decomposition is performed, then the detail coefficients are thresholded (with dynamic threshold calculated at each level of decomposition. Finally, a wavelet reconstruction is applied so that the contrast of the resulting images is visibly enhanced (Fig. 4).

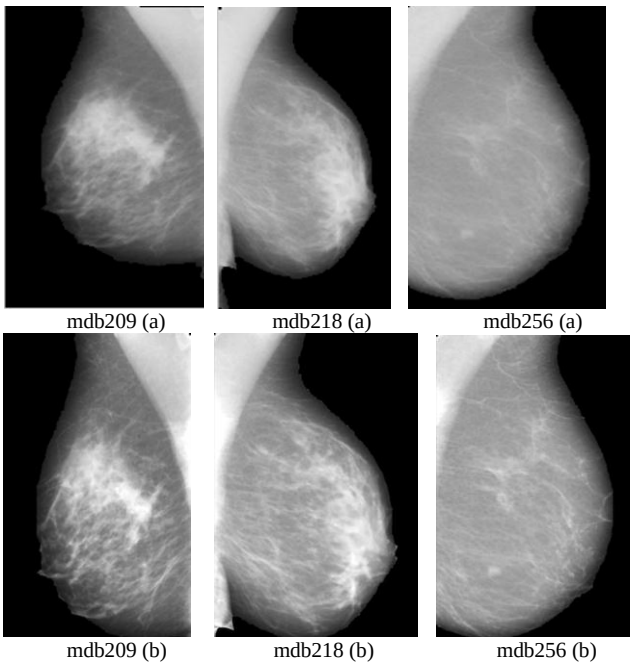


Fig. 4. Contrast enhancement and wavelet denoising (a) Images resulting from first preprocessing step and (b) Enhanced images

3) Pectoral muscle removal: The third step of the proposed preprocessing technique is the removal of the pectoral muscle. This muscle usually appears in MLO mammograms with intensities very similar to those of the microcalcifications. An adaptive thresholding algorithm is used to detect pectoral muscle and then remove it to keep only breast region as illustrated in Fig. 5.

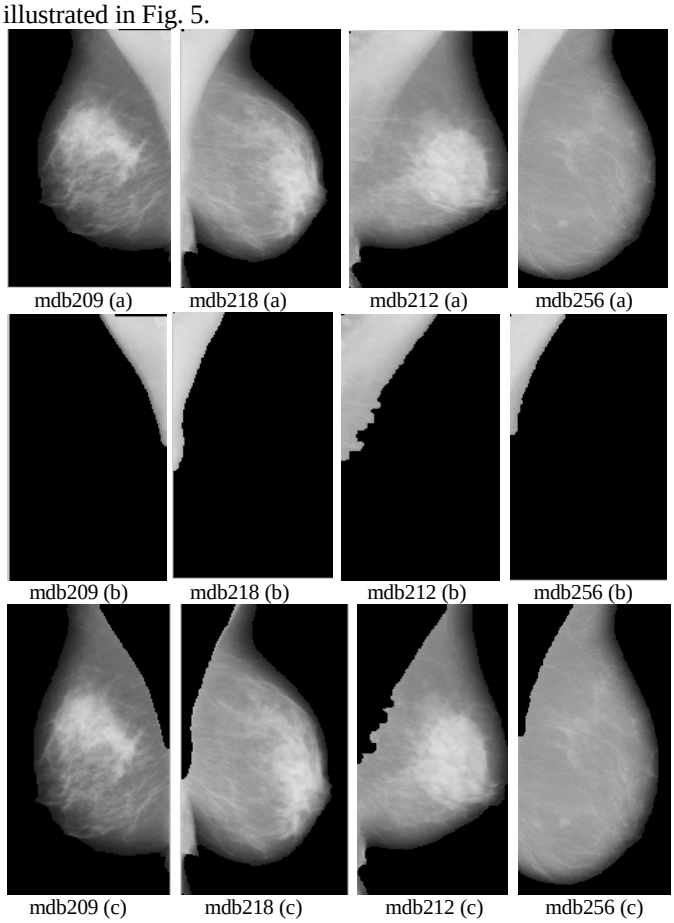


Fig. 5. Pectoral muscle removal (a) Images resulting from second preprocessing step, (b) Detected pectoral muscles and (c) Pectoral muscles removed

4) Detection and interpolation of the galactophorous tree: In order to increase mammograms contrast, especially for dense breasts, it is important to remove from the breast region, all the details that may interfere with detecting microcalcifications, including galactophorous tree.

Galactophorous tree has the structure of overlapped vessels with high gray levels intensities that connect lobules of the mammary gland to the tip of the nipple. Its presence could lead to wrongly suspected regions. Indeed, galactophorious tree is in the form of a lines network with variable thickness from a region to another. In this last preprocessing step, an oriented version of the top-hat transform is used to detect all pixels of the galactophorous tree.

The extracted elements width can be controlled by the structuring element choice. Since galactophorous vessels have different thicknesses and gray levels, the structuring element must be straight and oriented in different directions. The different tests led to use three straight segments of respective lengths 10, 20 and 30 pixels, oriented in 13 different directions rising from 0° to 360° by step of 30°. The galactophorous tree is obtained by summing all the 39 obtained images (13 orientations for each of the 3 segment lengths).

A morphological opening step is then applied to remove isolated regions. The final step consists on interpolating all the pixels belonging to this mask with the average value of their eight nearest neighbors. Examples of resulting images are given in Fig. 6.

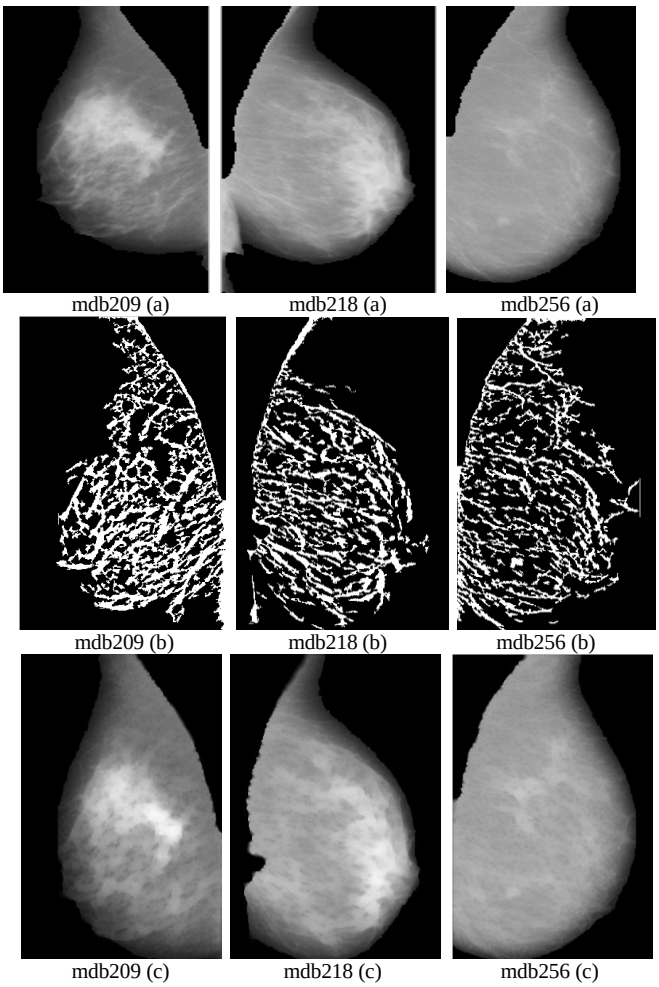


Fig. 6. Galactophorous tree interpolation (a) Images resulting from third preprocessing step, (b) Detected galactophorous trees and (c) Enhanced images

B. Proposed segmentation approach

In [8], an unsupervised masses detection approach based on Generalized Gaussian Density (GGD) was proposed. In this work, the GGD estimation is used with a supervised classifier to detect microcalcifications clusters.

The main principle of the Generalized Gaussian Density is wavelet decomposition.

Texture analysis using Generalized Gaussian Density was introduced by Do and Vetterli [24]. It consists on building the histogram showing the distribution of the coefficients extracted from the wavelet transform at a given sub-band (level). For each sub-band, a continuous law describing as faithfully as possible the histogram behavior is determinate.

Experimentally, the histogram distribution resembles a Gaussian distribution centered on 0, but for some textures, the peak at 0 is not very rounded and rather reminds a Laplace distribution [49].

Do and Vetterli [24] have proposed to model the wavelet coefficients behavior, at each scale, by a generalized Gaussian, parameterized by three factors μ, α and β (4) [73]:

$$P_{\mu,\alpha,\beta}(x) = \frac{\beta}{2\,\alpha\,\Gamma(\frac{1}{\beta})}\, e^{-\frac{|x-\mu|^{\beta}}{\alpha}} \quad (4)$$

Where:

$\Gamma(z) = \int_0^{+\infty} e^{-t}\, t^{z-1} dt$, $z > 0$: gamma function

μ, α and β: mean, scale and shape parameters respectively

The form factor β governs the shape, more or less sharp. The scale factor α governs the spread of the curve and corresponds to the standard deviation in the case of a classic Gaussian. Fig. 7 gives examples of GGD distributions for different values of α and β.

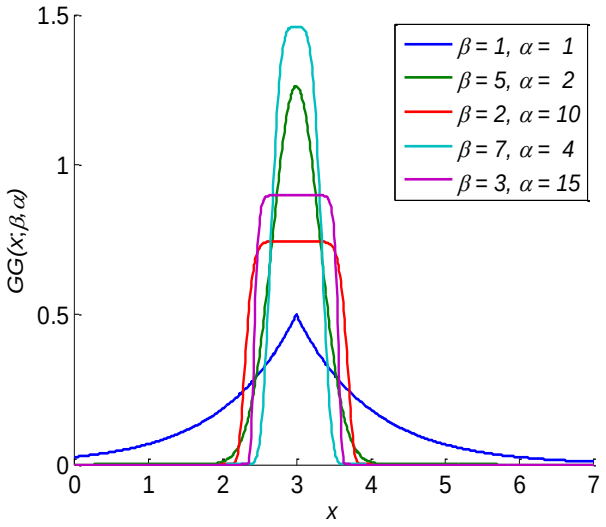


Fig. 7. GGD distribution for different values of α and β (μ=0)

In order to correctly decide whether a microcalcifications cluster exist or not in a given mammogram, two processes are ensured: a training process and a testing process [50].

- Training: in this step, known mammograms are considered. They are enhanced as described previously and then decomposed using a three-level redundant Haar wavelet transform. The multi-scale analysis is carried out by sweeping a 64 × 64 pixels window in the entire image. Then, several primitives are extracted (such as GGD, energy, mean and standard deviation) and then stored in a features' matrix.
- Testing: in this stage, an unknown mammogram is preprocessed, divided into blocks of 64 × 64 pixels and then decomposed using the same wavelets transform described above. Then, a set of features is extracted for each block and compared to the feature values stored in the features' matrix. To achieve this comparison, we use multi-layer perception Bayesian regularization

neural networks to train data set and classify extracted features.

Nowadays, neural networks are one of the most powerful tools in textural tissues recognition [40, 72], since they are able to learn from known examples. The elementary component of a neural network is the neuron. Each neuron is linked to some of its neighbors with varying coefficients of connectivity representing the strengths of these connections [26].

During the learning procedure, the connectivity coefficients are adjusted so that neurons can be grouped into layers. The majority of the published works use standard 3-layer architecture.

In this work, several tests have been carried out, using neural networks with one and two hidden layers, containing 3 to 15 neurons each. In addition, three different activation functions were tested: logistic sigmoid function (logsig), hyperbolic tangent function (tanh) and a simple linear function (pur). Fig. 8 shows results of these different tests.

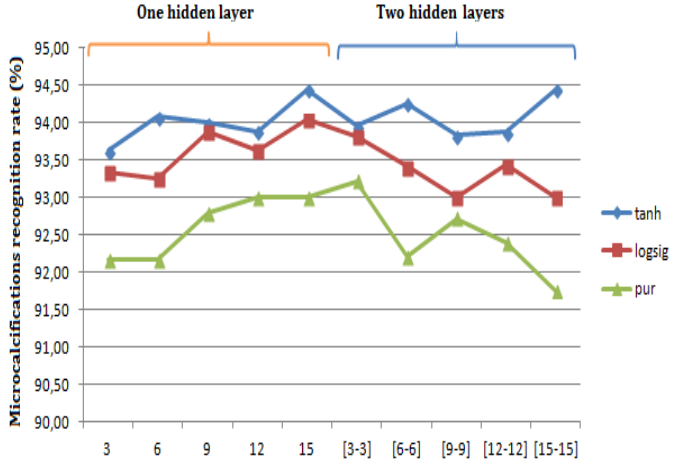


Fig. 8. Recognition rates for different activation functions and different numbers of neurons

Best microcalcifications segmentation rates are obtained with the hyperbolic tangent activation function, for a neural network containing two hidden layers with 15 neurons each.

Besides, back-propagation learning is used: it consists of minimizing an error function using an optimization method such as gradient descent, Quasi-Newton, and Levenberg–Marquardt method [50].

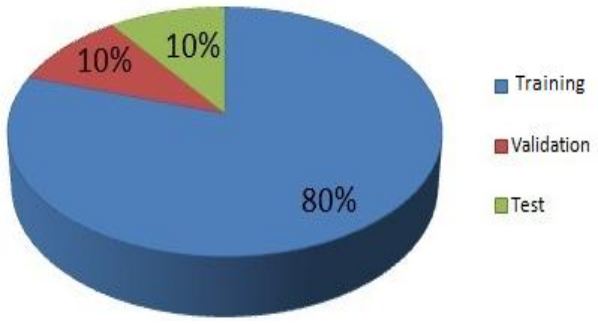
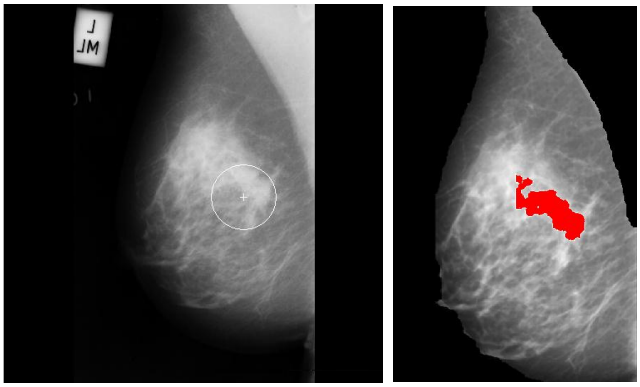


Fig. 9. Data sets partitioning

In addition, a 10-fold cross-validation process is used to avoid over fitting and improve the generalization ability of the back-propagation trained net-work. Therefore, the input data are randomly partitioned into 3 sets: a training set, a testing set and a validation set (Fig. 9). Each time, the neural network is trained with the training set and then verified with the generated validation set until the validation error starts increasing. The training procedure is then stopped and the network with minimum validation error is selected as the best model and used to classify the test set.

The process was performed 10 times, and then all the10 recognition rates are averaged to obtain the final performance of the proposed system. In order to create distinct data sets for cross-validation, none of the sets in the training folder appear in any of the remaining folders. This way, every network was trained to give the maximum value of 1 for the extracted microcalcifications cluster region and 0 for the other regions.

Fig. 10 gives an example of microcalcifications segmentation result on a digital mammogram.



mdb209 (a) mdb209 (b)

Fig. 10. Example of microcalcifications cluster detection (a) Original mammogram with ground truth and (b) Detected microcalcifications cluster

C. Proposed classification approach

In this work, three morphologic features are first extracted to describe the detected microcalcifications cluster distribution and size: area, compactness and eccentricity. Then a neuro-fuzzy network is exploited to classify characterized microcalcifications into malignant and benign classes.

Fuzzy logic has become a significant area of interest for researchers on artificial intelligence. Pr. Mamdani was the first to investigate the use of fuzzy logic to simulate human decision principles. Fuzzy models have the advantage of integrating the knowledge representation and reasoning mechanism with the priori expert experience and knowledge.

A fuzzy system is composed of a knowledge base (KB), and an inference engine module that includes a fuzzification interface, an inference system and a defuzzification interface.

The KB contains a Data Base (DB) and a Rule Base (RB): the Data Base contains all the sets considered in the linguistic rules and the membership functions defining the semantics of the linguistic labels and the Rule Base contains a collection of linguistic rules that are joined by some operators.

The structure of a fuzzy system is illustrated in Fig. 11.

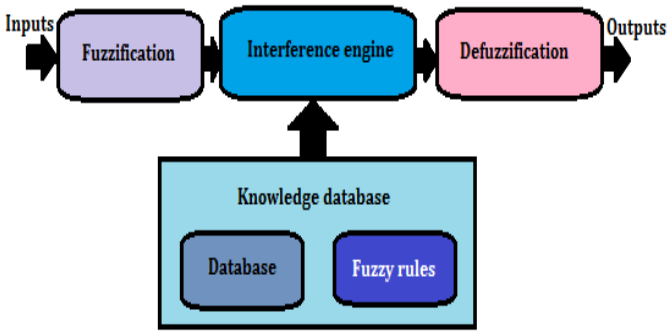


Fig. 11. Structure of a fuzzy system

Combining neural networks with fuzzy systems, called neuro-fuzzy systems, is a powerful alternative approach to develop fuzzy systems [29]. In fuzzy systems, relationships are represented explicitly in the form of if-then rules whereas, in neural networks, the same relationships are not explicitly given, but are given in the network by its parameters. Neuro-fuzzy systems combine the semantic transparency of rule-based fuzzy systems with the learning capability of neural networks [3]. In this work, an improved neuro-fuzzy system, known as adaptive network-based fuzzy inference system (ANFIS) is used [36]. It is a neuro-fuzzy network with five layers. It includes a knowledge representation and a reasoning mechanism resembling a human expert one.

The inference engine simulates the human expert reasoning based on fuzzy. In our case, we used 9 fuzzy rules that are destined to assist the ANFIS system in the classification decision of microcalcifications, based on the 5 classical membership functions of the deduction system: Very Negative (VNE), Negative (NE), Zero (Z), Positive (PO) and Very Positive (VP).

For each mammogram, fuzzy system is run 10 times, and the average of the ten classification rates is considered. The fuzzy rules used for microcalcifications classification are the following:

- Rule 1: if (e1 is PO) and (e2 is PO) and (e3 is PO) then (Malignant is VPO).
- Rule 2: if (e1 is NE) and (e2 is NE) and (e3 is NE) then (Malignant is VNE).
- Rule 3: if (e1 is PO) and (e2 is PO) and (e3 is NE) then (Malignant is PO).
- Rule 4: if (e1 is PO) and (e2 is NE) and (e3 is PO) then (Malignant is PO).
- Rule 5: if (e1 is NE) and (e2 is PO) and (e3 is PO) then (Malignant is PO).
- Rule 6: if (e1 is Z) and (e2 is Z) and (e3 is Z) then (Malignant is Z).
- Rule 7: if (e1 is PO) and (e2 is NE) and (e3 is NE) then (Malignant is NE).
- Rule 8: if (e1 is NE) and (e2 is PO) and (e3 is NE) then (Malignant is NE).
- Rule 9: if (e1 is NE) and (e2 is NE) and (e3 is PO) then (Malignant is NE).

Fig. 12 and Fig. 13 show the membership functions respectively of the three inputs and the one output of the ANFIS system used for microcalcifications malignant/benign classification.

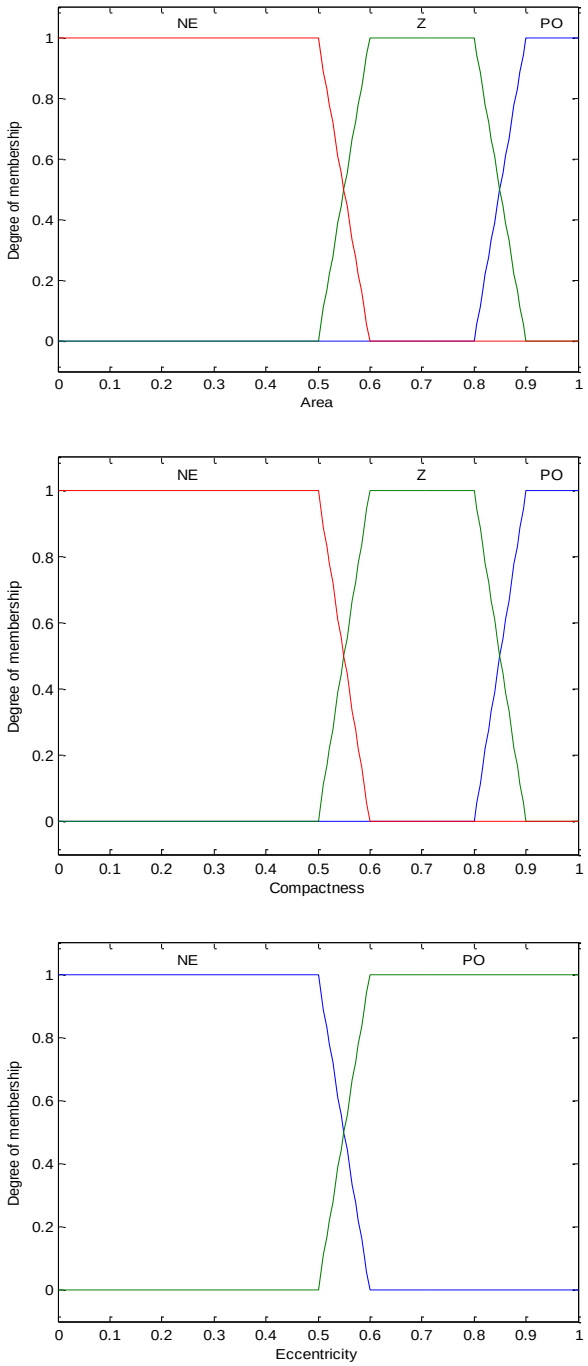


Fig. 12. Membership functions of the three inputs

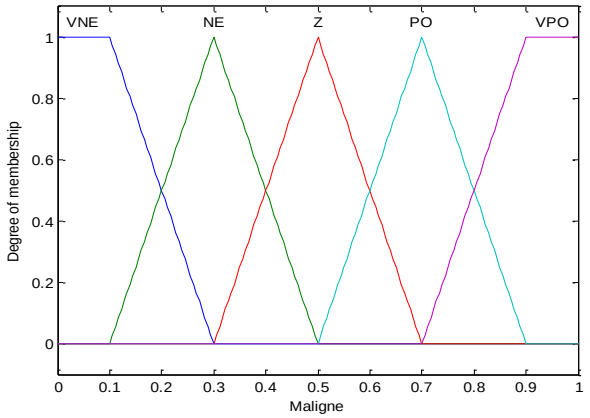


Fig. 13. Membership function of the output

IV. Experimentations and Results

The proposed microcalcifications CAD system was tested on mammograms coming from the MIAS database.

MIAS database is the most used mammographies' database since it can easily be downloaded and exploited. It contains 322 medio-lateral oblique (MLO) mammograms: those whose number is even are left MLO and those whose number is odd are right MLO (Fig. 14).

The 322 images cover all the possibilities of diagnosis: normal (208 images), masses (56 images), microcalcifications (25 images), architectural distortions (18 images) and unbalances (15 images).

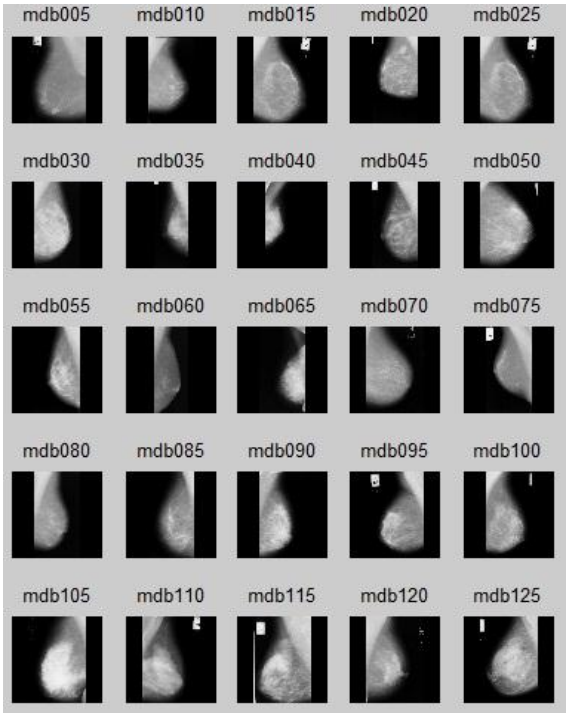


Fig. 14. Examples of mammograms from MIAS database

The efficiency of the proposed CAD system was tested on the 25 mammograms containing microcalcifications clusters.

These images are preprocessed as described previously: artifacts and film boundaries are removed, and then the contrast is enhanced using a soft wavelet coefficients thresholding. After that, the pectoral muscle is detected and removed from the mammograms. The final preprocessing step is the detection and interpolation of the galactophorous tree using an oriented top hat.

Hence enhanced, the mammogram is analyzed using a 64 by 64 pixels sliding window. For each defined block, several GGD features are extracted and then used by a Bayesian back-propagation neural network to detect pixels belonging to microcalcifications cluster.

For each mammographic image, a cross-validation is used, so that the blocks classification is executed 10 times. Each time, training set, validation set and testing set are randomly selected.

The average recognition rate of the 10 tests of the proposed microcalcifications segmentation technique has reached **94.44%**, which is promising compared to segmentation rates given in several works.

In fact, using a neural network provided a recognition rate of 70.8% with textural primitives in [2] and 84% with morphologic features in [63].

In [1], authors used wavelet transform to characterize breast tissue and an SVM classifier and achieved only 79.58% of good detection.

Using neural network with textural primitives provided S. Krishnaveni et al. with a detection rate of 96.25% [45].

TABLE I. Comparison of Microcalcifications Detection Results

Reference	Primitives	Technique	Rate
[2]	Textural	Neural Network	70.8%
[1]]	Wavelets	SVM	79.58%
[63]	Morphological	Neural Network	84%
Proposed approach	GGD	Neural Network	**94.44%**
[45]	Textural	Neural Network	96.25%

The segmented clusters are next characterized and classified into benign and malignant classes as described in the previous section. The proposed classification approach was tested 10 times on segmented microcalcifications clusters from MIAS database. The average classification rate has reached **99%**, which is perfect as a result for a CAD system.

M.J. Bottema et al. used the analysis of the density for the classification of microcalcifications, marking a rate of 69% [8].

C. Anuradha and P. Preeti proposed in [1] malignant/ benign microcalcifications classification approach based on wavelet analysis. They compared two supervised classifiers: SVM whose classification rate reached 69% and an artificial neural network which has provided 96% of good classification.

In [51], Malar et al., authors have developed a classification algorithm which reached a rate of 94%. They used descriptors from wavelet analysis and a supervised classifier (ELM for Extreme Learning Machine).

TABLE II. COMPARAISON OF CLASSIFICATION RESULT WITH OTHER WORKS

Reference	Technique	Rate
[8]	density analysis	69%
[51]	extreme learning machine	94%
[1]	svm	96%
Proposed approach	ANFIS system	**99%**

V. CONCLUSION

In this paper, a new microcalcifications clusters CAD system is proposed. The process begins with a preprocessing step aiming the removal of unnecessary components (artifacts, film boundaries and pectoral muscle) and the enhancement of the mammograms contrast based mainly on detecting and interpolating the galactophorous tree structure.

Microcalcifications segmentation is processed by dividing the enhanced mammograms into 64 by 64 pixels overlapped blocks. Each block is characterized using GGD analysis, and calculated features are used to separate microcalcifications clusters from normal breast tissue via a Bayesian back-propagation neural network. The detection rate has reached **94.44%**.

Finally, three morphologic features are calculated to characterize segmented microcalcifications and an ANFIS system is used to classify these detected lesions into benign and malignant classes. **99%** of microcalcifications clusters were correctly classified.

All tests were carried out with MATLAB R2014 with an Intel Core i5 CPU, 2.53 GHZ and 4GB of RAM.

The proposed approach has proven its efficiency, not only for microcalcifications segmentation and classification, but for breast masses diagnosis as well [50].

The proposed CAD system could be ameliorated by combining 2D-mammograms with 3D-mammograms to boost segmentation and classification accuracy.

REFERENCES

[1] C. Anuradha and P. Preeti, "Detection and classification of microcalcifications using discrete wavelet transform", International Journal of Emerging Trends Technology in Computer Science, 2013, Vol.2, no.4.

[2] K. Arai, I. Abdullah and H. Okumura, "Automated detection method for clustered microcalcification in mammogram image based on statistical textural features", International Journal of Advanced Computer Science and Applications, 2012, Vol.3, no.5, pp.12-16.

[3] R. Babuka and H. Verbruggen, "Neuro-fuzzy methods for linear system identification", Annu. Rev. Control, Vol. 27, 2003, pp.73-85.

[4] M. Bottema and J. Slavotinek, "Detection and classification of lobular and dcis (small cell) microcalcifications in digital mammograms", Pattern Recognition Letters, vol. 21, 2012, pp.1209-1214.

[5] A. Boucher, P.E. Jouve., F. Cloppet, and N. Vincent, "Segmentation du muscle pectoral sur une mammographie", Congrès des jeunes chercheurs en vision par ordinateur, ORASIS'09, Trégastel - France, 2009

[6] H. Boulehmi, H. Mahersia, and K. Hamrouni, " Rehaussement d'images mammographiques pour la segmentation des masses", Conférence Internationale en Traitement et Analyse de l'Information, Methodes et Application-TAIMA, 2015.

[7] H. Boulehmi, H. Mahersia, K. Hamrouni, S. Boussetta and N. Mnif, "Breast cancer detection : A review on mammograms analysis techniques", International IEEE Multi-Conference on Systems, Signals and Devices-SSD, 2013, pp.1-6.

[8] H. Boulehmi, H. Mahersia, K. Hamrouni, S. Boussetta and N. Mnif, "Unsupervised masses segmentation technique using generalized gaussian density", International Journal of Image Processing and Graphics-IJIPG, Vol.1, no.2, 2014, pp.15-24.

[9] S. Bouyahia, J. Mbainaibeye, and N. Ellouze, "Wavelet based microcalcifications detection in digitized mammograms", ICGST-GVIP Journal, January, ISSN 1687-398X, Vol.8, No.5, 2009.

[10] D. Cascio, F. Fauci, R. Magro, G. Raso, R. Bellotti, F. De Carlo, S. Tangaro, G. De Nunzio, M. Quarta, G. Forni, A. Lauria, M.E. Fantacci, A. Retico, G.L. Masala., P. Oliva, S. Bagnasco, S.C. Cheran, and E. Lopez-Torres., "Mammogram segmentation by contour searching and mass lesions classification with neural network", IEEE Trans. Nucl. Sci. Vol.53, No.5, 2006, pp.2827-2833.

[11] C. Castella, "Breast texture synthesis and estimation of the role of the anatomy and tumor shape in the aadiological detection process: from digital mammography to breast tomosynthesis", phd thesis - Ecole Polytechnique Fédérale de Lausanne, thèse no 4347-2009, 20 march 2009

[12] D.M. Catarious, A.H. Baydush, and C.E. Floyd, "Characterization of difference of Gaussian filters in the detection of mammographic regions", Med. Phys., Vol.33, No.11, 2006, pp.4104-4114.

[13] P.C. Chen and T. Pavlidis, "Segmentation by texture using a co-occurrence matrix and a split-and-merge algorithm", Comput. Graph. Image Process., Vol. 10, 1979, pp.172-182.

[14] H. Cheng, X. Cai, X. Chen, L. Hu and X. Lou, "Computer-aided detection and classification of microcalcifications in mammograms : a survey", Pattern Recognition, 2003, Vol. 36, pp. 2967- 2991.

[15] I. Christoyianni, E. Constantinou, and E. Dermatas, "Automatic detection of abnormal tissue in bilateral mammograms using neural networks", Methods and Applications of Artificial Intelligence, 2004, pp.267-275.

[16] L. Clarke, M. Kallergi, W. Qian, H. Li, R. Clark and M. Silbiger, "Three structured nonlinear filter and wavelet transform for microcalcification segmentation in digital mammography",Cancer Letters, vol.77, 1994, pp.173-181.

[17] A.S. Constantinidis, M.C. Fairhust and A.F.R. Rahman, "A new multi-expert decision combination algorithm and its application to the detection of circumscribed masses in digital mammograms", Pattern Recognition, Vol.34, No.8, 2001, pp.1527-1537.

[18] L. Cordella, F. Tortorella and M. Vento, "Combing experts with different features for classifying clustered microcalcifications in mammograms", 15th International Conference on Patten Recognition, 2000, pp.324-327.

[19] J. Daponte and P. Sherman, " Classification of ultrasonic image texture by statistical discriminant analysis and neural networks", Comput. Med. Imag. Graphics, vol.15, 1991, pp.3-9.

[20] D. Davies and D. Dance, "Automatic computer detection of clustered calcifications in digital mammograms", Physics Medicine and Biology, vol.35, 1990, pp.1111-1118.

[21] E.R. Davies, Machine Vision, second ed. Academic Press, London, UK, 1997.

[22] A. Dhawan, Y. Chitre, C. Kaiser-Bonasso and M. Moskowita, "Analysis of mammographic microcalcifications using grey-level image structure features", IEEE Transactions in Medical Imaging, vol.15, 1996, pp.246-259.

[23] J. Dheeba and J. Wiselin, "Detection of microcalcification clusters in mammograms using neural network", International Journal of Advanced Science and Technology, 2010, Vol.19.

[24] M.N. Do, M. Vetterli, Wavelet-based texture retrieval usinggeneralized Gaussian density and Kullback–Leibler distance,IEEE Trans. Image Process., Vol.11, February 2002, pp.146-158.

[25] F. Eddaoudi, F. Regragui, A. Mahmoudi and N. Lamouri, "Masses detection using SVM classifier based on textures analysis", Applied Mathematical Sciences, 2011, Vol.5, no.8, pp.367- 379.

[26] T. Efendigil, S. Önüt, and C. Kahraman, A decision support system for demand forecasting with artificial neural networks and neuro-fuzzy

models: a comparative analysis, Expert Syst. Appl., Vol.36, no.3, 2009, pp.6697-6707.

[27] M. Eltoukhy, I. Faye and B. Samir, "A comparison of wavelet and curvelet for breast cancer diagnosis in digital mammogram", Computers in Biology and Medicine, vol.40, 2010, pp.384-391.

[28] F. Fauci, S. Bagnasco, R. Bellotti, D. Cascio, S.C. Cheran, F. De Carlo, G. De Nunzio, M.E. Fantacci, G. Forni, A. Lauria, E. Lopez Torrez, R. Magro, G.L. Masala, P. Oliva, M. Quarta, G. Raso, A. Retico, and S. Tangaro, "Mammogram segmentation by contour searching and massive lesion classification with neural network", IEEE Nuclear Science Symposium Conference Record, vol.5, pp. 2695-2699, 2004

[29] M. Figueiredo, and F. Gomide, "Design of fuzzy systems using neuro-fuzzy networks", IEEE Trans. Neural Network, Vol.10, No.4, 1999, pp.815-827.

[30] J. Freixenet, A. Oliver, X. Llado, R. Marti, J. Pont, E. Perez, E. Denton, and R. Zwiggelaar, "Eigendetection of masses considering false positive reduction and breast density information", Med. Phys., Vol.35, No.5, 2008, pp.1840-1853.

[31] R. Gordon and R.M. Rangayyan, "Feature enhancement of film mammograms using fixed and adaptive neighborhoods", Appl. Opt., Vol.23, No.4, 1984, pp.560-564.

[32] T.O. Gulsrud, K. Engan, and T. Hanstveit, "Watershed segmentation of detected masses in digital mammograms", IEEE Conference on Engineering in Medicine and Biology Society, 2005, pp.3304-3307.

[33] M.R. Hejazi and Y.S. Ho, "Automated detection of tumors in mammograms using two segments for classification", Lecturer Notes in Computer Science, vol.3767, 2005, pp.910-921.

[34] B.W. Hong and M. Brady, "A topographic representation for mammogram segmentation", Lecturer Notes in Computer Science, vol. 2879, 2003, pp.730-737.

[35] J.S. Jasmine, A. Govardhan, and S. Baskaran, "Classification of Microcalcification in Mammograms using Non-subsampled Contourlet Transform and Neural Network", European Journal of Scientific Research, ISSN 1450-216X, Vol.46, No.4, 2010, pp.531-539.

[36] J.S.R. Jang, "ANFIS: adaptive network based fuzzy inference system", IEEE Trans. Syst. Man Cybern, Vol.23, No.3, 1993, pp.665-684.

[37] L. Jiang, E. Song, X. Xu, G. Ma, and B. Zheng, "Automated Detection of Breast Mass Spiculation Levels and Evaluation of Scheme Performance", Academic Radiology, Vol.15, No.12, December 2008, pp.1534-1544.

[38] I. Kachouri, "Description et classification des masses mammaires pour le diagnostic du cancer du sein", thesis, University Evry-Val d'Essonne, 2012.

[39] K. Kavitha and N. Kumaravel, "A comparative Study of Various MicroCalcification Cluster Detection Methods in Digitized Mammograms, Systems", Signals and Image Processing, 2007, pp.405-409.

[40] S.B. Kazmi, M. QuratulAin, and A. Jaffar, "Wavelets based facial expression recognition using a bank of neural networks", 5th International Conference on Future Information Technology (FutureTech), 2010, pp.1–6.

[41] P. Kestener, "Analyse multifractale 2d et 3d à l'aide de la transformation en ondelettes : application en mammographie et en turbulence développée", thesis, University of Bordeaux I- Ecole doctorale de sciences physiques et de l'ingénieur, 2003.

[42] J. Kim and H. Park, "Statistical textural features for detection of microcalcifications in digitized mammograms", IEEE Trans. Med. Imaging, vol.18, 1999, p.231-238.

[43] G. Kom, A. Tiedeu, M. Kom, C. Nguemgne, and J. Gonsu, "Détection automatique des opacités dans les mammographies par la méthode de minimisation de la somme de l'inertie", Elsevier ITBM-RBM, Vol.26, 2005, pp.347–356.

[44] D. Kramer and F. Aghdasi, "Texture analysis techniques for the classi"cation of microcalcifications in digitized mammograms", Proceedings of the fifth IEEE AFRICON Conference Electrotechnical Service for Africa, September 28–October 1, 1999, pp.395–400.

[45] S. Krishnaveni, R. Bhanumathi and T. Pugazharasan, "Study of mammogram microcalcification to aid tumour detection using naive bayes classifier", International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering, 2014, Vol.3, no.3.

[46] N. Kumar, S. Amutha, and D.R. Ramesh, "Enhancement of mammographic images using morphology and wavelet transform", Int. J. Comput. Technol. Appl., Vol.3, No.1, 2012, pp.192-198.

[47] A.F. Laine, S. Schuler, J. Fan, and W. Huda, "Mammographic Feature Enhancement by Multiscale Analysis", IEEE Transaction on Medical Imaging, vol.13, No.4, 1994, pp.725-740.

[48] L. Li, R.A. Clark and J.A. Thomas, "Computer-aided diagnosis of masses with fullfield digital mammography", Acad. Radiol., Vol.9, No.1, 2002, pp.4-12.

[49] H. Mahersia, "Contribution à l'analyse de texture par des méthodes spatio-frquentielles", National Engineering Schoolof Tunis, Tunisia, 2010, PhD thesis.

[50] H. Mahersia, H. Boulehmi and K. Hamrouni, "Development of intelligent systems based on bayesian regularization network and neuro-fuzzy models for mass detection in mammograms: A comparative analysis", Computer Methods and Programs in Biomedicine, 2015.

[51] E. Malar, A. Kandaswamy, D. Chakravarthy and A. GiriDharan, "A novel approach for detection and classification of mammographic microcalcifications using wavelet analysis and extreme learning machine", Computers in Biology and Medicine, vol.42, 2012, pp.898-905.

[52] J. Marti, J. Batle, X. Cufi, and J. Espanol, "Microcalcification evaluation in computer assisted diagnosis for digital mammography", Colloquium on Digital Mammography, 1999, pp.1-6.

[53] J. Martý, J. Freixenet, X. Muñoz, and A. Oliver, "Active region segmentation of mammographic masses based on texture, contour, and shape features", Lecturer Notes in Computer Science, vol.2652, 2003, pp. 478-485.

[54] M.B. McSweeney, P. Sprawls, and R.L. Egan, "Enhanced image mammography", AJR 140, 1983, pp.9-14.

[55] A.J. Méndez, M. Souto, P.G. Tahoces, and J.J. Vidal, "Computer aided diagnosis for breast masses detection on a telemammography system", Comp. Med. Imag. Grap., Vol.27, 2003, pp.497–502.

[56] W.M. Morrow, R.B. Paranjape, R.M. Rangayyan, and J.E.L. Desautels, "Region-based contrast enhancement of mammograms", IEEE Trans. Med. Imag., Vol.11, No.3, 1992, pp.392-406.

[57] R. Nakayama, Y. Uchiyama and K. Namba, "CAD scheme using filter bank for detection of Mcc in mammograms", IEEE Transactions on Bio Medical Engineering, 2006, Vol.53, no.2, pp.273-283.

[58] L. Nanni, S. Brahnam, and A. Lumini, "Combining different local binary pattern variants to boost performance", Expert Systems with Applications, Vol. 38, 2011, pp.6209-6216.

[59] T. Netsch and H.O. Peitgen, "Scale-space signatures for the detection of clustered microcalcifications in digital mammograms", IEEE Trans. Med. Imag., Vol.18, No.9, 1999, pp.774-786.

[60] A. Oliver, J. Freixenet, R. Marti and R. Zwiggelaar, "A comparison of breast tissue classification techniques", MICCAI, 2006, pp.872-879.

[61] A. Oliver, J. Freixenet, R. Marti, J. Pont, E. Perez, E. Denton, and R. Zwiggelaar, "A Novel Breast Tissue Density Classification Methodology", IEEE Transactions on Information Technology in Biomedicine, Vol.12, No.1, january 2008

[62] A. Oliver, J. Freixenet, R. Marti, E. Perez, J. Pont, E. Denton and R. Zwiggelaar, "A review of automatic mass detection and segmentation in mammographic images", Medical Image Analysis, vol.14, 2010, pp.87-110.

[63] A. Papadopoulos, I. Fotiadisb and A. Likas, "Characterization of clustered microcalcifications in digitized mammograms using neural networks and support vector machines", Artificial Intelligence in Medicine, 2004.

[64] A. Pandey and S. Singh, "New performance metric forquantitative evaluation of enhancement in mammograms", 2nd International Conference on Information Management in the Knowledge Economy (IMKE), 2013, pp.51-56.

[65] S. Paquerault, N. Petrick, H.P. Chan, B. Sahiner, and M.A. Helvie, "Improvement of computerized mass detection on mammograms: fusion of two-view information", Med. Phys., Vol.29, No.2, 2002, pp.238–247.

[66] S. Petroudi, T. Kadir and M. Brady, "Automatic classification of mammographic parenchymal patterns : A statistical approach", Engineering in Medicine and Biology Society, vol.1, 2003, pp.798-801.

[67] S. Sharma and A. Oberoi, "A new approach for Classification and Detection of Suspicious Lesions in Mammograms based on Adaptive Thresholding", International Conference on Advanced Computing, Communication and Networks'11, 2011, pp.427-431.

[68] J. Sharma, J.K. Rai, and R.P. Tewari, "Identification ofpre-processing technique for enhancement of mammogramimages", International Conference on Medical Imaging, m-Health and Emerging Communication Systems (MedCom), 2014, pp.115-119.

[69] T. StojiC, I. Reljin, and B. Reljin, "Local contrast enhancement indigital mammography by using mathematical morphology", International Symposium on Signals, Circuitsand Systems, ISSCS 2005, vol.2, 2005, pp.609–612.

[70] Y. Sun, J. Suri, Z.Y. Rangaraj, M. Rangayyan, and R. Janer, "Effect of adaptive neighborhood contrast enhancement on the extraction of the breast skin line in mammograms", 27th IEEE Annual conference on Engineering in Medicine and Biology Society, 2005, pp.3475-3478.

[71] K.S. Woods and K.W. Bowyer, "Computer detection of stellate lesions", International Workshop on Digital Mammography, 1994, pp.221–229.

[72] Y. Xiao, L. Ma, and K. Khorasani, "A new facial expression recognition technique using 2-D DCT and neural networks based decision tree", International JointConference on Neural Networks, 2006, pp.2421-2428.

[73] S. Yu, A. Zhang, and H. Li, "A review of estimating the shape parameter of generalized Gaussian distribution", J. Comput.Inf. Syst., Vol.8, No.21, 2012, pp.9055-9064.

[74] Z. Zhang, J. Lu and J. Yip, "Computer aided mammography", School of Computing and Engineering Researchers Conference, 2008.

[75] Y. Zheng, "Breast cancer detection with Gabor features from digital mammograms", Algorithms, vol.3, 2010, pp.44-62

Secure High Dynamic Range Images

Secure HDR Images

Med Amine Touil, Noureddine Ellouze
Dept. of Electrical Engineering
National Engineering School
Tunis, Tunisia

***Abstract*—In this paper, a tone mapping algorithm is proposed to produce LDR (Limited Dynamic Range) images from HDR (High Dynamic Range) images. In the approach, non-linear functions are applied to compress the dynamic range of HDR images. Security tools will be then applied to the resulting LDR images and their effectiveness will be tested on the reconstructed HDR images. Three specific examples of security tools are described in more details: integrity verification using hash function to compute local digital signatures, encryption for confidentiality, and scrambling technique.**

Keywords—high dynamic range; tone mapping; range compression; integrity verification; encryption; scrambling; inverse tone mapping; range expansion

I. Introduction

Thanks to recent advances in computer graphics and in vision, HDR imaging has become a new generation technology becoming a new standard representation in the field of digital photography. Advances in techniques, equipment acquisition and display, handsets with the powerful increasing of processors in professional and consumer devices, as well as the continued efforts to get content more photo-realistic with higher quality image and video; have attracted attentions to HDR imaging.

Nowadays, several industrials offer cameras and displays capable of acquiring and rendering HDR images. However, the popularity and the public adoption of HDR images are hampered by the lack of file formats, compression standards and security tools.

In this paper, mechanisms suitable for HDR images are developed to protect privacy and to minimize risks to confidential information.

The structure of this paper is the following. The existing standard is first reviewed in Section 2. Three specific use cases are then discussed dealing with integrity verification, encryption and scrambling in Section 3. Some conclusions are finally drawn in Section 4.

II. Overview

The protection of privacy is important in our civilization and is also essential in several social functions. However, this fundamental principle is rapidly eroding due to the intrusion tolerated by some modern information technology. In particular, the protection of privacy is becoming a central issue in the transfer of images through open networks and especially in video surveillance systems.

The digital images are distributed via the network so they can be easily copied and modified legally and/or illegally. In this spirit, there has been a strong demand for a security solution JPEG2000 images. To meet this demand, the JPEG [1][2] committee has created an extension of the JPEG2000 [3][4] encoder by integrating security tools such as integrity verification, encryption and scrambling. This extension is part 8 of standard JPEG2000 coder (JPEG2000 Part 8) designated by JPSEC. JPSEC [5][6] defines the framework, concepts and methods for the safety of JPEG2000. It specifies a specific syntax for the encoded data and provides protection JPEG2000 bit stream. The syntax defines the security services associated with the image data, the tools required for each service and how to apply its tools, and parts of the image data to be protected.

The problem of protection of visual privacy in digital image and video data has attracted much interest lately. The capacity of HDR imaging to capture fine details in contrasting environments, making dark and bright areas clear, has a strong implication on privacy. However, the point at which the HDR representation affects privacy if used instead of the SDR (Standard Dynamic Range) is not yet clear and the scenarios of use are not fully understood. Indeed, there is no mechanism of protection of privacy specific to HDR representation. Currently, many challenges are open for research related to the intrusion in privacy for HDR images.

As part of this paper, mechanisms adapted to HDR images are developed to protect privacy. These mechanisms should essentially meet the expectations of consumers concerned about the respect of their privacy and ethics.

III. Methodology

In this section, a system based on DCT (Discrete Cosine Transform) is proposed to secure HDR images.

For tone mapping (Fig. 1), sub-bands architectures [7] are applied using a multi-scale decomposition with Haar pyramids splitting a signal into sub-bands which are rectified, blurred, and summed to give an activity map. A gain map is derived from the activity map using parameters which are to be specified. Each sub-band coefficient is then multiplied by the gain at that point, and the modified sub-bands are post-filtered and summed to reconstruct the result image.

(a) (b)

Fig. 1. Tone mapping (a) HDR image, (b) LDR image

Hereafter, three specific examples of security tools are described in more details: integrity verification, encryption and scrambling.

A. *Integrity Verification*

To detect manipulations to the image data, the integrity verification (Fig. 2) is used where a bit exact verification is considered in this use case as a technique applied in the transform-domain based on hash function and digital signature. More specifically, hash function SHA-1 [8] is applied on the DCT coefficients generating a 160 bits value which is then encrypted by the public-key algorithm RSA [9] to generate a digital signature. It is possible to use other hash functions and encryption algorithms obviously. If new hash values computed and compared with those decrypted are not equal or if the digital signature is missing then an attack is detected. Enabling to locate a potential attack, the integrity verification is performed for each macro-block.

While a single digital signature is computed to verify the integrity of the whole image, multiple ones are computed to identify locations in the image data where the integrity is in doubt. As an alternative, a digital signature is generated for each macro-block composed of several DCT blocks not for each 8x8 DCT block because of a significant increase of the overall bit-rate resulted from a very large number of digital signatures and added bytes.

An original and a tampered image are shown in Fig. 3. Integrity verification is performed on macro-blocks composed of 100 DCT blocks corresponding to square shaped regions of 80x80 pixels. The attack is identified in the upper left 160x80 pixels by a comparison between the hash values obtained from the two images.

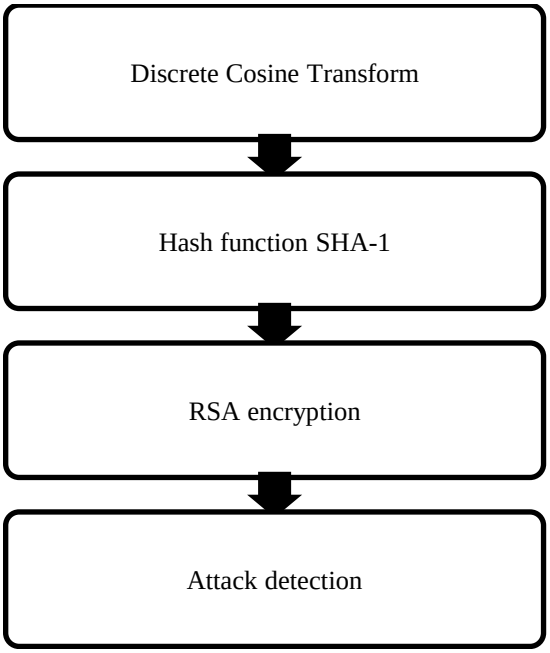
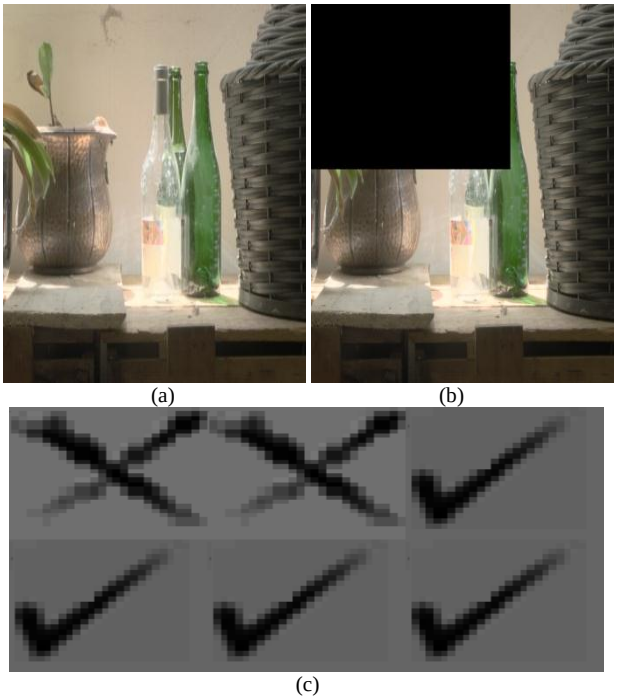


Fig. 2. Flowchart of integrity verification



(a) (b)

(c)

Fig. 3. Example of integrity verification (a) original image, (b) tampered image, (c) digital signature verification

B. Encryption

For confidentiality, the use case of encryption (Fig. 4) is now considered. The preferred approach is to apply encryption in the transform-domain. More specifically, encryption is applied on the quantized DCT coefficients. Authorized users are able to decrypt and recover the original data. The whole image or alternatively the ROI (Region of Interest) AES [10] encryption is restricted to selected DCT blocks.

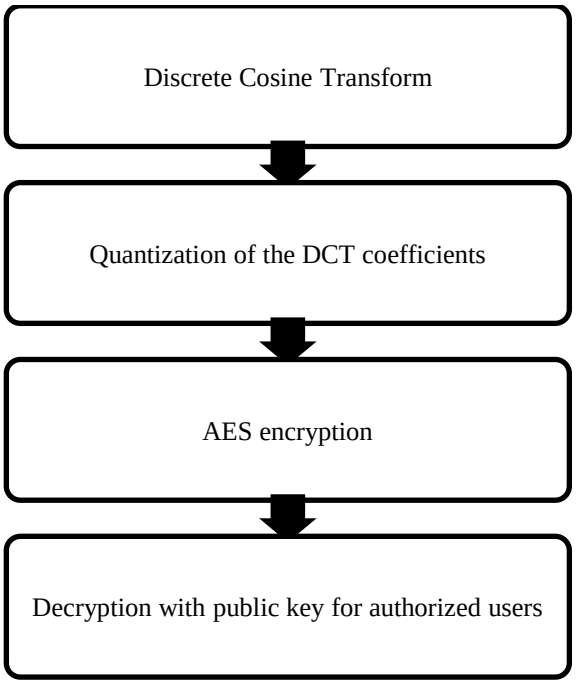
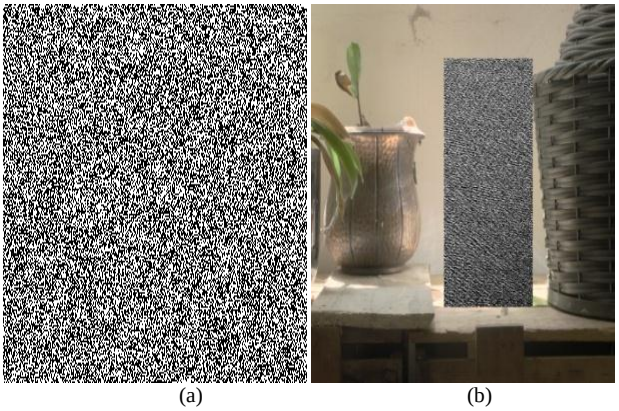


Fig. 4. Flowchart of encryption

An example of a whole image or a ROI encryption is shown in Fig. 5 by restricting the shape of the encrypted region to match the 8x8 DCT blocks boundaries.



(a) (b)

Fig. 5. Example of encryption (a) whole image encryption, (b) region of interest encryption

C. Scrambling

Image and video content is characterized by a very high bit-rate and a low commercial value when compared to other types of information such as banking data and confidential documents. Conventional encryption techniques entail a non-negligible complexity increase and are therefore not optimal in this case.

While keeping complexity very low, scrambling (Fig. 6) is a compromise to protect image and video data. A scrambling technique applied on the quantized DCT coefficients is considered in this use case. Authorized users perform unscrambling of the coefficients allowing for a fully reversible process for them.

In the way confidentiality will be guaranteed, a pseudo-random noise consisting in pseudo-randomly inverting the sign of the quantized DCT coefficients is introduced. The whole image or alternatively the ROI scrambling is restricted to fewer DCT coefficients as in the previous case. The technique requires negligible computational complexity as it is merely flipping signs of selected DCT coefficients where other extensions such as flipping of least or most significant bits of the quantized DCT coefficients can be considered.

Initialized by a seed value, PRNG (Pseudo Random Number Generator) is used to drive the scrambling process. Multiple seeds can be used to improve the security of the system where they are encrypted using RSA in order to communicate the seed values to authorized users.

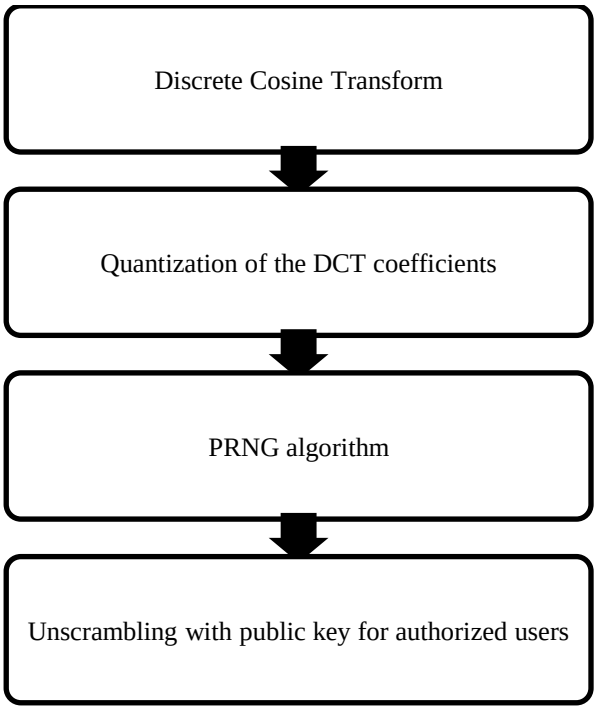


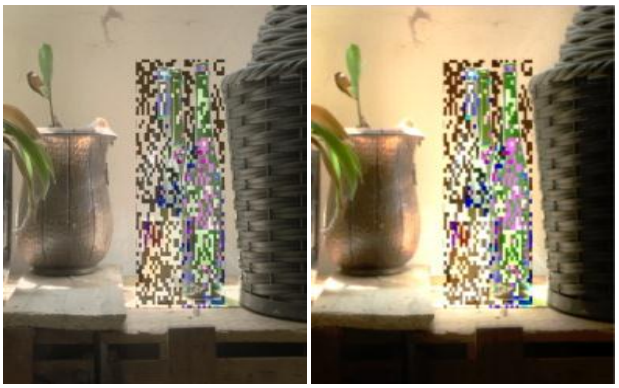
Fig. 6. Flowchart of scrambling

An example of a whole image or a ROI scrambling is shown in Fig. 7 by restricting the shape of the scrambled region to match the 8x8 DCT blocks boundaries.

A reconstructed HDR image is shown in Fig. 8 after inverse tone mapping of the resulting LDR image.

(a) (b)

Fig. 7. Example of scrambling (a) whole image scrambling, (b) region of interest scrambling



(a) (b)

Fig. 8. Inverse tone mapping (a) resulting LDR image, (b) reconstructed HDR image

IV. Conclusions

In this paper, a system consisting of security tools was introduced to provide services similar to those in the standard JPSEC. As illustrative use cases, three specific examples of security tools are described in more details: integrity verification, encryption and scrambling techniques. Indeed, the system allows the use of different tools in support of a number of security services.

As perspective, the scrambling technique implemented will be integrated in a video coding system adapted to HDR image sequences. Specifically, the scrambling process will be directly applied to the DCT coefficients after quantization and before entropy coding. Authorized users perform unscrambling (inverse scrambling) of the resulting DCT coefficients of entropy decoding at the decoder side. Different results will be presented in terms of subjective and objective measure of the quality and scrambling force.

References

[1] G. K. Wallace, "The JPEG Still Picture Compression Standard", Communications of the ACM, vol. 34, no. 4, pp. 31-44, 1991.

[2] W. B. Pennebaker and J. L. Mitchell, "JPEG: Still Image Data Compression Standard", Van Nostrand Reinhold, New York, 1993.

[3] A. Skodras, C. Christopoulos and T. Ebrahimi "The JPEG 2000 Still Image Compression Standard", IEEE Signal Processing Magazine , vol. 18, no. 5, pp. 36-58, Sept. 2001.

[4] D. Taubman and M. Marcellin, "JPEG 2000: Image Compression Fundamentals, Standards and Practice", Kluwer Academic Publishers, 2002.

[5] "JPSEC Final Draft International Standard", ISO/IEC JTC1/SC29/WG1/N3820, Nov. 2005.

[6] J. Apostolopoulos, S. Wee, F. Dufaux, T. Ebrahimi, Q. Sun and Z. Zhang, "The Emerging JPEG 2000 Security (JPSEC) Standard", in IEEE Proc. Int. Symp. on Circuits and Systems (ISCAS), Island of Kos, Greece, May 2006.

[7] Y. Li, L. Sharan and E. H. Adelson, "Compressing and Companding High Dynamic Range Images with Subband Architectures", ACM Transactions on Graphics (TOG), 24(3), Proceedings of SIGGRAPH, 2005.

[8] FIPS PUB 180-1, "Secure Hash Standard (SHS)", NIST, April 1995.

[9] R. L. Rivest, A. Shamir and L. M. Adleman, "A Method for Obtaining Digital Signatures and Public-Key Cryptosystems", Communications of the ACM, vol. 21, no. 2, pp. 120-126, 1978.

[10] FIPS PUB 197, "Advanced Encryption Standard (AES)", NIST, November 2001.

A Subset Feature Elimination Mechanism for Intrusion Detection System

Herve Nkiama
Faculty of Computing
University Technology Malaysia
Skudai 81310, Johor Barhu - Malaysia

Syed Zainudeen Mohd Said
Faculty of Computing
University Technology Malaysia
Skudai 81310, Johor Barhu - Malaysia

Muhammad Saidu
Faculty of Computing
University Technology Malaysia
Skudai 81310, Johor Barhu – Malaysia

***Abstract*—several studies have suggested that by selecting relevant features for intrusion detection system, it is possible to considerably improve the detection accuracy and performance of the detection engine. Nowadays with the emergence of new technologies such as Cloud Computing or Big Data, large amount of network traffic are generated and the intrusion detection system must dynamically collected and analyzed the data produce by the incoming traffic. However in a large dataset not all features contribute to represent the traffic, therefore reducing and selecting a number of adequate features may improve the speed and accuracy of the intrusion detection system. In this study, a feature selection mechanism has been proposed which aims to eliminate non-relevant features as well as identify the features which will contribute to improve the detection rate, based on the score each features have established during the selection process. To achieve that objective, a recursive feature elimination process was employed and associated with a decision tree based classifier and later on, the suitable relevant features were identified. This approach was applied on the NSL-KDD dataset which is an improved version of the previous KDD 1999 Dataset, scikit-learn that is a machine learning library written in python was used in this paper. Using this approach, relevant features were identified inside the dataset and the accuracy rate was improved. These results lend to support the idea that features selection improve significantly the classifier performance. Understanding the factors that help identify relevant features will allow the design of a better intrusion detection system.**

***Keywords*—classification; decision tree; features selection; intrusion detection system; NSL-KDD; scikit-learn**

I. INTRODUCTION

With the recent advance in technologies where concepts like Cloud Computing, Big Data, and Social Media Network have emerged, our society produce enormous quantity of data. Finding useful information among this immense data generated by these technologies became critical for marketers, data scientist and even business corporate. With this amount of data transmitted over a network or internet, security becomes a major concern, although multiple intrusion prevention technologies have been built in the past decade to eliminate potential threats despite that, attacks still continue and increase in complexity, this is the reason there is a need of a mechanism to detect any suspicious or unwanted traffic which may cause damage on a particular network.

This security mechanism can be implemented using an Intrusion Detection System (IDS) which can be describe as a collection of software or hardware device able to collect, analyze and detect any unwanted, suspicious or malicious traffic either on a particular computer host or network[1]. Therefore to achieve its task, an IDS should use some statistical or mathematical method to read and interpret the information it collects and subsequently reports any malicious activity to the network administrator [2].

There still exist one main issue regarding the actual intrusion detection technique that is the involvement of human interaction when it comes to label the traffic between an intrusion and a normal one, another major concern is the new challenge of "Big Data" and "Cloud Computing". These two ubiquitous technologies produce a large amount of data that must be collected and analyzed by the intrusion detection engine dynamically and often the IDS needs to deal with a multi-dimensional data generated by these large quantities of data. It is necessary to consider that the intrusion dataset can be huge in size, not only the number of observations grown, but the number of observed attributes can also increase significantly and may generated a considerably amount of false positives results as it can contain many redundant or duplicate records [3].

A data clean process can require a tremendous human effort, which is an extensive time consuming and expensive [4]. A machine learning approach and data mining technique which is the application of machine learning methods to large database are widely known and used to reduce or eliminate the need of a human interaction.

Machine learning helps to optimize performance criterion using example data or past experience using a computer program, models are defined with some parameters, and learning is the execution of the programming computer to optimize the parameters of the model using a training data. The model can be predictive to make predictions in the future, or descriptive to gain knowledge from data. To perform a predictive or descriptive task, machine learning generally use two main techniques: Classification and Clustering. In classification, the program must predict the most probable category, class or label for new observation into one or multiple predefined classes or label while clustering, the classes are not predefined during the learning process.

However if the purpose of the IDS is to differentiate between normal or intrusion traffic, classification is recommended and if we seeks to identify the type of intrusion, clustering can be more helpful [5].

However, a lot of researchers have suggested to use the KDD dataset to detect an attack [6][7] in the past. Unfortunately, these proposal have failed to ensure a good performance in terms of detection rate. Moreover those existing IDS aims to analyze all features which can result a misclassification of intrusion and quite amount of time when building the model, despite some concern and critic about the system evaluation of the KDD dataset [8], research still use it to test their model. Thus, in this paper a method has been suggested for selecting and identifying relevant features on the NSL-KDD dataset which is an improvement of the previous one [9].

The rest of this paper is divided as followed: Section II – Description of the NSL-KDD Dataset, Section III- Previous works, Section IV – Methodology, Section V- Experiment and Results and finally the Section VI- Conclusion and Further Works.

II. Description of NSL-KDD

The KDD 1999 dataset was developed by the MIT Lincoln Labs [10] and was extensively used by researchers during the last decade. The entire dataset is very large in size and contains many attributes variables. Therefore to improve the machine learning computation, 10 % of it was extracted and adopted as training dataset in the intrusion detection process. However, some inherent drawback was made about this dataset [8][9]. The KDD 99 contains an important quantities of redundant records which has as consequence to prevent the learning algorithm to perform well. In addition, duplicate records found in the test dataset cause the evaluation result to be biased by the method used during the detection rates results.

To resolve some issues found in the previous KDD 99, an improved version was created, the NSL KDD dataset which can be available at [11]. The reason behind the use of this dataset has been reported at [9] among them the following are relevant to mention:

- Elimination of redundant records in the training set will help our classifier to be unbiased towards more frequent records.
- No presence of duplicate records in the test set, therefore, the classifier performance will not be biased by the techniques which have better detection rates on the frequent records.
- The training and test set contains both a reasonable numbers of instances which is affordable for experiments on the entire set without the need to randomly choose a small portion.

The NSL KDD dataset contains four main files as describe in the Table 1.

TABLE I. NSL KDD Dataset Description

Name of the Files	Description
KDDTrain+.TXT	It is the full training set including attack-type labels and difficulty level in csv format
KDDTest+.TXT	It is the full test set including attack-type labels and difficulty level in csv format
KDDTrain+_20Percent.TXT	20% subset of the KDDTrain+.txt
KDDTest-21.TXT	A subset of the KDDTest+.txt file which does not include records with difficulty level of 21 out of 21

In this paper, the KDDTain+.TXT and the KDDTest+.TXT which consists of 126,620 and 22,850 records respectively were used. The training and test set contain both 41 features labeled as normal traffic or specific attack types, all these features are subdivided in 4 categories [12][13]: basic features, time-based traffic features, content features and host-based traffic features.

All categories are described below:

- Basic features: It contains all features which derived from TCP/IP connection such as Protocol_type, Service, duration and etc.
- Time-based traffic features: It is used to capture those features which are mature over a 2 second temporal window (e.g. count, srv_count, Rerror_rate and etc.)
- Content features: Those features use domain knowledge to access the payload of the original TCP packets (e.g. hot, num_root, is_guest_login and etc.)
- Host-based traffic features: all attacks which span longer than 2 second intervals that have the same destination host as the current connection are access using these features (e.g. dst_host_count, dst_host_srv_count and etc.)

The classes or labels in the NSL KDD dataset are divided into four categories which represent the attack class and one as normal traffic [12]:

1) Denial of Service (DoS): This attack aims to block or restrict a computer system or network resources or services.

2) Probe: here the intruder aims to scan for information or vulnerabilities in a network or computer system which later on will be used to launch attacks.

3) Remote to Local (R2L): Here the intruder gain remotely unauthorized access to a computer system over a network by sending data packet to that system.

4) User to Root (U2R): Here the intruder gains access to a user with normal privilege and later on try to access a user with administrator or root privilege.

The Table 2 and 3 describe and explain the analysis of the attack classes and types in the NSL_KDD dataset in details and shows the number of individual instances and records, both in the training and testing set.

TABLE II. NUMBER OF INSTANCES IN THE TRAINING SET

Attack Classes or Labels	Attack types (number of instances)	Total of instances
DoS	back (956), land(18), neptune(41,214), pod(201), smurf(2,646), teardrop(892)	45,927
Probe	satan(3,633), ipsweep(3,599), nmap(1,493), portsweep (2,931)	11,656
R2L	guess_passwd(53), ftp_write(8), imap(658), phf(4), multihop(7), warezmaster(20), warezclient(890), spy(2)	1,642
U2R	buffer_overflow(30), loadmodule(9), rootkit(10), perl(3)	52
Grand Total	59,277	

The normal traffic contains 67,343 instances which brings a total of 126,620 instances in the training set.

TABLE III. NUMBER OF INSTANCES IN THE TEST SET

Attack class or label	Attack types (number of instances)	Total of instances
DoS	back(359), land(7), neptune(4,657), apache2(737), pod(41), smurf(665), teardrop(12), udpstorm(2), processtable(685), worm(2), mailbomb(293)	7,460
Probe	Satan(735), ipsweep(141), nmap(73), portsweep(157), mscan(996), saint(319)	2,421
R2L	guess_passwd(1,231), ftp_write(3), imap(307), xsnoop(4), phf(2), multihop(18), warezmaster(944), xlock(9), snmpguess(331), snmpgetattack(178), httptunnel(133), sendmail(14), named(17)	3,191
U2R	Buffer_overflow(20), loadmodule(2), xterm(13), rootkit(13), perl(2), sqlattack(2), ps(15)	67
Grand Total	13,139	

The normal traffic contains 9,711 instances which brings a total of 22,850 instances in the test set. More details on the features names and descriptions can be found at [13]

III. PREVIOUS WORK

Most of the proposed research system could effectively utilize feature selection process to improve detection rate of their system and minimize considerably the false alarm rate. Research usually missed to detect new intrusions, especially when the intrusion mechanism used differed from the previous intrusion.

In 2009, Shi-Jinn [14] works revealed that not all research carried out feature selection before they trained their classifier, however based on [15][16], this processes takes a significant part to different types of intrusion identification and features can be excluded without the performance of the IDS to be dropped. Juan Wang et al., in their work [17] proposed a decision tree based algorithm for intrusion detection, even if during their experiments the C4.5 algorithm was achieving a good detection accuracy, the error rate was remaining identical.

Back in 2010, Farid et al. [18], used a decision tree based learning algorithm to retrieve important features set from the training dataset for intrusion detection. Their techniques found relevant features using a combination of ID3 and C4.5 decision tree algorithms. They assigned a weight value to each features. The weight is determined where the minimum depth of the decision tree at which each feature is checked inside the tree and the weights of features that do not appear in the decision tree are allocated a value of zero. Ektefa et al. [19], used different data mining method for intrusion detection and they found that the decision tree classifier was performing better than the SVM learning algorithm.

Geetha Ramani et al. [20] used in their paper in 2011, a statistical method for analyzing the KDD 99 dataset. They identified the important features by studying the internal dependences between features.

In their paper proposed in 2012, S. Mukherjee and N. Sharma [21] designed a technique called Feature- Vitality Based Reduction Method (FVBRM) using a Naïve Bayes classifier. FVBRM identifies important features by using a sequential search approach, starting with all features, one feature is removed at a time until the accuracy of the classifier reaches some threshold. Their method shows an improvement of the classification accuracy but takes more time and still complex when detecting the U2R attacks.

In 2013, support vector machine classifier was used by Yogita B. Bhavsar et al. [22], for intrusion detection using the NSL KDD dataset. The drawback with this technique is the extensive training time required by the classifier, so to reduce the time, they applied a radial basis function (RBF) to reduce the extensive time.

In 2014, O. Y. Al-Jarrah et al. [23], used an ensembles of decision-tree based voting algorithm with forward selection / backward elimination feature raking techniques using a Random Forest Classifier. Their method shows an improvement of detection accuracy when selected important features and it can be suitable for large-scale network.

N. G. Relan and D. R. Patil [24] in their papers have tested two decision tree approach to classify attacks using the NSL KDD dataset. They have found that the C4.5 with pruning offers better accuracy than the C4.5 without pruning and it was necessary to reduce the number of features because using all features degrades the performance of the classifier also its time consuming. After analyzing some previous works, the reasons most of researchers are interested in selecting and identifying relevant features are described as follow:

- In most learning algorithms, the complexity depends on the number of input dimensions, *d*, as well as on the size of the data sample, *N*, and for reduced memory and computation, researchers are interested in selecting

relevant and important feature to reduce the dimensionality of the problem. Decreasing *d* also decreases the complexity of the inference algorithm during testing.

- When an input is decided to be unnecessary, the cost of extracting it is saved.
- Simpler models are more robust on small datasets. Simpler models have less variance, that is, they vary less depending on the particulars of a sample, including noise, outliers, and so forth.
- When data can be explained with fewer features, better idea about the process that underlies the data can be obtained and this allows knowledge extraction.

When data can be represented in a few dimensions without loss of information, it can be plotted and analyzed visually for structure and outliers.

IV. METHODOLOGY

A. Scikit-Learn Description

As stated before, during this experiment *scikit-learn* [25] was used, which is a machine learning library written in python. Most of the learning algorithm implement in *scikit-learn* required data to be stored in a two-dimensional array or matrix. The size of the expected matrix is [samples, features].

The first parameter defines the number of samples, each sample is an item to be processed and the second parameter is the number of features that can be used to describe each item in a quantitative manner, generally real-valued but may be Boolean or discrete-valued in some cases. Data in *scikit-learn* is represented as a feature matrix and a label vector. Fig. 1 shows the data representation in scikit-learn.

$$\text{feature matrix}: \quad \mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1D} \\ x_{21} & x_{22} & \cdots & x_{2D} \\ x_{31} & x_{32} & \cdots & x_{3D} \\ \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{ND} \end{bmatrix}$$

$$\text{label vector}: \quad \mathbf{y} = [y_1, y_2, y_3, \cdots y_N]$$

Fig. 1. Data representation in *scikit-learn*

Here, *N* are samples and *D* features

B. Expriment Methodology

The experiment methodology used in this paper, is illustrated in the Fig. 2 and describe as follow:

Step 1: Data Cleaning and Pre-processing

Basically in this step the dataset has to go through a cleaning process to remove duplicate records, as the NSL KDD dataset was employed which has already been cleaned, this step is not anymore required. Next a Pre-processing operation has to be taken in place because the dataset contains numerical and non-numerical instances. Generally the estimator (classifier) defines in the *scikit-learn* works well with numerical inputs, so a one-of-K or one-hot encoding method is used to make that transformation. This technique will transforms each categorical feature with *m* possible inputs to *n* binary features, with one active at the time only

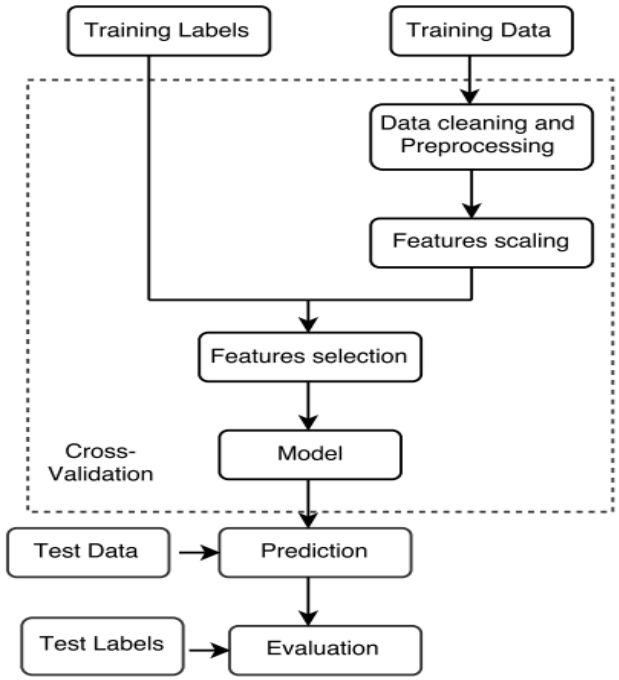


Fig. 2. Experiment methodology

Step 2: Features scaling

Features scaling is a common requirement of machine learning methods, to avoid that features with large values may weight too much on the final results. For each feature, calculate the average, subtract the mean value from the feature value, and divide the result by their standard deviation.

After scaling, each feature will have a zero average, with a standard deviation of one.

Step 3: Features Selection

Feature selection is used to eliminate the redundant and irrelevant data. It is a technique of selecting a subset of relevant features that fully represents the given problem alongside a minimum deterioration of presentation [26], two possible reason were analyzed why it would be recommended to restrict the number of features:

Firstly, it is possible that irrelevant features could suggest correlations between features and target classes that arise just by chance and do not correctly model the problem. This aspect is also related to over-fitting, usually in a decision tree classifier. Secondly, a large number of features could greatly increase the computation time without a corresponding classifier improvement.

The feature selection process starts with a univariate feature selection with ANOVA F-test for feature scoring, univariate feature selection analyzes each feature individually to determine the strength of the relationship of the feature with

labels. The *SelectPercentile* method in the *sklearn.feature_selection* module were used, this method select features based on a percentile of the highest scores. Once, the best subset of features were found, a recursive feature elimination was applied which repeatedly build a model, placing the feature aside and then repeating the process with the remained features until all features in the dataset are exhausted. As such, it is a good optimization for finding the best performing subset of features. The idea is to use the weights of a classifier to produce a feature ranking.

Step 4: Model

Here, a decision tree model was built to partition the data using information gain until instances in each leaf node have uniform class labels. This is a very simple but yet an effective hierarchical method for supervised learning (classification or regression) whereby the local space (region) is recognized in a sequence of repetitive splits in a reduced number of steps (small). At each test, a single feature is used to split the node according to the feature's values. If after the split, for every branches, all the instances selected belong to the similar class, the split is considered complete or pure.

One of the possible method to measure a good split is entropy or information gain. Entropy is an information-theoretic measure of the 'uncertainty' found in a training set, because of the existence of more than one possible classification. The training set entropy is represented by *H*. It is calculated in 'bits' of information and it is described as:

$$H = -\sum_{i=1}^{n} \mathrm{P}(c_i)\ log_2\ \mathrm{P}(c_i)$$

The generation process of a decision tree done by recursively splitting on features is equivalent to dividing the original training set into smaller sets recursively until the entropy of every one of these subsets is zero (i.e everyone will have instances from a single class target).

A Decision Tree is made up internal decision nodes and terminal leaves. A test function is implemented by each decision node with a discrete results labelling the branches. Providing an input, at every node, a test is constructed and based on the outcome, one of the branches will be considered. Here the learning algorithm starts at the root and until a leaf node is reached, the process will be done recursively at which moment the value represented in the leaf node is the output. Every leaf node possesses an outcomes label, which it is the class target in case of classification and numeric value for regression. A leaf node can describe a localized space or region where instances finding in this input space (region) possess the same labels for classification and similar numeric value for regression

Step 5: Prediction and Evaluation

The test data was used to make prediction of our model and for evaluation, multiple settings was considered such as the accuracy score, precision, recall, f-measure and a confusion matrix. A 10-fold cross-validation was performed during all the process.

V. Experiment and Results

A. Experiment

The Decision Tree learning algorithm was used in the experiment. Decision Tree tends sometimes to over-fitting, so to find the best parameters to fit the model, an exhaustive grid search parameters tuning was computed and information gain is used to select features. Hence, building from the training data, a tree was obtained with its leaves being class labels. When building a decision tree, only one feature is used at a time to split the node and partition the data. Hence, features are used in a univariate manner.

After obtaining the adequate number of features during the univariate selection process, a recursive feature elimination (RFE) was operated with the number of features passed as parameter to identify the features selected. During the RFE process, first, the classifier is trained on the original set of features and weights are attributed to each features. Then, features whose absolute weights are the smallest are pruned from the current set features. That process is recursively repeated on the pruned set until the desired number of features to choose is finally reached

B. Discussions and Results

Feature selection is utilized to discriminate the redundant and irrelevant data. It is a technique of selecting a subset of relevant attributes that completely represent the given problem alongside a minimum deterioration of presentation. As consequence, working with a small number of feature may bring better results.

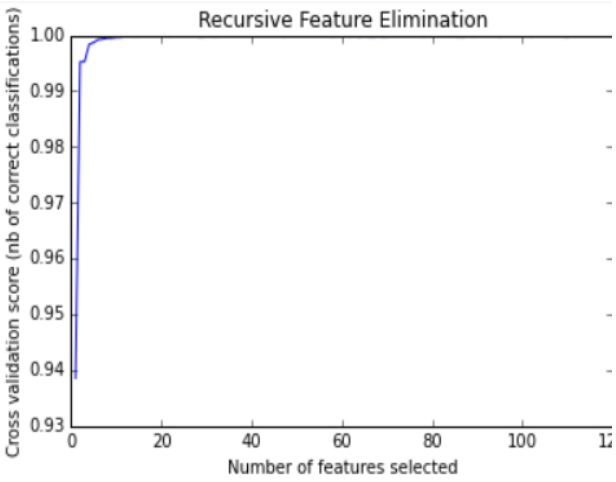
A general technique to achieve feature selection will be to retrieve the smallest set of features that can exactly characterize the training data. If an attribute always coincides with the label class (that is, it is an exact predictor), it is enough to characterize the data. On the other hand, if an attribute always has the same value, its prediction power will decrease and can be very low. A recursive feature elimination method which repeatedly build a model, placing the feature aside and then repeating the process with the remained features until all features in the dataset are exhausted. The objective of the recursive feature is to retrieve features by recursively keeping smaller and smaller group of features.

A good feature ranking criterion does not necessarily produce a good feature subset generation. The some criteria estimate the effect of removing one feature at a time based on the goal to achieve. They become very sub-optimal when it comes to removing several features at a time, which is necessary to obtain a small feature subset. This problem can be overcome by using the following iterative procedure that is Recursive Feature Elimination:

- Train the classifier (optimize the weights of features with respect to criterion).
- Compute the ranking criterion for all features.
- Remove the feature with smallest ranking criterion.

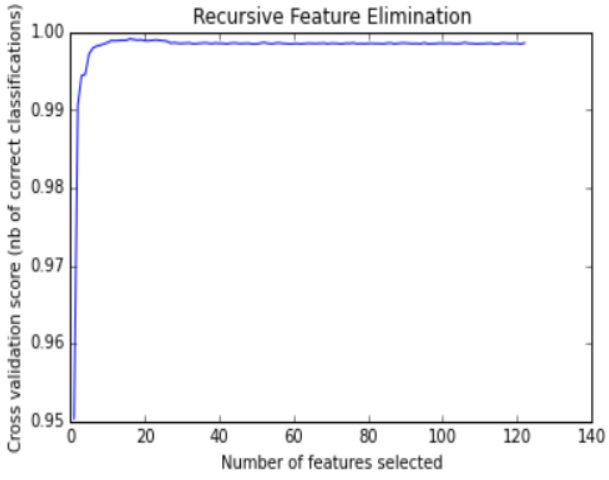
Firstly, the classifier is trained on the initial group of attributes and weights are assigned on each attributes. Then, absolute weights of some attributes that are the smallest are pruned from the current sets of attributes. That technique is recursively repeated on the pruned set until the desired number of attributes to select is eventually reached. As such, it is a good optimization for finding the best performing subset of features. In should be noted that RFE has no effect on correlation methods since the ranking criterion is computed with information about a single feature.

An analysis was performed to determine the accuracy of our estimator after selecting relevant features as illustrate in the Fig. 3, 4, 5 and 6 and the detail is summarized in the Table 4. When comparing the result alongside the performance evaluation with all features describe in the Table 5, a significant improvement of the overall performance of the proposed model has been observed.



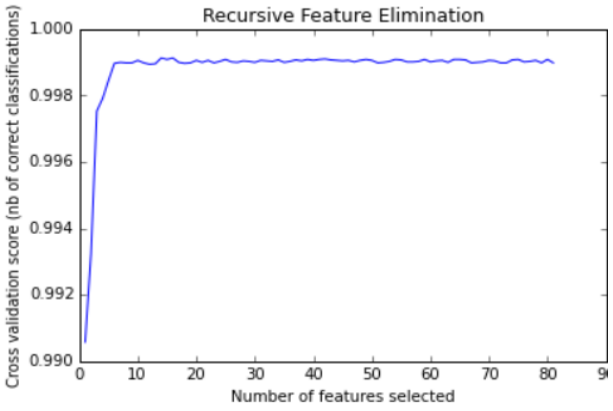

```
0.999095272483
```

Fig. 3. DOS Recursive Feature Elimination



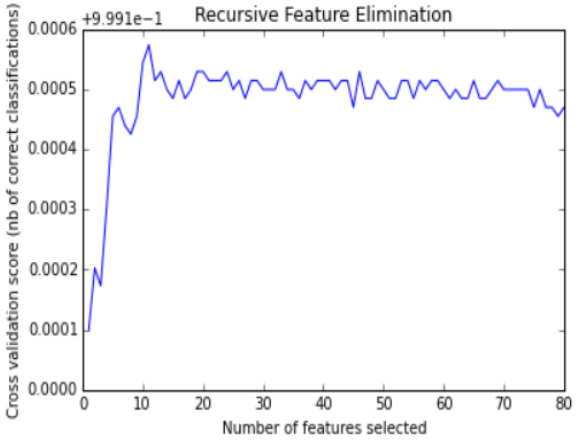

```
0.998014087806
```

Fig. 4. Probe Recursive Feature Elimination




```
0.99881074737
```

Fig. 5. R2L Recursive Feature Elimination




```
0.999583790242
```

Fig. 6. U2R Recursive Feature Elimination

TABLE IV. PERFORMANCE EVALUATION WITH SELECTED FEATURES

Accuracy	Precision	Recall	F-measure	N of Features	Class
99.90	99.69	99.79	99.74	12	Dos
99.80	99.37	99.37	99.37	15	Probe
99.88	97.40	97.41	97.40	13	R2L
99.95	99.70	99.69	99.70	11	U2R

TABLE V. PERFORMANCE EVALUATION WITH 41 FEATURES

Accuracy	Precision	Recall	F-measure	N of Features	Class
99.66	99.505	99.71	99.61	41	Dos
99.57	99.04	98.84	98.94	41	Probe
97.03	95.83	95.59	95.71	41	R2L
99.64	99.66	99.61	99.65	41	U2R

Table 6 shows a 2x2 confusion matrix after features selection on the dataset for a combination of two target classes (normal class and an attack class).

TABLE VI. DETAILS CONFUSION MATRIX AFTER FEATURES SELECTION

Confusion Matrix		Predicted Label: Normal	Predicted Label: DoS		
True Label	Normal	9676	25	Positive predictive value	99.74 %
	DoS	15	7445	Negative predictive value	99.79 %
Confusion Matrix		Predicted Label: Normal	Predicted Label: Probe		
True Label	Normal	9652	59	Positive predictive value	99.39 %
	Probe	30	2391	Negative predictive value	98.76 %
Confusion Matrix		Predicted Label: Normal	Predicted Label: R2L		
True Label	Normal	9594	117	Positive predictive value	98.79 %
	R2L	87	2798	Negative predictive value	96.98 %
Confusion Matrix		Predicted Label: Normal	Predicted Label: U2R		
True Label	Normal	9683	28	Positive predictive value	99.71 %
	U2R	7	60	Negative predictive value	89.53 %

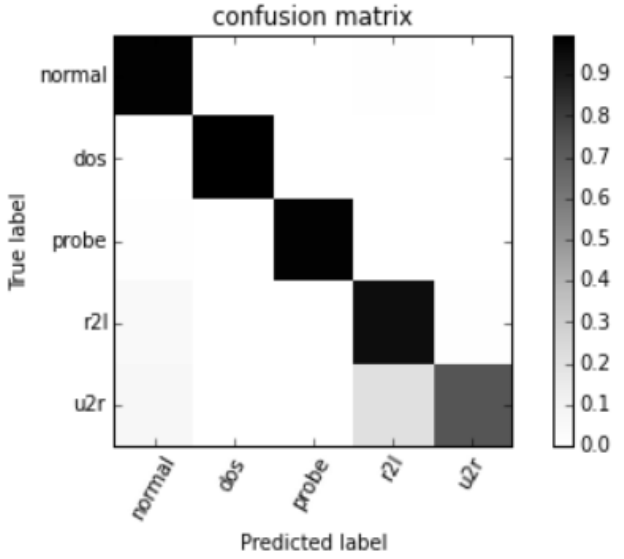


Fig. 7. Confusion Matrix

TABLE VII. CONFUSION MATRIX DETAILS

99.84	2.368e-03	1.544e-03	1.153e-02	4.119e-04
3.485e-03	99.691	9.38337802e-04	6.702e-04	0
8.674e-03	2.891e-03	99.45	2.891e-03	4.130e-04
4.540e-02	1.039e-03	2.079e-03	96.10	6.239e-03
5.970e-02	0	0	2.089e-01	87.5

The completed confusion matrix as illustrate in the Figure 7 and Table 7 show the number of correct and incorrect predictions made by the classification model compared to the actual outcomes in the dataset.

The Table 8 present the relevant features after a recursive features elimination was operated on the NSL KDD dataset. Features are retrieved based on their rank, relevant (i.e., estimated best) features are attributed a rank 1. During the selection process, some features have participated in improving the accuracy of the model, which it is called the important features selected. The important features or attributes are calculated as the reduction of the criterion brought by that attributes. Each time a split of a node is performed on a particular attributes, the criterion for the two children nodes is inferior to their parent.

TABLE VIII. RELEVANT FEATURES

Target	Features selected
Dos	diff_srv_rate,dst_bytes,dst_host_serror_rate,dst_host_srv_serror_rate ,flag_S0, rerror_rate,same_srv_rate,service_ecr_i,service_http,service_private,src_bytes, wrong_fragment'
Probe	src_bytes, service_http, dst_bytes, service_ftp_data, dst_host_rerror_rate, service_smtp,service_finger, service_private , rerror_rate , dst_host_diff_srv_rate , dst_host_same_srv_rate , service_telnet , dst_host_count , service_auth , count
R2l	dst_bytes , dst_host_same_src_port_rate , dst_host_same_srv_rate , dst_host_srv_count , dst_host_srv_diff_host_rate , duration , hot , num_access_files , num_fail_login , num_root , service_ftp_data , service_r2l4 , src_bytes
U2r	src_bytes , service_other , service_ftp_data , root_shell , num_shells ,num_file_creations , hot , dst_host_same_srv_rate , dst_host_count , dst_bytes , count

The Figure 8, 9, 10, 11 illustrate the subset features after the feature elimination process has been performed, the aim of this process is to elimination non-relevant features and only print out the relevant or important one.

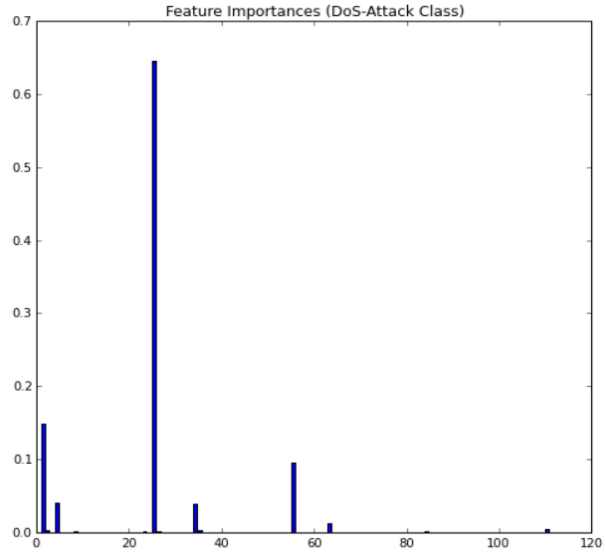


Fig. 8. DOS Class Features Selected and Important

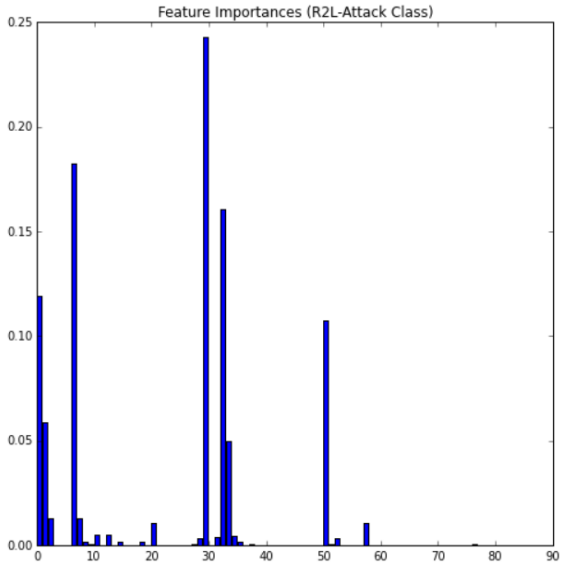


Fig. 10. R2L Class Features Selected and Important

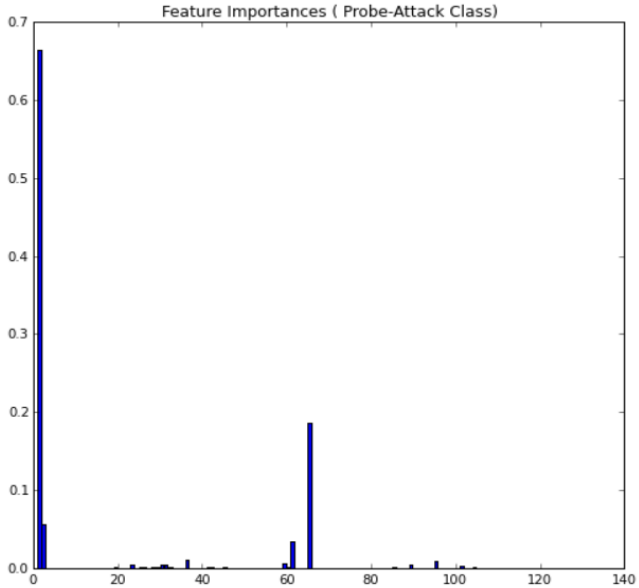


Fig. 9. Probe Class Features Selected and Important

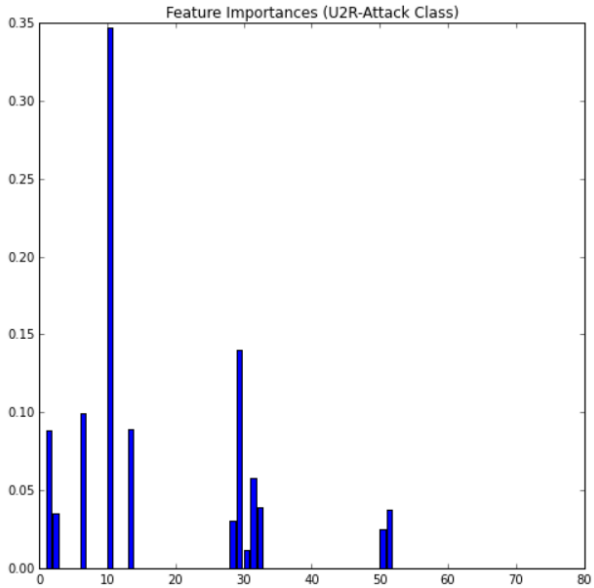


Fig. 11. U2R Class Features Selected and Important

The analysis for feature selection has been done in terms of the class that achieved good levels of entropy or Gini index from others in the training set and the analysis of feature relevancy in the training set. The importance of a feature is computed as the (normalized) total reduction of the criterion brought by that feature. From the Figure 8, 9, 10, 11 and the Table 7 shows that the most relevant features for Dos, Probes, R2L and U2R are respectively "same_srv_rate", "src_bytes", "dst_host_srv_count" and "root_shell"

With the improvement the accuracy, the proposed model demonstrated that it performs well after selecting relevant features. And from the Figure 12, it is evident that the time taken to build classifier is decreased through feature selection, especially the proposed approach. Building Decision Tree classifier on the dataset with features selected by our approach takes only 0.956 seconds for DoS attack class, which is faster than building on the dataset with all of the 41 features by 14.541s as shown on Table 9.

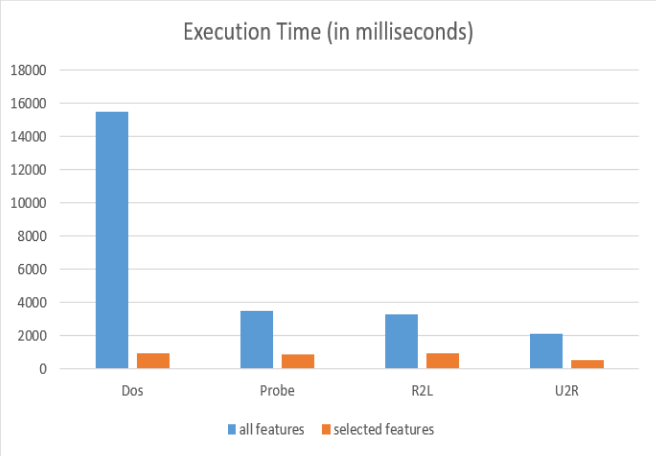


Fig. 12. Execution time of the classifier

TABLE IX. TIME TAKEN TO BUILD CLASSIFIER ON EACH FEATURE SUBSETS

Attack Classes	With all Features	With selected Features
Dos	15.5 seconds	0.959 seconds
Probe	3.48 seconds	0.868 seconds
R2L	3.31seconds	0.929 seconds
U2R	2.10 seconds	0.530 seconds

This result provided new insight using a classification learning algorithm and reduction technique to selection relevant and important feature in order to improve the accuracy detection rate of the system and to identify possible features which may contribute to this improvement.

As our goal was to determine whether or not a features selection process will improve the accuracy detection of a model on different set of attacks class found in the dataset used for our experiment, each set of attacks class were treated individually as they present different characteristics and are different by nature.

This decision was also made in order to identify all relevant features for each different attacks class and to compare the accuracy improvement from the original set of features. As relevant features which are appropriated to classify those different attacks class have been found, the result analysis has shown that the performance of the model has really been improved.

As an example the R2L attack accuracy detection has been improvement from 97.03% to 99.88% as well as its execution time as shown on the Table 9. The Table 10 shows a comparison between our method and some previous one.

TABLE X. COMPARISON WITH OTHER FEATURES SELECTION TECHNIQUES

Author	Method used	Classifier used	Accuracy for Attack Classes (N^{br} of selected features)			
			DOS	Probe	R2L	U2R
(Dhanabal & Shantharajah 2015) [27]	Correlation based Feature Selection method	J48	99.1 % (6)	98.9 % (6)	97.9 % (6)	98.7 %(6)
(Senthilnayaki et al. 2015) [28]	Optimal Genetic Algorithm	SVM	99.15 % (10)	99.08 % (10)	96.50 % (10)	97.03 % (10)
(Zhang & Wang 2013) [29]	Sequential search	Naïve Bayes	99.3 % (11)	97.4 % (11)	95.0 % (11)	59.6 % (11)
(Alazab et al. 2012) [30]	Information gain	J48	99.7 % (12)	97.8 % (12)	91.3 % (12)	97.2 % (12)
(Mukherjee & Sharma 2012) [31]	Feature vitality based Method	Naïve Bayes	98.7% (24)	98.8 % (24)	96.1 % (24)	64% (24)
(Parsazad et al. 2012) [32]	Correlation Coefficient	K-nearest neighbor	98.34 % (30)	98.38 % (30)	97.03 % (30)	83.33 % (30)
(Parsazad et al. 2012) [32]	Fast feature Reduction	K-nearest neighbor	98.28 % (10)	98.50 % (10)	97.79 % (20)	82.00 % (10)
(Parsazad et al. 2012) [32]	Least Square Regression Error	K-nearest neighbor	98.34 % (30)	98.98 % (20)	97.62 % (20)	82.61 % (20)
(Parsazad et al. 2012) [32]	Maximal Information Compression Index	K-nearest neighbor	98.03 % (30)	98.92 % (10)	98.05 % (20)	90.70 % (20)
The proposed method	Recursive Feature Elimination	Decision Tree Classifier	**99.90 % (12)**	**99.80 % (15)**	**99.88 % (13)**	**99.95 % (11)**

VI. CONCLUSION

In this paper, the significance of using a set of relevant features with an adequate classification learning algorithm for modelling an IDS has been demonstrated.

A presentation and proposition of a feature selection method which consist of a univariate features selection associated with a recursive feature elimination using a decision tree classifier to identify important features have been done. This process repeatedly builds a model placing the feature aside and then repeating the process with the remaining features until all features present in the dataset are exhausted. The evaluation the effectiveness of the method using different classification metric measurement has been made and it has been proved that by reducing the number of feature, the accuracy of the model was improved. The feature selection method proposed in this paper had achieved a high result in term of accuracy and features were identified based on information gain and ranking technique.

REFERENCES

[1] A. Alazab, M. Hobbs, J. Abawajy, and M. Alazab, "Using feature selection for intrusion detection system," 2012 Int. Symp. Commun. Inf. Technol., pp. 296–301, 2012.

[2] M. P. K. Shelke, M. S. Sontakke, and A. D. Gawande, "Intrusion Detection System for Cloud Computing," Int. J. Sci. Technol. Res., vol. 1, no. 4, pp. 67–71, 2012.

[3] S. Suthaharan and T. Panchagnula, "Relevance feature selection with data cleaning for intrusion detection system," 2012 Proc. IEEE Southeastcon, pp. 1–6, 2012.

[4] S. Suthaharan and K. Vinnakota, "An approach for automatic selection of relevance features in intrusion detection systems," in Proc. of the 2011 International Conference on Security and Management (SAM 11), pp. 215-219, July 18-21, 2011, Las Vegas, Nevada, USA.

[5] L. Han, "Using a Dynamic K-means Algorithm to Detect Anomaly Activities," 2011, pp. 1049-1052.

[6] R. Kohavi, et al., "KDD-Cup 2000 organizers report: peeling the onion," ACM SIGKDD Explorations Newsletter, vol. 2, pp. 86-93, 2000.

[7] I. Levin, "KDD-99 Classifier Learning Contest: LLSoft s Results Overview," SIGKDD explorations, vol. 1, pp. 67-75, 2000.

[8] J. McHugh, "Testing intrusion detection systems: a critique of the 1998 and 1999 darpa intrusion detection system evaluations as performed by lincoln laboratory," ACM Transactions on Information and System Security, vol. 3, no. 4, pp. 262–294, 2000.

[9] M. Tavallaee, E. Bagheri, W. Lu, and A. Ghorbani, "A Detailed Analysis of the KDD CUP 99 Data Set," Submitted to Second IEEE Symposium on Computational Intelligence for Security and Defense Applications (CISDA), 2009.

[10] KDD 99 dataset, Accessed December 2015, http://kdd.ics.uci.edu/databases/kddcup99

[11] NSL KDD dataset, Accessed December 2015, https://github.com/defcom17/NSL_KDD

[12] P. Ghosh, C. Debnath, and D. Metia, "An Efficient Hybrid Multilevel Intrusion Detection System in Cloud Environment," IOSR J. Comput. Eng., vol. 16, no. 4, pp. 16–26, 2014.

[13] Dhanabal, L., Dr. S.P. Shantharajah, "A Study on NSL_KDD Daaset for Intrusion Detection System Based on Classification Algorithms," International Journal of Advanced Research in Computer and Communication Engineering, vol. 4, issue 6, pp. 446-452, June 2015

[14] C. F. Tsai, et al., "Intrusion detection by machine learning: A review," Expert Systems with Applications, vol. 36, pp. 11994-12000, 2009.

[15] V. Bolón-Canedo, et al., "Feature selection and classification in multiple class datasets: An application to KDD Cup 99 dataset," Expert Systems with Applications, vol. 38, pp. 5947-5957, 2011.

[16] F. Amiri, et al., "Improved feature selection for intrusion detection system," Journal of Network and Computer Applications, 2011.

[17] Juan Wang, Qiren Yang, Dasen Ren, "An intrusion detection algorithm based on decision tree technology," In the Proc. of IEEE Asia-Pacific Conference on Information Processing, 2009.

[18] Dewan Md. Farid, Nouria Harbi, and Mohammad Zahidur Rahman, "Combining Nave Bayes and Decision Tree for Adaptive Intrusion Detection," International Journal of Network Security & Its Applications, Vol. 2, No. 2, April 2010, pp. 12-25.

[19] Ektefa M, Memar S, Sidi F, Affendey L., "Intrusion detection using data mining techniques," 2010 International Conference on Information Retrieval & Knowledge Management(CAMP). 2010.doi:10.1109/infrkm.2010.5466919.

[20] Geetha Ramani R, S.SivaSathya, SivaselviK, "Discriminant Analysisbased Feature Selection in KDD Intrusion Dataset," , International Journal of Computer Application VoI.31,No.ll, 2011

[21] S. Mukherjee and N. Sharma, "Intrusion Detection using Naive Bayes Classifier with Feature Reduction," Procedia Technol., vol. 4, pp. 119–128, 2012.

[22] Bhavsar Y. B, Waghmare K. C. "Intrusion Detection System Using Data Mining Technique: Support Vector Machine," International Journal of Emerging Technology and Advanced Engineering, Vol.3, Issue 3, pp.581-586(2013).

[23] O. Y. Al-Jarrah, a. Siddiqui, M. Elsalamouny, P. D. Yoo, S. Muhaidat, and K. Kim, "Machine-Learning-Based Feature Selection Techniques for Large-Scale Network Intrusion Detection," 2014 IEEE 34th Int. Conf. Distrib. Comput. Syst. Work., pp. 177–181, 2014.

[24] N. G. Relan and D. R. Patil, "Implementation of Network Intrusion Detection System using Variant of Decision Tree Algorithm," 2015 Int. Conf. Nascent Technol. Eng. F., pp. 3–7, 2015.

[25] Scikit-Learn, Accessed December 2015, http://scikit-learn.org/stable/index.html

[26] V. Bol´oN-Canedo, N. S´aNchez-Maro˜no, and A. Alonso-Betanzos, "Feature selection and classification in multiple class datasets: an application to kdd cup 99 dataset," Expert System Application, vol. 38, pp. 5947–5957, 2011.

[27] Dhanabal, L., Dr. S.P. Shantharajah, "A Study on NSL_KDD Daaset for Intrusion Detection System Based on Classification Algorithms," International Journal of Advanced Research in Computer and Communication Engineering, vol. 4, issue 6, pp. 446-452, June 2015.

[28] Senthilnayaki, B., Venkatalakshmi, D.K. & Kannan, D.A., " Intrusion Detection Using Optimal Genetic Feature Selection and SVM based Classifier," 2015 3rd International Conference on Signal Processing, Communication and Networking (ICSCN) Intrusion, pp.1–4.

[29] Zhang, F. & Wang, D., "An Effective Feature Selection Approach for Network Intrusion Detection," 2013 Eighth International Conference on Networking, Architecture and Storage, IEEE.

[30] Alazab, A. et al.," Using feature selection for intrusion detection system," 2012 International Symposium on Communications and Information Technologies (ISCIT), pp.296–301. Available at: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6380910.

[31] Mukherjee, S. & Sharma, N., " Intrusion Detection using Naive Bayes Classifier with Feature Reduction," 2012 Procedia Technology, 4, pp.119–128. Available at: http://dx.doi.org/10.1016/j.protcy.2012.05.017.

[32] Parsazad, S., Saboori, E. & Allahyar, " Fast Feature Reduction in intrusion detection datasets, " MIPRO 2012 Proceedings of the 35th International Convention, pp.1023–1029.

Systematic Evaluation of Social Recommendation Systems: Challenges and Future

Priyanka Rastogi
Research Scholar, Department of CSE & IT
The NorthCap University
Gurgaon, India

Dr. Vijendra Singh
Associate Professor, Department of CSE & IT
The NorthCap University
Gurgaon, India

***Abstract*—The issue of information overload could be effectively managed with the help of intelligent system which is capable of proactively supervising the users in accessing relevant or useful information in a tailored way, by pruning the large space of possible options. But the key challenge lies in what all information can be collected and assimilated to make effective recommendations. This paper discusses reasons for evolution of recommender systems leading to transition from traditional to social information based recommendations. Social Recommender System (SRS) exploits social contextual information in the form of social links of users, social tags, user-generated data that contain huge supplemental information about items or services that are expected to be of interest of user or about features of items. Therefore, having tremendous potential for improving recommendation quality. Systematic literature review has been done for SRS by categorizing various kinds of social-contextual information into explicit and implicit user-item information. This paper also analyses key aspects of any generic recommender system namely Domain, Personalization Levels, Privacy and Trustworthiness, Recommender algorithms to give a better understanding to researchers new in this field.**

Keywords—Social Recommender System; Social Tagging; Social Contextual Information

I. INTRODUCTION

Exponential growth and sophistication of information on the web is the result of diminishing lines of produces and consumers of data as well as latest growing trend of pervasive computing of "information anywhere, anytime" .In order to deal with information overload, progressive evolution of Recommender systems has taken place over the years. There are zillions of different items available, users cannot be expected to browse through all of them to find what they might like, and therefore, filtering has become a popular technique to connect supply and demand [1].

In mid-90 a lot of research was done to improve Collaborative Filtering (CF) [2], [3] [4], [5] , [6] one of the most popular methods of recommendation, and even now. One of the major problem with CF is Cold start problem which occurs due to initial lack of ratings to make any reliable recommendation. To overcome this, new methods of recommendations were explored like demographic filtering, content-based filtering (CBF) [3], [6]. At this point in time, Recommender systems used only the explicit ratings from the users along with demographic information (e.g., Sex, age, country) and limited content based information or item attributes (e.g. genre, album, singer etc. for music) available with recommender engine designers. In some domains (e.g., videos, photos, blogs) it is very difficult to generate reliable attributes for items. Therefore, pure Content based Filtering (CBF) implementations are rare to find [3] since they are based on content analysis of items. Also, one of the major drawback of CBF is overspecialization problem. Because of its inherent nature, it tries to recommend similar type of items to users, thereby losing on novelty factor in making recommendation. In order to overcome the short coming of the existing methods, Hybrid methods [7] came to existence which exploited the merits of each of these techniques. Constant effort of improving hybrid methods still went on. But, the data sparsity problem inherent in the traditional recommendation systems adversely affects the recommendation quality. Also, many of the traditional recommendation algorithms could not be applied on large datasets [3], [8].

Basic premise of traditional recommender systems is that, it considers users to be independent people, ignoring the social trust relationships among the users, which happens to be quite an important key aspect and distributed across identically. With the help of social contextual information (e.g., user's social trust network, tags issued by users or associated with items, etc.) more accurate suggestions could be made. This led to the second phase of the evolution in recommendation system. With the rapid expansion of Web 2.0, these systems incorporated social information [8] [9], [10], [11], [12] along with information used in traditional recommendation system, leading to the development of Social based Recommendation systems [13], [14], [15], [2] [16], [17] [18]. This social information was related to the virtual social circle of the user. Simultaneously, users- generated information (e.g., comments, post, tags, photos, videos) in social network too started being used for the purpose of recommendation [11]. Bobadilla et. al.(2013) in [3] have shown the evolution of recommendation system from first phase which is based on traditional Web to the present second phase based on Social Web, which has almost progressed to third phase based on Internet of things. From the evolution so far, it was evident that as we assimilate and integrate more and diverse types of information, the gradual development in these systems is bound to happen.

As seen so far, in this paper we have provided an overview of how recommender systems evolved over the years and highlights the reasons that are leading to its evolution. The rest of the paper is structured in the following manner: Section 2 focuses on analyzing recommendation systems over 8 key

dimensions for better understanding. Section 3 focuses on Social Recommendation systems (SRS) which is based on social information (e.g., tags, post, opinions, and social links of user) going as input to recommendation engine. A systematic literature review has been done for the same. Section 4 provides an overview of the next generation recommendation systems and highlights the challenges posed by existing systems. Lastly, Section 5 concludes the paper stating the performance of social –contextual recommendation systems over traditional recommendation systems based on the review done. This paper would give researchers a deep insight into SRS and acquit them about the latest advancements and finally provide a foundation on which the future work of these systems could be based.

II. Dimensions of Recommender System

In order to design a new recommendation system or improve an existing or simply understand it, one needs to understand the generic framework of any recommender system. The key elements in any generic recommendation systems (User, Items, Ratings, Community) are linked as depicted in Figure 1. Users make preferences for items in a system. They express their opinion in the system via ratings (e.g., on a scale of 0-5, ratings in form of stars, fun boards). The space where these key elements make sense is called community.

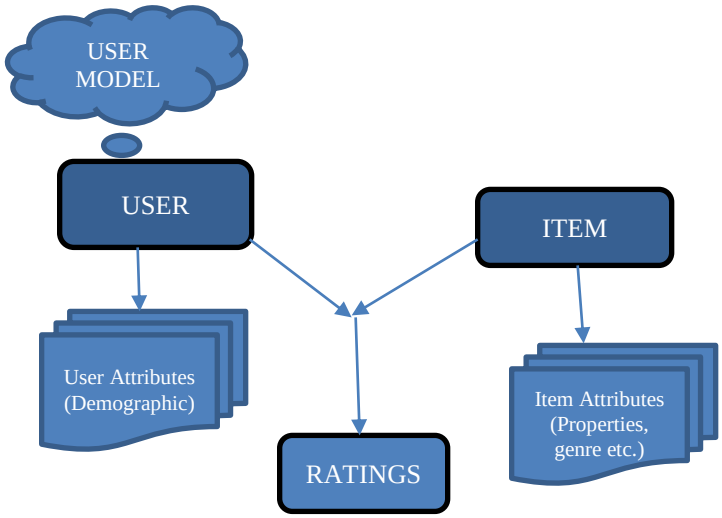


Fig. 1. Linking of key elements of Recommender System

Konstan [2] discussed 8 dimensions of analysis for Recommendation system. They are various aspects to these systems, which makes the understanding and functioning of it easier to researcher. Further these dimensions have been discussed in present scenario to explore future opportunities as the commercial recommendation systems strive in the market to offer best content and quality in recommendations as well as greatest variety of services [3].

- **Domain** –Recommendation systems has felt its importance in diverse areas and with the popularity of internet, the number is still growing [3]. Based on the research carried out in paper [19] in 2012, most of the articles were related to Movie recommendations (46 out of 164 articles, 28.04%) owing to easy availability of the movies dataset MovieLens. The second most sought after domain is E-commerce (33 out of 164 articles, 20.9%). However, a large volume of recommendation systems literature is focused on varied topics such as Entertainment and Beyond e.g., Books, Music , Mobile App downloads; Match Making; Social Media e.g., Suggesting Friends, Face Recognition for picture tags; Tourism e.g. tripadvisor.com; e-news; digital library.
- **Purpose** –The compelling reason for implementing recommendations in E-commerce domain is that they have turned out to be serious business tools to enhance the sales by improving cross-sell by suggesting additional products and gaining customer loyalty resulting in repeat business [20]. In university digital library, recommender system is proposed to disseminate information based on quality to help users access relevant research resources among the thousands of resources that are available but yet hard to find [21], [22].
- **Recommendation Context** –It refers to the context in which the recommendation is being made. It answers the question - What the user is doing when the recommendation is being made. Examples could include like e.g. hanging out with friends, looking for an eating joint in a user's nearby location. Recommendation systems that consider group of users as input to these system, are starting to expand and are used in different areas like tourism, music, web etc. Currently, mobile applications use GPS feature to fetch the current geographic location of user, and employ Recommender systems [23], [24] to utilize this information for generating recommendations e.g., Zomato app. Moon-Hee Park , Jin-Hyuk Hong , Sung-Bae Cho (2007) proposed to model user preference in restaurants by using context-aware information and user profile by implementing map-based Personalized Recommendations using Bayesian Network [25].
- **Who's Opinion** – It refers to people on whose opinions, recommendations are made e.g., Experts, Friends, Friends of Friend, PHOAKS. SRS uses User's trust network which is the social network of user (friends, friends of friends etc.) to make recommendations.
- **Personalization Levels** - Recommendations have many variants. They could be in the form of Non-personalized summary stats (e.g., Best Seller books, popular movies), Demographic personalization based on target group (e.g., Male/female, different age groups), Ephemeral personalization based on current navigation (e.g., item generally brought with another item – Product associated recommendation), Persistent personalization based on preferences and behavior (e.g., based on combination of user's historical purchases, rating given by him for products and his browsing history).
- **Privacy and Trustworthiness** –Privacy is an important issue because these systems exploit information from social networking sites which contain a lot of

information about its registered users. How much of the user's personal information to be revealed? For the sake of privacy preservation, a certain level of ambiguity must be introduced into the predictions. A tradeoff must be maintained between the accuracy and predictions [3].

Recommendation systems are highly vulnerable to external manipulations especially in E-commerce where rating biasness can be introduced by companies who wish to recommend their products more than their competitors (Shilling attacks).

- **Interfaces** – The output of recommendation algorithm could be in the form of e.g., predictions, recommendations, and filtering of information. While the input for these algorithm could be broadly categorized into User data and Item data. Initially, these algorithms made use of explicit information (e.g., user rating for various items) to filter out items that could be recommended to other users of similar interests. But this was not sufficient to make reliable recommendations due to initial lack of ratings for new item, new user or new community. This is known as cold start problem. Then, they incorporated implicit information typically by monitoring user's behavior (e.g., songs heard, books read, applications downloaded). And now input from diverse areas is being used to make accurate recommendations. Fig 2 shows a snapshot of different input data which has been sub categorized into explicit - implicit data and user - item data. Indicated in the fig, aggregated explicit – implicit user and item data is used in traditional recommendation systems. And input data used in SRS is superset of data used in traditional recommendation systems including some additional data.

- **Recommendation Algorithms** - In general, recommendation algorithms are based on 2 basic filtering techniques: Collaborative Filtering, Content-Based Filtering. These two approaches can be combined in different ways forming Hybrid technique [3].These filtering techniques (Collaborative, Content-Based, Hybrid) can be applied on databases (Nonpublic Commercial databases or Public databases) to yield accurate predictions and recommendations of items to the taste of users.

Among these techniques, Collaborative Filtering (CF) is has been the most popular in recommendation algorithm. It is based on the assumption that an active user preferences would be in accordance with other similar user preferences. It allows users to give ratings about a set of items, generating spare matrix of user-item. Based on the matrix, first the similarity between users can be retrieved (e.g., using k-nearest neighbor). Second, predict rating for an item for an active user who has not rated this item earlier and leverage user neighbor's ratings for the item (Fill in missing values).Third, select promising items for recommendation based on user's similarity with other users. This is generic CF procedure.

Collaborative filtering techniques could be implemented in 2 ways: User – based Collaborative filtering where in neighborhood of similar-taste people is selected and their opinions are used for making predictions. Another is, Item – based Collaborating filtering where similarity among various items via ratings is pre- computed and user's own ratings are used to triangulate for recommendation. In other words, item based CF is usage of user – item matrix represented by its column vectors.

Content based filtering(CBF) makes recommendations based on user choices made in the past (e.g., if a user rated a rom-com movie positively over a movie recommendation site, the system would probably recommend more of recent rom-com movies that he has not yet seen or rated). A user model is created using the user ratings (for the watched movies) and item attributes (in case of movies, attributes like cast, types of movie, genre etc.). This model is applied to predict which kind of movie would the user like in future. It is also known as Information filtering.

There is another variant to CBF which is knowledge based filtering where in, an item attributes form model in item space and users navigate that space. As in the case of personalized news feeds, user reads certain news articles, recommender systems read user's preferences and based on the item model, recommends similar news articles to the user.

Hybrid technique uses a different combinations of CF and CBF [3] to exploit the merits of each of these techniques. Hybrid techniques are usually based on probabilistic methods like Genetic algorithms, Bayesian networks, Clustering etc.

Fig. 2. Categorization of Input Data

Recommender Systems can also be divided into Memory based (Similarity Measures, Aggregation approaches) or Model based approaches (Clustering methods, Genetic algorithms, Bayesian classifiers, Neural Network, Fuzzy systems, latent features) in widely accepted taxonomy. Memory-based approach can be applied only on user-item matrix while in the case of Model-based approach, data is used to model the system.

III. Social Recommender System

Ever since the rapidly increasing popularity of Web 2.0 applications and advent of Social Web, exploiting social contextual information (e.g., social links of users in the form of friend list, followed and followers, user's interest groups etc.) contain huge supplemental information about items or services that are likely to be of interest of user or about features of items. Therefore, providing tremendous opportunity to improve the recommendation quality.

There have been constant efforts for exploring social contextual information (e.g., user's social trust network, tags issued by users or associated with items, etc.) and devising methods to capture that information and incorporate it into recommender systems. It works on the principle that a user would trust their network of "elective affinities" more than generic suggestions made by impersonal entities unknown to them [13]. In simple words, when asking your trusted friend about a book he would recommend you to read or movie he would recommend you to watch, you would rely on the recommendations given by your trusted friend rather suggestions given by some acquaintance you don't know or trust. This is a kind of verbal social recommendation indulgence. On similar lines, in users' social trust network, users more likely to go by the interests of the friends/people they trust.

Knowledge and content sharing systems (e.g., news, articles, bookmarks) too have been gaining momentum and generating huge amounts of shared data along with user created data in the form of comments, blogs, ratings, labels etc. Discovering relevant content in such shared data space has become a night mare. It's like finding a needle in haystack. In such systems, like E-news website, where the users can read news articles from around the globe. There should be some practical means to assess the quality and authenticity of the news going into the personalized news feeds of the users. Also, some parameters to check the trustworthiness of the sites publishing news articles and accessing reputation of sites before making recommendations.

So, we see that there is a lot of importance of trust and reputation in social web. Network of trust is a social network where nodes are inter-connected based on their trust relations [26]. Many researchers have devised various approaches to measure trust. User trust and Item trust both can be measured either implicitly or explicitly.

User Trust can be computed through explicit information (e.g., trust networks [14], [26] distrust analysis [27], personality based similarity measure [4] etc.) or through implicit information obtained in the form of social network (e.g., trust propagation mechanism [5]).Item Trust can be explicitly obtained by assessing the reputation of items through feedback of users in online community or implicitly obtained by studying the relationship between the user and the items [3].

A. Types of Information Sources in Social Context

As, the traditional recommender algorithms exploit explicit user feedback as an information source, on which recommendation to similar users or items could be based. Similarly, various explicit and implicit information source that aid in capturing social information for User and Item are depicted in the Table 1.

TABLE I. DIFFERENT SOCIAL INFORMATION SOURCES

	User Data	**Item Data**
Explicit data	Friends, followed, followers, trusted and untrusted	Social Tags (folksonomies)
Implicit data	Implicit trust	Number of tweets, Number of posts, Number of descriptors

Researchers have tried to propose different ways in which the social information of user could be captured and used for recommendation process. In social media communities, explicit social networks are created by complex web of relationships amongst users making friendship with other users and/or by following/being followed by and/or joined by some common interest group. They are useful for forecasting users' inclinations, because the users' interests may be governed by their friends or neighbors in interest groups. A lot of work has already been undertaken for utilizing friendship relations for recommendation [9]. The social filtering of links in social network to discover user's trust network constitutes the inherent implicit data of user.

B. Social Links

Wolfgang Woerndl and Georg Groh (2007) added social context of user as another dimension to the item-user matrix of CF, thereby broadening of domain of mapping the Rating (R) to 3 - dimensional space represented by U,I,C (U: User, I :item, C:Context). They used real data set where in subset of users from Lokalisten4 - a Munich-based German community for making friends, rated some restaurants via online survey. Their evaluation showed that the proposed social neighborhood based recommender outperformed old-fashioned collaborative filtering algorithms (using kNN method) in this scenario. Its limitation is, it remains doubtful whether these results can be generalized in all domains [28].

Fengkun Liu, Hong Joo Lee (2010) used social network information and CF methods for recommending suggested neighbor groups. The methodology followed involved collecting data about users' preference ratings for homepage skin (digital item) and their social relationships from a social networking Web site -Cyworld5, a social networking community in South Korea. Next, they developed approaches for selecting neighbors using Pearson's correlations and augmented it with friends' data. As a result, the model generated recommendations about items using proposed CF with suggested neighbor sets [29].

Kazienko et. al. (2011) in their paper analyzed multimedia sharing systems (MSS), 'Flickr6' photo sharing system as multi relation social network (MSN) wherein they aimed at exploring the various relation layers based on contact list, tags, group, favorites, opinions. Eventually, aggregating these layers to form a comprehensive multidimensional social relationship between users. This enabled the successful merging of both the semantic and social background from which the concerned user hailed. The model was used to recommend other users' to the active user in MSS. Additionally some system and personal weights were adjusted for better accuracy. The experiment was conducted in two stages which lead to the generation of two separate recommendations. The initial suggestion being computation with an assumption that applied equal values of personal weight for layers, i.e., for each layer k and each user ui: By using adaptation mechanism the suggestions were provided which were adjusted according to each user and they were expected to rate it., and this is how it lead to the generation of the second recommendation list. Thereafter, layer contributions were applied after the first lists were rated. After adaptation personal weight values were analyzed directing towards the revelation that the social layer based on indirect reciprocal contact list R_{coc} and author-opinion R_{ao} gained in their contribution much after adaptation, by 220% and 65%, respectively, where other layers lost in their importance. The tag-based layer R_t increased in average by 8% [16].

Xin Liu and Karl Aberer (2013) proposed SoCo (social network aided context-aware recommender system). First they partitioned the original user-item rating matrix into groups based on similar contexts of ratings by using random decision trees. Next, they predicted missing preference of a user for an item in the portioned matrix by using Matrix factorization. A social regularization term was added to the matrix factorization objective function which inferred user's

preference for an item by learning interests from his/her friends who are expected to share similar tastes. The model was experimented on Real dataset, Douban7, largest Chinese social platforms for sharing reviews and recommendations for books, movies and music. It contains time/date related information, other inferred contextual information and social relationships information. SoCo outperformed compared to the contemporary context-aware recommender system and social recommendation model by 15.7% and 12.2% respectively [30].

C. Social Tagging

With the popularity of Web 2.0, there has been a progressive growth in creation, modification and sharing of online content over social network communities like Youtube, Facebook, Flickr etc. and social tagging systems (STS) provides powerful way for users to organize, administer, consolidate and search for innumerable kinds of resources. These tags [8], [9] , [12] , [15] , [17], [18] carry interesting information about the preference of users who make the tags and of course about the labelled items itself. For example, Last.fm allows users showcase their preferences by tagging artists, albums or music tracks and Del.icio.us allows users to tag webpages. Users annotate an item such as photos, videos, blogs etc., for which is otherwise difficult to generate attributes, by introducing a tag. A set of triples -user, item, tag form information spaces referred to as folksonomies [12] .Recommending tags can serve various purposes, such as: increasing the chances of getting an entity annotated, reminding a user what an entity is about and consolidating the annotation across the users [15]. The collection of all his assignments is called his personomy, the collection of all personomies constitutes the folksonomy.

Jäschke et. al (2008) compared several approaches for tag recommendation in the domain of social bookmarking system. They evaluated an adaptation of user-based collaborative filtering, a graph-based recommender built on top of the Folk Rank algorithm and several simpler approaches based on tag counts. They computed the complexity and compared these algorithms over three real world folksonomy datasets from del.icio.us, BibSonomy and last.fm, and found that most popular tags ρ–mix approach outperformed all other approaches as it is can almost reach the grade of FolkRank (which was powerful but cost intensive) and is extremely cheap to generate [15]. They have been used on small datasets, their performance on big datasets has not be evaluated.

Stefan Siersdorfer, Sergej Sizov (2009) represented Web 2.0 folksonomies as IR-like Vector Space Model and implemented known recommender methodology namely - collaborative filtering and content based filtering using additional social relations obtained from folksonomy features such as posts, contacts, favorites over it. They provided a large-scale experimental study for photo and contact recommendations on Flickr6 comparing a variety of object representations. The study showed that the common relationship model between users, items, and annotations is often not sufficient for constructing accurate recommendation algorithms in folksonomies. Personalized models which consider user's personal data and the local neighborhood for modeling provide higher accuracy at the noticeably lower computational and modeling costs [12].

Nan Zheng, Qiudan Li (2011) investigated the usefulness of tag and time information in predicting user's preference and integrated this information into CF for building effective resource-recommendation model in Social Tagging Systems (STS). They realized this model in 3 phases where first they generated ratings based on tag-weight, time-weight and tag-time weight. In the second phase, used generated rating information to calculate user similarity finally in third phase, recommended the resource. They supported their research with empirical results by using a real-world dataset. Further they proposed to evaluate their model using other datasets [18].

Ma, H., Zhou, T. C., Lyu, M. R. and King, I (2011) proposed a generic framework by amalgamating user item rating matrix and users' social trust network by performing probabilistic matrix factorization analysis. Further, they extended the framework incorporating social tag information. They conducted the experiments on two different datasets: Epinions for social trust network, Movielens for tag information. The experimental results show that their approach outperformed the other contemporary CF algorithms, and the complexity analysis indicated scalability to huge datasets. The limitation when consolidating the social trust network information is they ignore the diffusion or propagation of information between users. Also, a more general framework could be designed to incorporate tags with users and items simultaneously, than associate tags with users and items individually [8].

Tan, S., Bu, J., Chen, C., Xu, B., Wang, C., and He, X (2011) proposed music recommendation hypergraph (MRH) algorithm wherein they incorporated various kinds of social media based information and music acoustic-based content. They used hypergraph to advance into a unified framework taking into account all objects and relations. Recommendation was reduced to a ranking problem on this hypergraph. To evaluate their algorithm, they collected data from Last.fm. They also compared the MRH algorithm with MRH- variant algorithms and some traditional methods. They found that the proposed algorithm significantly outperforms its variants and traditional recommendation algorithms [9].

Jian Jin and Qun Chen (2012) proposed a Top-K recommender system which is based on social tagging network. The tag information is the representation of the item. Feature matrix is constructed by gathering information on all items annotated by tags (Item-tag). So the more tags an item has, the more complete semantic information it has. This matrix formed the basis for Item similarity computation. Then a User-tag matrix is constructed which gathered information about the number of times User i uses tag ij when he tags item j in the item set.

TABLE II. SYSTEMATIC REVIEW OF LITERATURE OF SOCIAL RECOMMENDER SYSTEM

Type of data	Year	Author	Approach/Algorithm	Dataset used	Domain of Recommendation	Result
Social Links data	2007	Wolfgang Woerndl , Georg Groh [28]	Social neighbourhood based recommender	Lokalisten – social community	Restaurant Recommendation	Outperformed traditional collaborative filtering algorithms
	2010	Fengkun Liu , Hong Joo Lee [29]	Hybrid approaches utilizing social network information in CF methodologies	Cyworld, a social networking Web site in South Korea	Digital item (homepage skin) recommendation	Enhancing Recommendation performance.
	2011	Kazienko et. al. [16]	Multidimensional social network.	'Flickr'- photo sharing system	Recommendation of other users' to active user of MSS	Contribution of certain layers including tag layer more than other layers.
	2013	Xin Liu and Karl Aberer [30]	SoCo recommender system	Douban- largest Chinese social platforms	Item(Book, Movie, Music) Recommendation	SoCo outperformed compared to the benchmarked context-aware and social Recommender systems
Social tags	2008	Jäschke et. al. [15]	Most popular tags ρ–mix tag recommender	Del.icio.us Dataset, BibSonomy dataset, Last.fm dataset	Tag recommendation in Bookmark Recommender System	Outperforms of user-based collaborative filtering, a graph-based recommender
	2009	Stefan Siersdorfer , Sergej Sizov [12]	Recommender techniques(CF,CBF) built over Vector Space Model representation	Flickr portal	User and Photo Recommendation	Emphasized importance of Personalized models to provide higher accuracy
	2011	Nan Zheng, Qiudan Li [18]	Integration of Tag and Time Information into CF	Real world dataset of in Social Tagging Systems(STS)	Resource recommendation	Tag, time and both tag and time outperform traditional log-based model
	2011	Ma, H., Zhou, T. C., Lyu, M. R. and King, I. [8]	Integration of social contextual information and the user-item rating matrix, based on a probabilistic matrix factorization	Epinions , MovieLens	Movies Recommendation	Outperformed the other state-of-the-art collaborative filtering algorithms
	2011	Tan, S., Bu, J., Chen, C., Xu, B., Wang, C., et. al. [9]	Modelling high-order relations in social media information by hypergraphs	Last.fm dataset	Music Recommendation	Outperforms its variants and traditional recommendation algorithms
	2012	Jian Jin and Qun Chen [17]	MWalker(modified- Tustwalker algorithm	Last.fm dataset	Music Recommendation	Overcomes the problem of explicitly stating the trust values in the networks by users.
	2014	Bastian, Mathieu, et al. [31]	Skill inference algorithm using CF	LinkedIn	Skill Recommendation	Led to far greater numbers of members adding skills to their profiles than before.
User-generated data	2012	Yung-Ming Li, et. al. [11]	SNMC analysis model	Online forum community	Discussion thread/ Expert recommendation	Better precision and recall rates than other standard methods.

This matrix is used to calculate user similarity based on tagging the same item. The trust value between two friends could be abstracted and trust-based social network could be perfected. The recommendation algorithm used is MWalker (modified- Tustwalker) for Last.fm dataset. This approach outperformed the two traditional CF algorithms. It overcomes the problem of explicitly stating the trust values in the networks by users, which are subjective processes and the Cold start problem of traditional CF. The one limitation reported is there is redundancy in between tags [17].

Bastian, Mathieu, et al (2014) have presented their experiences developing "Skills and Expertise" which allows its users to tag themselves with subjects expressing their areas of proficiency, a data-driven feature on LinkedIn,. Herein, they developed large-scale topic extraction pipeline on Hadoop platform in which they constructed a folksonomy of skills and expertise to assist the users in standardizing information in the skills section and provide a type-ahead. And then create a skill inference algorithm which is a collaborative filtering (CF) skills recommender system based on profile attribute data which would directly suggest additional skills to members through a recommender system.

A large number of members adding skills to their profiles led by the recommender system was one of the major benefit. Author also suggested that the extending it to include other foreign languages will be a compelling challenge [31].

D. User Generated Information

The users-generated content (e.g., comments, blogs, posts, opinions etc.) along with their social network links have stirred a new trend in improving the recommendation results. Semantic resemblance, social closeness and popularity are some of the additional aspects that could be employed as information source for measuring social information.

Yung-Ming Li, Tzu-Fong Liao, Cheng-Yang Lai (2012) modelled SNMC (Social network-based Markov Chain) by integrating semantic similarity, expertise, social intimacy for knowledge sharing to generate discussion threads and expert recommendations into analysis in online forums [11]. The systematic review of literature of Social Recommender system is summarized in Table II.

IV. Future of Recommender System and Challenges

From the discussion so far, the success of any good recommender system is based on a comprehensive consideration set of information sources. The kind of information source used has a great impact on the recommendation quality. Therefore with the advent of web 3.0, context-aware information (e.g., geo-social information) and information from a variety of sensors (e.g., sensors for measuring various health data) along with the above information would be incorporated. Currently, only the geographic location [24] of the user is included in recommendation system. Other expected data that could be incorporated is RIFD, surveillance data etc. [3]. The future of recommender systems lies in internet of things.

The understanding of dependencies and correlations between preferences in different domains led to transference and exploitation of user acquired knowledge from one domain to several other domains. Tobias et.al [32] conducted a survey have highlighted various Cross Domain Recommender systems.

The growing size of folksonomies poses interesting challenges and opportunities for searching and mining useful content and finding other users sharing the same interests [12]. Analysis of such “Big” data is the one of the key issues faced by designers of recommender systems.

Another is Privacy, an important issue because these systems exploit information from social networking sites which contain a lot of information about its registered users. Sharing of such information by companies may pose identity threats.

Other issues are Difficulty in acquiring feedback from users due to the fact that users don’t really rate the items (as in CF), therefore almost impossible to determine whether the recommendation made was correct or not. Also, Recommender systems (mainly in E-commerce) experience shilling attacks which generate rating biasness for example while products from competitors receive negative ratings, the product of company X receives positive ratings. These systems are highly susceptible to such external influences.

V. Conclusion

From the literature review, it can be concluded that Social information aided Recommender System have outperformed most traditional systems in making effective recommendations. Social link information have been captured from real time social networking sites and used in devising Hybrid approaches utilizing fundamental CF methodologies [16], [28], [29], [30]. As the online content is progressively being created, edited and shared over social network communities social tagging provides a powerful way for users to organize, administer, consolidate and search for innumerable kinds of resources. Tags are considered as an expression of user’s preference towards a certain resource over time, a denotation of user’s interest drift [18]. To best of my capacity, limited literature is available on social tag recommendation in different domains areas. It has been explored in areas like –Bookmarking, Bibliographic references, Music, Movies, Books, Skills. Recommending better tags is dependent on creation of improved folksonomy. In addition to it, a rich set of structures and annotations that can be used for mining in variety of purposes include a range of descriptive metadata, such as author, a textual narration of media item, tags expressing theme of an item, historical and geographic information pertaining to an item, and comments and preference logs by other users.

References

[1] Dell'Amico, Matteo and L. Capra, "Sofia: Social filtering for robust recommendations," Trust Management II Springer US, pp. 135-150, 2008.

[2] Konstan, A. Joseph and J. Riedl, "Recommender systems: from algorithms to user experience.," User Modeling and User-Adapted Interaction, vol. 22.1, no. 2, pp. 101-123, 2012.

[3] J. Bobadilla, F. Ortega, A. Hernando and A. Gutiérrez., "Recommender systems survey," Knowledge-Based Systems 46, pp. 109-132, 2013.

[4] R. Hu and P. Pu., "Enhancing collaborative filtering systems with personality information," in In Proceedings of the fifth ACM conference on Recommender systems (RecSys '11), ACM, New York, NY, USA, 2011.

[5] P. Massa and P. Avesani, "Trust-aware collaborative filtering for recommender systems," On the Move to Meaningful Internet Systems 2004:CoopIS, DOA, and ODBASE, pp. 492-508, 2004.

[6] M. Tim Jones, "Recommender systems, Part 1,Part 2: Introduction to approaches and algorithms," IBM DeveloperWorks, 12 December 2013. [Online]. Available: http://www.ibm.com/developerworks/library/os-recommender1/. [Accessed 30 June 2015].

[7] R.Burke, "Hybrid recommender systems: survey and experiments," User Modeling and User-Adapted Interaction 12, p. 331–370, 2002.

[8] H. Ma, T. C. Zhou, M. R. Lyu and I. King, "Improving recommender systems by incorporating social contextual information," ACM Trans. Inf. Syst. 29, vol. 2, no. 9, p. 23, 2011.

[9] S. Tan, J. Bu, C. Chen, B. Xu, C. Wang and X. and He, "2011. Using rich social media information for music recommendation via hypergraph model.," ACM Trans. Multimedia Comput. Commun. Appl. 7, vol. 1, no. 22, p. 22, 2011.

[10] W. Carrer-Neto, M. L. Hernández-Alcaraz, R. Valencia-García and a. F. García-Sánchez., "Social knowledge-based recommender system. Application to the movies domain.," Expert Syst. Appl. 39, pp. 10990-11000, 2012.

[11] Y.-M. Li, T.-F. Liao and C.-Y. Lai., "A social recommender mechanism for improving knowledge sharing in online forums," Information Processing & Management, vol. 48, no. 5, pp. 978-994, 2012.

[12] S. Siersdorfer and S. Sizov, "Social recommender systems for web 2.0 folksonomies," in Proceedings of the 20th ACM conference on Hypertext and hypermedi, Torino, Italy, 2009.

[13] G. Ruffo and R. Schifanella, "A peer-to-peer recommender system based on spontaneous affinities.," ACM Trans. Intern. Tech. 9, vol. 1, no. 4, p. 34, 2009.

[14] G. Pitsilis and S. Knapskog, "Social trust as a solution to address sparsity-inherent problems of recommender Systems," in Proceedings of the 2009 ACM Conference on Recommender Systems, 2009.

[15] R. Jäschke, L. Marinho, A. Hotho, L. Schmidt-Thieme and G. Stumme, "Tag recommendations in social bookmarking systems.," AI Commun. 21, pp. 231-247, 2008.

[16] P. Kazienko, K. Musial and T. Kajdanowicz, "Multidimensional Social Network in the Social Recommender System," Systems, Man and Cybernetics, Part A: Systems and Humans IEEE Transactions, vol. 41, no. 4, pp. 746-759, 2011.

[17] J. Jin and Qun Chen, "A trust-based Top-K recommender system using social tagging network,," in 9th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), 2012.

[18] N. Zheng and Qiudan Li, "A recommender system based on tag and time information for social tagging systems," Expert Systems with Applications, vol. 38, no. 4, pp. 4575-4587,, 2011.

[19] Park, D. Hee and e. al, "A literature review and classification of recommender systems research," Expert Systems with Applications, Vols. 39-11, pp. 10059-10072, 2012.

[20] Schafer, J. Ben, J. Konstan and a. J. Riedl., " Recommender systems in e-commerce," in Proceedings of the 1st ACM conference on Electronic commerce. ACM, 1999.

[21] Tejeda-Lorente, Álvaro and e. al., "A quality based recommender system to disseminate information in a university digital library," Information Sciences, vol. 261, pp. 52-69, 2014.

[22] G. Meghabghab and A. Kandel, Search Engines, Link Analysis, and User's Web Behavior, Berlin Heidelberg: Springer-Verlag, 2008.

[23] K. Oku, R. Kotera and K. Sumiya, "Geographical recommender system based on interaction between map operation and category selection.," in Proceedings of the 1st International Workshop on Information Heterogeneity and Fusion in Recommender Systems, Barcelona, Spain, 2010.

[24] C. Matyas and Christoph Schlieder, "A Spatial User Similarity Measure for Geographic Recommender Systems.," in Proceedings of the 3rd International Conference on GeoSpatial Semantics, Mexico City, Mexico., 2009.

[25] M.-H. Park, J.-H. Hong and S.-B. Cho, "Location-based recommendation system using Bayesian user's preference model in mobile devices," in Proceedings of the 4th international conference on Ubiquitous Intelligence and Computing, Hong Kong,China, 2007.

[26] W. Yuan, D. Guan, Y.-K. Lee, S. Lee and S. J. Hur, " Improved trust-aware recommender system using small-worldness of trust networks," Knowledge-Based Systems, vol. 23, no. 3, pp. 232-238, 2010.

[27] P. Victor and e. al., "Gradual trust and distrust in recommender systems," Fuzzy Sets and Systems , vol. 160, no. 10, 2009.

[28] W. Woerndl and G. Groh, "Utilizing Physical and Social Context to Improve Recommender Systems," in Proceedings of the 2007 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology, 2007.

[29] F. Liu and Hong Joo Lee, "Use of social network information to enhance collaborative filtering performance," Expert Systems with Applications: An International Journal,, vol. 37, no. 7, pp. 4772-4778, 2010.

[30] Liu, Xin and K. Aberer., "SoCo: a social network aided context-aware recommender system," in Proceedings of the 22nd international conference on World Wide Web. International World Wide Web Conferences Steering Committee, 2013.

[31] Bastian, Mathieu and e. al., "LinkedIn skills: large-scale topic extraction and inference," in Proceedings of the 8th ACM Conference on Recommender systems, 2014.

[32] F.-T. Ignacio and e. al., ""Cross-domain recommender systems: A survey of the state of the art," in Spanish Conference on Information Retrieval.

Novel Approach to Estimate Missing Data Using Spatio-Temporal Estimation Method

Aniruddha D. Shelotkar
Research Scholar,
Department of Electronics Engineering
SGB Amravati University, Amravati
Maharashtra, India

Dr. P. V. Ingole
Principal,
G. H. Raisoni Institute of Engineering
and Management, Amravati
Maharashtra, India

***Abstract*—With advancement of wireless technology and the processing power in mobile devices, every handheld device supports numerous video streaming applications. Generally, user datagram protocol (UDP) is used in video transmission technology which does not provide assured quality of service (QoS). Therefore, there is need for video post processing modules for error concealments. In this paper we propose one such algorithm to recover multiple lost blocks of data in video. The proposed algorithm is based on a combination of wavelet transform and spatio-temporal data estimation. We decomposed the frame with lost blocks using wavelet transform in low and high frequency bands. Then the approximate information (low frequency) of missing block is estimated using spatial smoothening and the details (high frequency) are added using bidirectional (temporal) predication of high frequency wavelet coefficients. Finally inverse wavelet transform is applied on modified wavelet coefficients to recover the frame. In proposed algorithm, we carry out an automatic estimation of missing block using spatio-temporal manner. Experiments are carried with different YUV and compressed domain streams. The experimental results show enhancement in PSNR as well as visual quality and cross verified by video quality metrics (VQM).**

Keywords—Error concealment; Wavelet Transform; Missing Data estimation

I. INTRODUCTION

With enhancement in wired and wireless networks, more and more users are demanding video services, including video conferencing and video streaming over the internet. However, the Internet does not provide guaranteed quality of service (QoS). Loss of data packets occur due to traffic congestion [1]. In wireless networks, packet loss happens frequently due to shadowing, multipath fading, and noise disturbance of wireless channels [3]. Video transmission uses compressed video streams for transmission so that video data can be transmitted even with poor network bandwidth situations [2]. A loss of packet over transmission in compressed stream introduces severe distortion because the compression algorithms use spatial estimation methods and temporal to improve compression efficiency. Therefore a single distorted block within a frame may occur errors not only in present frame but also propagate error over several frames. Many decoder error concealment techniques and error resilience have been proposed to control amount of error in reconstructed frame [4]. A simple error resilience approach is to use feedback channels and request for retransmission whenever there is error. This is the most prosperous technique and the recovered data would always be correct. However, it involves halting decoding process till error block of data is received again. This is an inefficient approach in terms of delay involved in process. Another way to avert errors is to embed error checks in encoded video bit streams and transmit over the channels. This method though bypasses retransmission of video; it affects compression efficiency of the encoder and thus increased usage of network bandwidth.

Hence, in this paper a post processing algorithms on the decoder side are proposed for error concealment. The preference with decoder error concealment is that it does not require any change in encoding or decoding process. It simply appends a post processing block which retrieves erroneous data. Hence, there is no increase in bit rate or delay. Fig.1 shows block diagram for process in which packet loss occurs in channel and video sequence to recover loss of macroblocks. Therefore these methods can be used in real time video applications like video-voice over internet and streaming applications.

The complete video coding system can be organized according to the blockdiagram in Fig. 2 The building blocks of the video coding system including post processing at the decoder are summarized below which give complete idea about video coding system[17].

- Video Acquisition—Source of the video sequence which is output in a digital form. Following are the processing steps.
- Pre-Processing—Operations on the raw uncompressed video source material, such as color correction, or de-noising trimming, color format conversion.

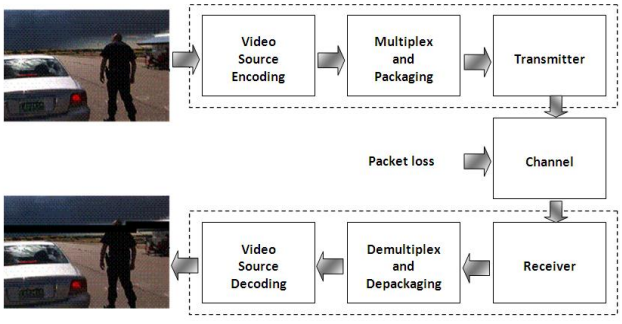


Fig. 1. Block diagram of process in which packet loss occurs in channel

- Encoding— The aim of encoding is to generate a compact representation of the input video sequence which is applicable for the transmission method in the given application scenario.
- Transmission—Transmission section includes sending and delivery of the video data to the receiver side.

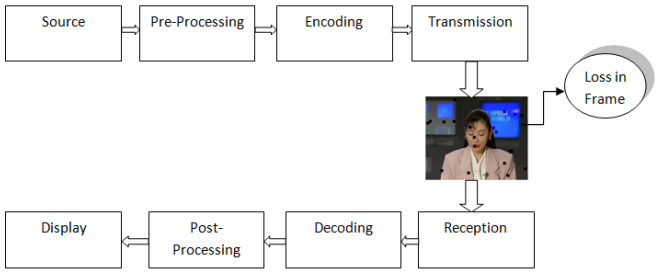


Fig. 2. Block diagram of video coding system including post-processing at the decoder

In transmission packet loss occur in video. In fig.2 loss occur in frame while transmission due to lossy network or compression of video sequences.

- Decoding—Encoding usually use compression to achieve the target transmission bitrate constraints due to which losses occur. The decoded video constitutes a resemblance of the original source video. If inaccessible transmission losses have occurred, the decoder implements concealment strategies to recover the corrupted video sequence as much as possible.
- Post-Processing—Post-processing perform operations on the reconstructed video sequence for enhancement or for adaptation of the sequence for display. These operations can e.g. include trimming, or re-sampling and color correction.
- Display—Display use for presentation of the video sequence for viewing. In real applications, example multiple pre-processing /post-processing steps in conjunctions with iterative de- and re-encoding and transmission may be used.

Therefore in this paper, we propose an approach based on combination of spatial and temporal concealment using spatial smoothing filter and wavelet transform for missing block estimation. Spatial smoothing is used to estimate missing block using spatial information present in the frame as, most of the information lies in low spatial frequency. And to add details we used wavelet based bidirectional estimation of high frequency information. The wavelet transform is used because it has ability to represent information in logarithmic way (low to high details) and hence can be used to estimate details (edges) separately.

The rest of the paper is organized as follows,

Section 2 reviews the theory behind the algorithm, section 3 Theory based on proposed new algorithm to improve error concealment, section 4 shows proposed algorithm section 5 shows simulation results and section 6 concludes the paper.

II. Previous Work on Error Concealment Techniques

A. Forward Error- Concealment Technique

There are different error resilience techniques: forward, concealment, and interactive techniques [18]. Almost all forward techniques boost the bit rate since they add redundancy to data. Few of them require modification to the encoder. Interactive methods need a feedback channel between the encoder and the decoder. Interactive methods will also announce some delay and may, therefore, be improper for real-time applications like video communications. On the other hand, concealment techniques do not boost the bit rate, do not require any modifications of the encoder, and do not boost any delay. This makes them a very alluring choice for video communications [19]. An error concealment method plays primary role at the encoder side. When the transport channel is not lossless, there are two kinds of distortion observed at the decoder. The first one is quantization noise imported by the waveform coder and the second is the distortion due to transmission errors. Both quantization and transmission errors are minimized by using optimal pair of transport coder and source coder such as transport protocol, packetization and FEC. Quantization error is minimizing by video codec given in available bandwidth. Shannon proposed separation theorem useful for source and channel that are memeoryless and stationary [20] and later extended to more general class of source and channel [21]. Joint design of source and channel coder achieves better performance. To accomplish forward error concealment they all add redundancy either in source coder and transport coder. All these techniques achieve error resilience by adding certain amount of redundancy such as layered coding, multiple description coding, robust waveform coding, robust entropy coding. Summary of Forward error concealment techniques proposed Y.Wang [10]. It explains Layered Coding with prioritized transport includes frequency domain partitioning, Successive amplitude refinement and spatial/temporal resolution refinement. Note that these techniques are not mutually exclusive; rather, they can be used together in a complementary way [10-16].

B. Error Concealment by Post-processing /Error Re-concealment

The loss of transmitted data packets affects the quality of the received video and this problem is caused by the limited channel bandwidth used by the mobile communication networks. It is not possible to retransmit lost packets in real time application. Generally post processing techniques is used to reduce the visual artifacts caused by bit stream error [22]. Error concealment methods which will be implemented on the receiver side to restore the missing and corrupted video frame using the previously decoded video data [4]. It will be noted that various post processing technique are successfully implemented such as Motion compensated temporal prediction, Spatial Interpolation, Maximally Smooth recovery, POCS, Fuzzy logic which is review by Y.Wang [10-16].

III. THEORY BASED ON PROPOSED ALGORITHM

In proposed algorithm we estimate the missing block using hybrid approach. Wherein, we first decompose the video frame using wavelet transform. To estimate the missing data we apply smoothing function on approximate wavelet coefficients and use bi-directional prediction on the detail wavelet coefficients. The reason behind using bidirectional wavelet prediction is, the detail information (edges) cannot be approximated and needs to be predicated using the available details in neighborhood frames. Once spatio-temporal predication is done we reconstruct the frame using inverse wavelet transform.

In this section we quickly revisit the basics of wavelet transform and spatial smoothing.

A. Wavelet transform

Wavelet transform have a wide range of applications ranging from analysis of image signal to data compression [5]. In general wavelet transform of time varying signal x(t) is calculated by taking inner product of signal against family of wavelets. These wavelets, $\varphi_{a,b}(t)$ are labeled by scale and time location parameters a and b. In continuous wavelet transform, the wavelet corresponding to scale a and time location b is given as,

$$\varphi_{a,b(t)} = \frac{1}{\sqrt{a}} * \varphi\left(\frac{t-b}{a}\right)$$

Where $\varphi(t)$ is a wavelet prototype function which can be thought of as a bandpass function. The continuous wavelet transform is given by,

$$CWT\{x(t); a, b\} = \int x(t) * \varphi^*_{a,b}(t)dt$$

Where $\varphi^*(t)$ denotes complex conjugate of $\varphi(t)$. Time t and time-scale parameters a, b vary continuously. Time remains continuous but time-scale parameters and sampled on a dyadic grid in time-scale (a, b) space. This is defined as,

$$C_{j,k} = CWT\{x(t); a = 2^j, b = k * 2^j\} for\ j, k \in \mathbb{Z}$$

The wavelets in this case are

$$\varphi_{j,k}(t) = 2^{-j/2} * \varphi(2^{-j} * t - k)$$

The original signal can be recovered through

$$x(t) = \sum_{j\in\mathbb{Z}} \sum_{k\in\mathbb{Z}} C_{j,k} * \varphi^{\sim}_{j,k}(t)$$

Wavelet functions $\varphi(t) and\ \varphi^{\sim}(t)$ form orthogonal basis.

The discrete wavelet transform (DWT) corresponds to continuous wavelet transform of a sampled sequence $x_n = x(nT)$, where T is a sampling period. The discrete wavelet transform applies to discrete time signals where both time and scale parameters are discrete. The DWT is represented by

$$DWT\{x(n); 2^j; k * 2^j\} = C_{j,k} = \sum_n x_n * h^*_j(n - 2^j * k)$$

Where $h^*_j(n - 2^j * k)$ denotes complex conjugate of $h_j(n - 2^j * k)$. The residual coefficient at J are given by

$$r_{j,k} = \sum_n x_n * g^*_j(n - 2^j * k)$$

Where $g_j(n - 2^j * k)$ is the analysis scaling sequence. It is used to bring input signal from initial scale to 2^j. One of the main concept of wavelet theory is the interpretation of wavelet transform in terms of multiresolution decomposition. The input signal x(n) is decomposed into approximate and detailed coefficients using a set of low pass(H) and high pass(G) filters followed by a decimator. These filters are quadrature mirror filters and are related by,

$$h(n) = (-1)^n * g(M - 1 - n)$$

where M is filter length.

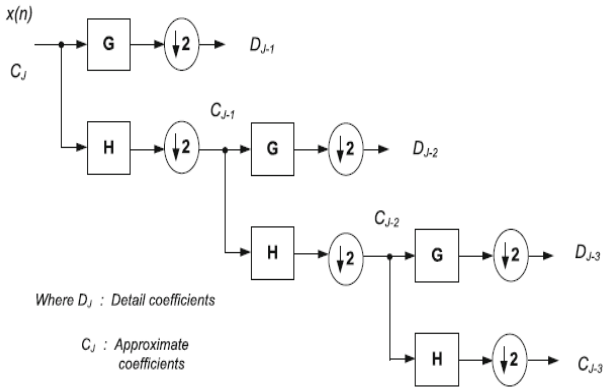


Fig. 3. Computation of DWT

In our algorithm we use wavelet decomposition and then use prediction on detail coefficients to estimate details of missing blocks (high frequency information in form of edges).

B. Smoothing function

In statistics and data analysis smoothing function is used to reduce noise or small scale information while keeping most imprints of the datasets [6,7]. Mathematically, noisy data can be represented as follows,

$$y = y_m + \in \quad (1)$$

Where $\in$ represents Gaussian noise with zero mean and unknown variance and y_m is the mean of the signal. Accuracy of estimation of y depends upon accuracy of y_m. Also, y_m is supposed to be smooth function that is derivatives of y_m of particular order generally greater than 2 are continuous. Smoothing of y relies upon smoothing of y_m. Here penalized least squares approach for smoothing of data is used. Mathematically, it can be expressed as,

$$F(y_m) = RSS + s * P(y_m) \quad (2)$$

Where, RSS is residual sum of squares which is expressed as,

$$RSS = ||y - y_m||^2$$

$P(y_m)$ is the penalized term, and S is scalar which indicates degree of smoothing, with increase in smoothing

parameter, degree of smoothing also increases. In [6], it is explained that term $P(y_m)$ can be expressed as,

$$P(y_m) = ||D * y_m||^2 \quad (3)$$

Where D is a tri-diagonal square matrix defined as,

$$D_{i,i-1} = \frac{2}{h_{i-1} * (h_{i-1} + h_i)}$$

$$D_{i,i} = \frac{-2}{h_{i-1} * h_i}$$

$$D_{i-1,i} = \frac{2}{h_i * (h_{i-1} + h_i)}$$

Where h_i represents step between $y_{m,i}, y_{m,i+1}$

To correctly estimate smoothing parameter, we minimize equation (2) with constraints of equation (3). Hence, smoothened data can be obtained as,

$$(I_n + s * D^T * D) * y_m = y \quad (4)$$

Where I_n is $n\ x\ n$ identity matrix and D^T is transpose of D.

As discussed above we formulate the missing block estimation problem in approximate wavelet sub band. We treat the approximate coefficients of missing block in the given frame as missing data and hence estimate using smoothing given in equation (4).

Effectively the concealments in done in low frequency and high frequency band separately. Low frequency coefficients are estimated as spatial smoothing (using missing data estimation given in section 3.2 and high frequency coefficients are predicted using high frequency bands of previous and next frames).

C. Error concealment

Temporal error concealment methods assume that motion in the video is not dramatic (Sum of Absolute difference between consecutive frames is less) and do follow some linear or quadratic model. We first detect the video shot boundary to detect the scene change. If the block loss happened in the frame is not the key frame then we can use bidirectional predication and else we restrict the algorithm as a unidirectional prediction. Literature gives multiple algorithms for the video shot boundary detection [8,9]. We use correlation between consecutive frames as a measure for shot boundary detection,

$$cc(i) = \frac{(\sum_{x,y}[f_{i(}(x,y) - \mu_i] * [f_{i+1}(x,y) - \mu_{i+1}])}{\sqrt{\sum_{x,y}[f_{i(}(x,y) - \mu_i]^2 * [f_{i+1}(x,y) - \mu_{i+1}]^2)}}$$

We apply adaptive threshold for detecting scene change based on correlation [4]. The threshold use is,

$$Th_n = \frac{\sum_N cc_n}{N}$$

Where, N is the size of cross correlation vector window use for calculation.

If the block loss happen in the frame which is not a key frame (frame after scene change) but having second frame after key frame then we use the first order bidirectional prediction on detail wavelet coefficients as,

$$IB_{avg}(i,j) = \frac{IB_{n-1}(i,j) + IB_{n+1}(i,j)}{2}$$

$$IB_n(i,j) = a * IB_{avg}(i,j) + b * IB_n$$

Where, $IB_n(i,j)$ represent lost block and n is frame index. IB is the low pass band in wavelet transform.

The high frequency components of the wavelet transformed image are restored using linear combination of reference frames.

$$HB_n = c * HB_{n-1} + d * HB_{n+1}$$

Where HB represents high frequency component of wavelet transformed image.

Therefore, to we propose a technique which uses multiple reference frames to recover frame data. Following figure explains proposed algorithm.

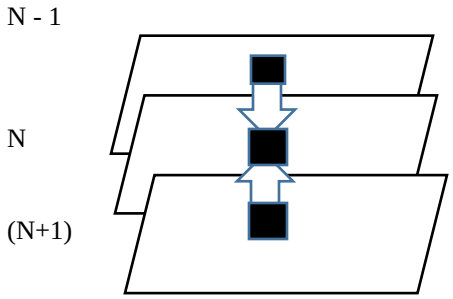


Fig. 4. Proposed block recovery technique

IV. PROPOSED ALGORITHM

The proposed algorithm is as follows. We use previous and next frame as reference frames

1. Read frames from incoming video streams.
2. Detect the video shot boundary.
3. Decompose frames using wavelet transform.
4. Estimate approximate wavelet coefficients (low frequency information) of transformed frames using robust smoothing of gridded data in one or higher dimension algorithm given in section 3.2.
5. Estimate detail coefficients (high frequency data in the current frame) by bi-liner predication given in section 3.1 and 3.3.
6. Restore the frame using inverse wavelet transform (with estimated low frequency coefficients using spatial smoothing and predicated high frequency coefficients using bidirectional predication of detail wavelet coefficients).

Following are the block diagram of proposed algorithm

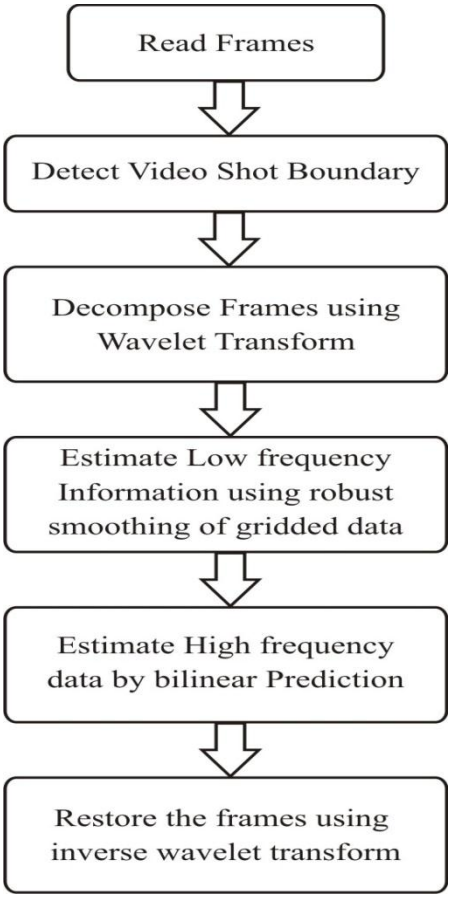


Fig. 5. Block diagram of proposed algorithm

V. SIMULATION RESULTS

This section describes simulation results for the proposed algorithm. The following assumptions are made while conducting experiments, video sequences used for video error concealment are of QCIF and CIF resolution at 30fps, YUV format and MP4, AVI videos for compressed video error concealment. The experiments were conducted on foreman and akiyo video sequences. For our experiments we selected the values of linear coefficients as follows,

$$a = 0.8, b = 0.2, c = 0.5 \; and \; d = 0.5$$

These parameters are chosen empirically and tuned to achieve the maximum reconstruction quality of missing blocks. We use Daubechies 1 wavelet for testing purposes [6].

We implemented for QCIF (176 X 144) and CIF (352 X 288) different video sequences in which error block is reconceal by our proposed algorithm.

Fig.6. shows visual result QCIF (176 X 144) Carphone video sequence at resolution 30 fps which having 380 frames in video shows a) Original Carphone video frame b) It shows corrupted video frame in which 20Macroblock is lost randomly in frame c) It shows Reconceal video frame with help of proposed method

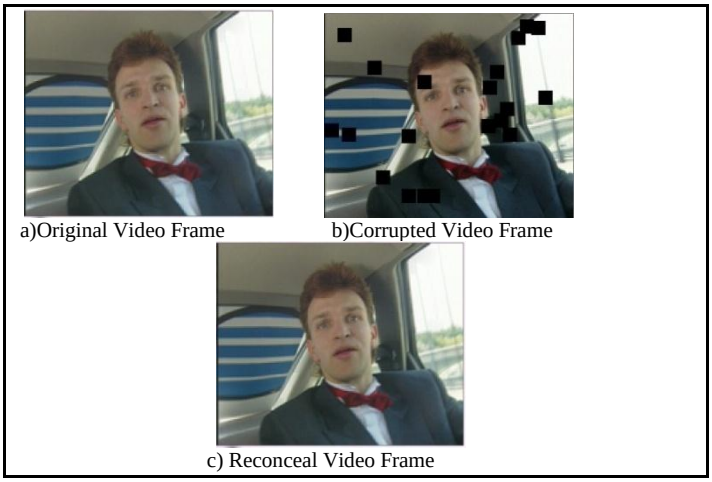
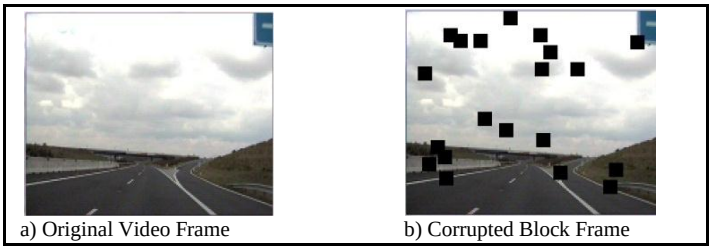


Fig. 6. Visual results shows QCIF "Carphone" video sequences at resolution 30 fps showing a) original video frame b) Corrupted video frame and c) Reconceal videoframe

Fig.7. shows visual result QCIF (176 X 144)Highway video sequence at resolution 30 fps which having 1998 frames in video shows a) Original Highway video frame b) It shows corrupted video frame in which 20Macroblock is lost randomly in frame c) It shows Reconceal Highway video frame with help of proposed method. A "Highway" sequence is high motion video which is concealing by our proposed algorithm effectively. Fig.8. shows visual result QCIF (176 X 144)Forman video sequence at resolution 30 fps which having 298 frames in video shows a) Original Forman video frame b) It shows corrupted video frame in which 40Macroblock is lost randomly in frame c) It shows Reconceal Forman video frame with help of proposed method. Forman video sequences is high motion video which is conceal by our proposed algorithm effectively when scene changes. Fig.9., Fig.10. and Fig.11. shows Error frame index Vs PSNR graph for Carphone, Highway and foreman QCIF (176 X 144) video sequence having different frame number 380, 1998 and 280 respectively. PSNR shows that our proposed algorithm conceals corrupted video frame effectively.

Now proposed algorithm implemented for CIF (352 X 288) video sequences in which error occur at randomly.

Fig.12. shows visual CIF (352 X 288) Akiyo video sequence at resolution 30 fps which having 380 frames in video shows a) Original Akiyo video frame b) It shows corrupted video frame in which 20Macroblock is lost randomly in frame c) It shows Reconceal Akiyo video frame with help of proposed method.




c) Reconceal Video Frame

Fig. 7. Results shows QCIF highway video sequences at resolution 30 fps showing original video, error block video and reconceal video

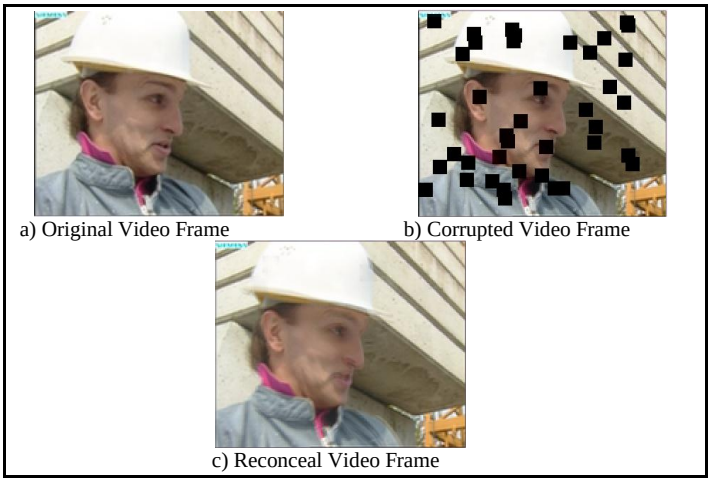

a) Original Video Frame b) Corrupted Video Frame
c) Reconceal Video Frame

Fig. 8. Results shows QCIF forman video sequences at resolution 30 fps showing original video, error block video and reconceal video

Fig.13. shows visual result CIF (352 X 288) Waterfall video sequence at resolution 30 fps which having 258 frames in video shows a) Original Akiyo video frame b) It shows Error block frame in which 30 Macroblock is lost randomly in frame c) It shows Reconceal Akiyo video frame with help of proposed method. Fig.no.14 shows visual result CIF (352 X 288) Tempete video sequence at resolution 30 fps which having 258 frames in video shows a) Original Tempete video frame b) It shows corrupted video frame in which 30 Macroblock is lost randomly in frame c) It shows Reconceal Tempete video frame with help of proposed method.

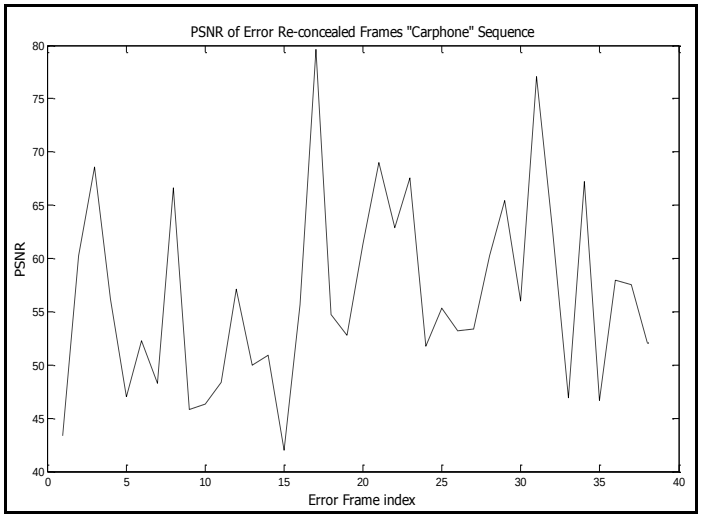


Fig. 9. Error Frame Index Vs PSNR for 'Carphone' Sequence

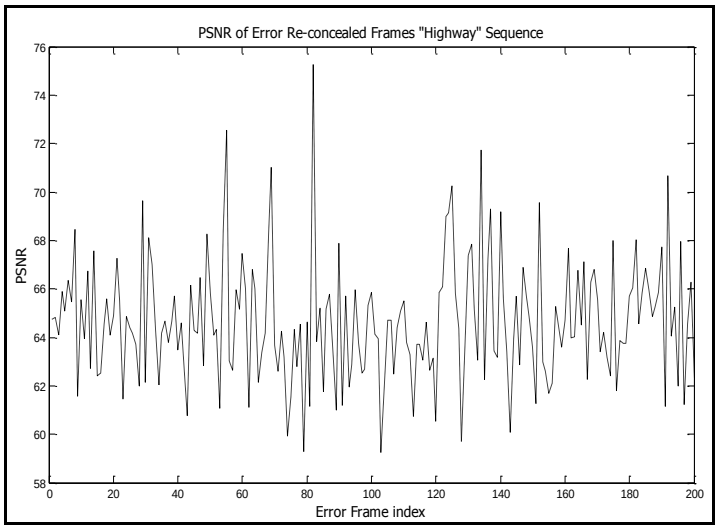


Fig. 10. Error Frame Index Vs PSNR for 'Highway' Sequence

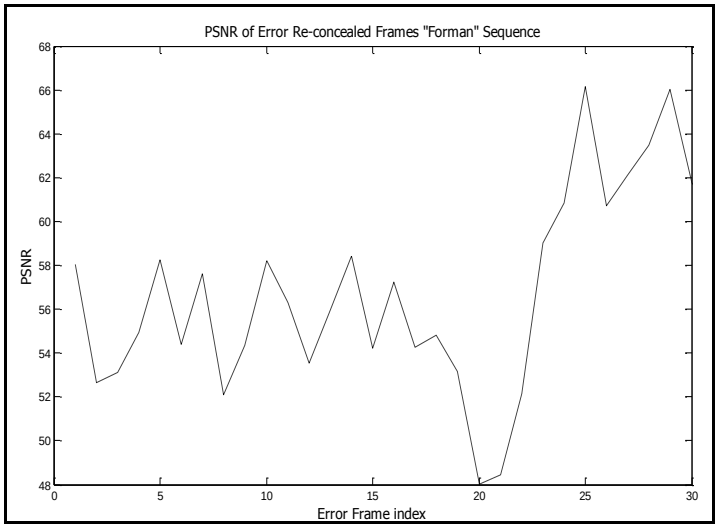


Fig. 11. Error Frame Index Vs PSNR for 'Foreman' Sequence

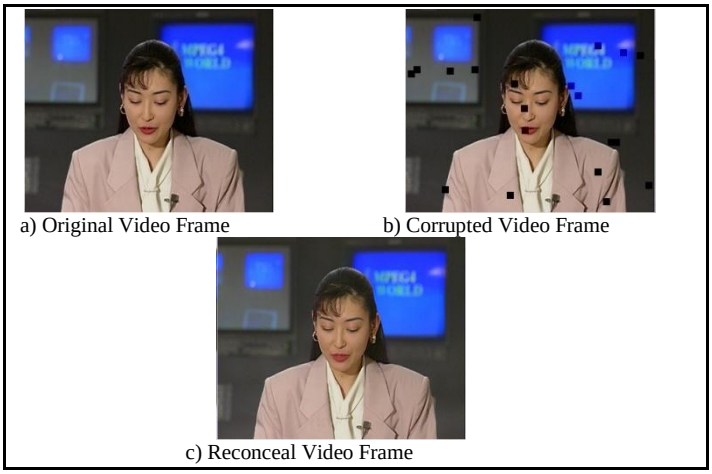
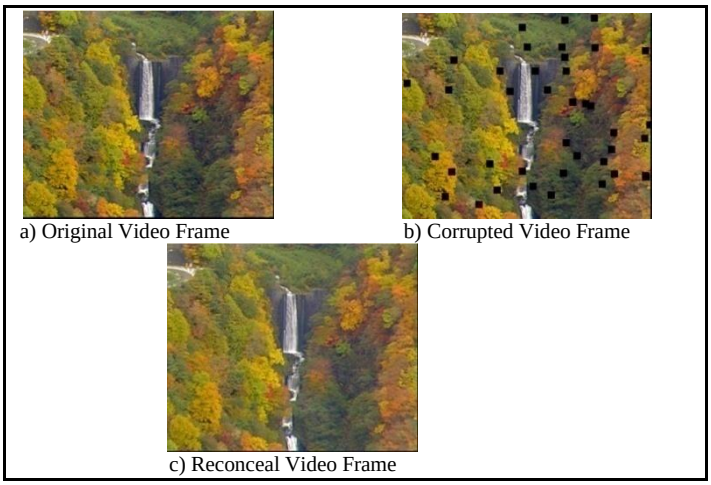

a) Original Video Frame b) Corrupted Video Frame
c) Reconceal Video Frame

Fig. 12. Results shows CIF Akiyo Video Sequences at resolution 30 fps showing Original video, Error block video and Reconceal video

Fig.15, Fig.16 and Fig.17. Shows Error frame index Vs PSNR graph for Akiyo, Waterfall and Tempete CIF (352 X

288) video sequence having different frame number 380,258 and 258 respectively. PSNR shows that our proposed algorithm conceals corrupted video frame effectively.

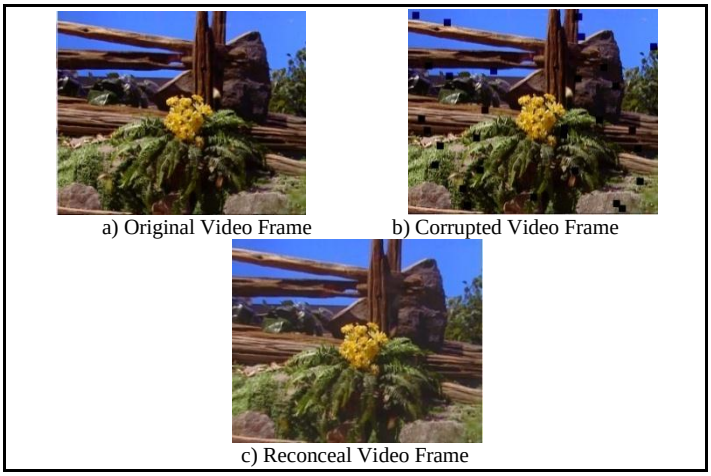


a) Original Video Frame b) Corrupted Video Frame

c) Reconceal Video Frame

Fig. 13. Results shows CIF Waterfall Video Sequences at resolution 30 fps showing Original video, Error block video and Reconceal video

Video Quality Metric is tool to measure quality of video effectively. PSNR_SEARCH is function VQM tool to calculate PSNR and video quality results effectively. We have also cross verified results using the Video Quality Metric Tool. We test various video sequence formats QCIF (172 X 144) for various video scene Carphone, Highway and Forman sequences and CIF (352 X 288) for various test sequences Akiyo, Waterfall and Tempete sequences. We have used PSNR_SEARCH measure for all the test video mention above. We have observed improvement in PSNR in Table I.



a) Original Video Frame b) Corrupted Video Frame

c) Reconceal Video Frame

Fig. 14. Results shows CIF Tempete Video Sequences at resolution 30 fps showing Original video, Error block video and Reconceal video

TABLE I. TEST RESULT USING VQM

Video Sequence Format	Test Video	No. of Frame	PSNR of Corrupted video	PSNR of Reconceal video of Proposed method
QCIF (172 X 144)	Carphone	380	30.19	36.55
	Highway	1998	26.02	36.94
	Forman	298	27.05	37.52
CIF (352 X 288)	Akiyo	298	36.50	40.6
	Waterfall	258	33.38	41.01
	Tempete	258	35.22	35.88

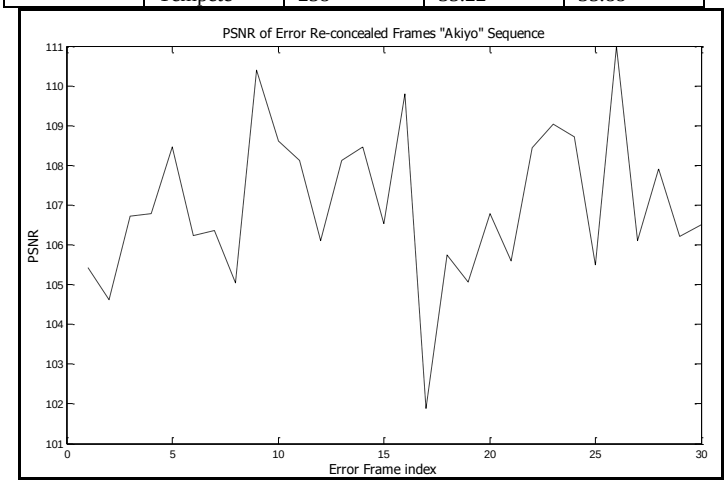


Fig. 15. Error Frame Index Vs PSNR for 'Akiyo' Sequence

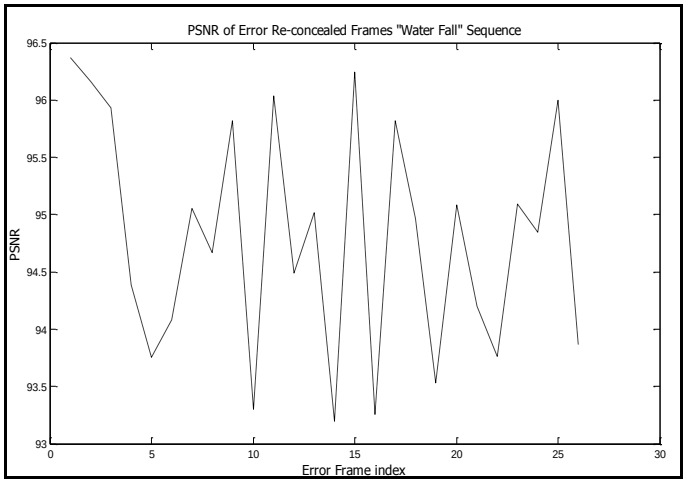


Fig. 16. Error Frame Index Vs PSNR for 'Waterfall' Sequence

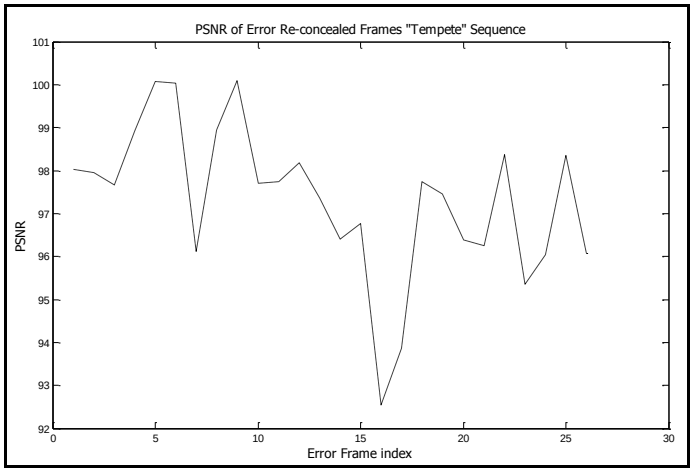


Fig. 17. Error Frame Index Vs PSNR for 'Tempete' Sequence

To show the comparative study we have used the algorithm given in [4], as it is conceptually similar to the proposed algorithm. Fig.18. Shows the concealment results for

one frame using reference and proposed algorithm. Fig.19 Shows the PSNR comparison.

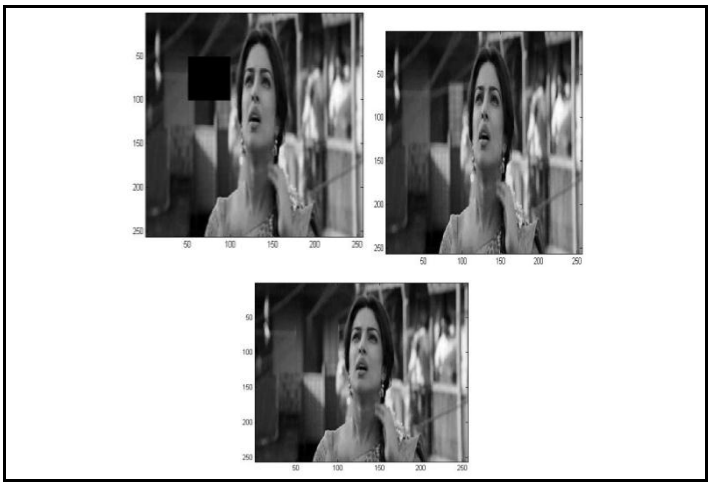


Fig. 18. Results for Compressed Sequence (Bottom right is proposed and bottom is as per algorithm in [4])

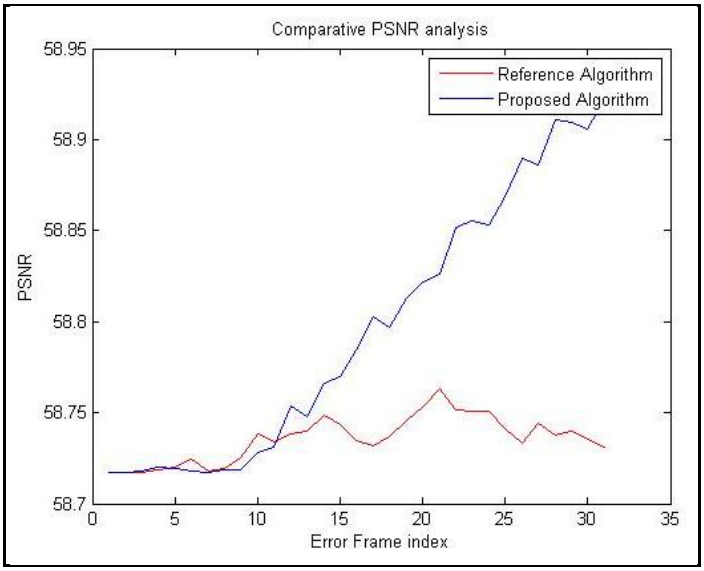


Fig. 19. Comparison between algorithm given in [4] and proposed algorithm

VI. CONCLUSION

In this paper, we propose a low complexity approach for bi-directional spatio-temporal error concealment. The proposed method exploits the information from multiple reference frames as well as information available in current frame to recover lost block of data. We have applied the method on Carphone, Highway, Forman, Akiyo, Waterfall and Tempete video sequence and also on compressed AVI video. Proposed Algorithm does not require very complicated computation and hence usable for various application. Results clearly indicate that the method outperforms existing methods in terms of PSNR as well as cross verified by video quality metrics (VQM) and visual quality.

REFERENCES

[1] Liu, Jing. "Spatial Error Concealment With an Adaptive Linear Predictor."Circuits and Systems for Video Technology, IEEE Transactions on 25.3 (2015): 353-366.

[2] Tang, Xuguo, et al. "Optimizing the MPEG media transport forward error correction scheme." Broadband Multimedia Systems and Broadcasting (BMSB), 2015 IEEE International Symposium on. IEEE, 2015.

[3] Gadgil, Neeraj, He Li, and Edward J. Delp. "Spatial subsampling-based multiple description video coding with adaptive temporal-spatial error concealment."Picture Coding Symposium (PCS), 2015. IEEE, 2015.

[4] Branislav H., Jan M. and Stanislav M., "Extended error concelmant algorithm for intra frame in H.264/AVC", Acta Electrotechnica et Informatica, Vol. 10, No. 4, 2010, pp. 59–63.

[5] Daubechies I., "Ten Lectures on Wavelets", Society for Industrial and Applied Mathematics , CBMS-NSF Regional Conference series in Applied Mathematics, Vol. 61, 1992.

[6] Damien Garcia, "Robust smoothing of gridded data in one and higher dimensions with missing values", Computational Statistics & Data Analysis, 2010; 54:1167-1178.

[7] Craven, P., Wahba, G., "Smoothing noisy data with spline functions. Estimating the correct degree of smoothing by the method of generalized cross validation", NumerischeMathematik1978.31, 377-403

[8] Warhade, Krishna K., S. N. Merchant, and Uday B. Desai. "Shot boundary detection in the presence of illumination and motion." Signal, Image and Video Processing 7.3 (2013): 581-592.

[9] Parul Arora Bhalotra, Bhushan D. Patil, "Video Shot Boundary Detection using Ridgelet Transform", Springer AISC series, Vol- 249, pp 163-171, 978-3-319- 03094-4, Sep 2013.

[10] Y. Wang and Q.-F. Zhu, "Error control and concealment for video communication: A review," Proceedings of the IEEE, vol. 86, no. 5, pp. 974-997, May 1998.

[11] W. Lam, A. Reibman, and B. Liu, "Recovery of lost or erroneously received motion vectors," Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), vol. 5, no. 12, pp. 417-420, April 1993, Minneapolis, MN.

[12] W-Y.Kung, C.-S. Kim, and C.-C. Kuo, "Spatial and temporal error concealment techniques for video transmission over noisy channels," IEEE Transactions on circuits and Systems for Video Technology, vol. 16, no. 7, pp. 789-802, July 2006.

[13] S. Shirani, F. Kossentini, and R. Ward, "A concealment method for video communications in an error-prone environment," IEEE Journal on Selected Areas in Communication, vol. 18, no. 6, pp. 1822-1833, June 2000.

[14] J. Seiler and A. Kaup, "Adaptive joint spatio-temporal error concealment for video communication," Proceedings of the IEEE Workshop on Multimedia Signal Processing, pp. 229-234, October 2008, Cairns, Australia.

[15] N. Gadgil, M. Yang, M. Comer, and E. Delp, "Multiple description coding," Academic Press Library in Signal Processing Vol. 5, R. Chellappa and S. Theodoridis, Eds. Oxford, UK: Elsevier Ltd., 2014.

[16] Y. Wang, A. R. Reibman, and S. Lin, "Multiple description coding for video delivery," Proceeding of the IEEE, vol. 93, no. 1, pp. 57-70, January 2005

[17] Wien, M."Video Coding Fundamental," High Efficiency video coding tools and specification, Proceeding of Springer ISBN: 978-3-662-44275-3, 2015

[18] "Narrow-band visual telephone systems and terminal equipment,"International Telecommunications Union, Geneva, Switzerland, ITU-T Recommendation H.320, 1993.

[19] "Terminal for low bitrate multimedia communication," International Telecommunications Union, Geneva, Switzerland, ITU-T Draft Recommendation H.324, Dec. 1997.

[20] C. E. Shannon, "A mathematical theory of communication," Bell Syst. Tech. J., vol. 27, pp. 379–423, 623–656, 1948.

[21] S. Vembu, S. Verdu, and Y. Steinberg, "The source-channel separation theorem revisited," IEEE Trans. Inform Theory, vol. 41, pp. 44–54, Jan. 1995

[22] MOCHNÁČ, J. – MARCHEVSKÝ, S.: Hybrid Concealment Mechanism, In: Acta Electrotechnica et Informatica, ISSN 1335-8243. Vol. 8, No. 1 (2008) pp. 11-15

Improvement of Sample Selection: A Cascade-Based Approach for Lesion Automatic Detection

Shofwatul 'Uyun[1,a], M. Didik R Wahyudi[1,c]
[1]Department of Informatics, Faculty of Science and Technology, State Islamic University Sunan Kalijaga, Yogyakarta, Indonesia

Lina Choridah[2,b]
[2]Department of Radiology, Faculty of Medicine, Gadjah Mada University, Yogyakarta, Indonesia

***Abstract*—Computer-Aided Detection (CADe) system has a significant role as a preventative effort in the early detection of breast cancer. There are some phases in developing the pattern recognition on the CADe system, including the availability of a large number of data, feature extraction, selection and use of features, and the selection of the appropriate classification method. Haar cascade classifier has been successfully developed to detect the faces in the multimedia image automatically and quickly. The success of the face detection system must not be separated from the availability of the training data in the large numbers. However, it is not easy to implement on a medical image because of some reasons, including its low quality, the very little gray-value differences, and the limited number of the patches for the examples of the positive data. Therefore, this research proposes an algorithm to overcome the limitation of the number of patches on the region of interest to detect whether the lesion exists or not on the mammogram images based on the Haar cascade classifier. This research uses the mammogram and ultrasonography images from the breast imaging of 60 probands and patients in the Clinic of Oncology, Yogyakarta. The testing of the CADe system is done by comparing the reading result of that system with the mammography reading result validated with the reading of the ultrasonography image by the Radiologist. The testing result of the k-fold cross validation demonstrates that the use of the algorithm for the multiplication of intersection rectangle may improve the system performance with accuracy, sensitivity, and specificity of 76%, 89%, and 63%, respectively.**

Keywords*—CADe system; haar; cascade classifier; mammogram

I. INTRODUCTION

Breast Cancer is the second most common forms of cancer in the world and has the first position as the most common form of cancer among women (World Cancer Research Fund International). One of eight women are under the risk of being diagnosed with breast cancer during their lifetime (WHO). The exact causes of the emergence of the cancer cells are not yet known. Therefore, a preventive action by performing the breast screening has a very significant role in reducing the number of victims [1]. The recommended imaging technology is mammography because it has more advantages than the other imaging. Furthermore, the Radiologist will give an assessment on the mammogram image. The interpretation screening on a mammogram is a challenge for the Radiologist because there are often some difficulties in finding the abnormal parts (the disorders) on the mammogram, which may happen because of many factors [2]. The researchers have developed some techniques to improve the Radiologist's performance, one of which is developing a system of computer-aided cancer detection on the mammogram commonly called the Computer-Aided Detection (CADe) System [3].

The CADe system is required to reduce the errors and to improve the Radiologist's ability in making the interpretation on mammography. Some CADe commercial systems have been used by the Radiologist widely. However, those cannot often function optimally yet (there is a positive phase case in the true positive case) [4]. The researchers are still trying to optimize the performance of the CADe system on using the mammogram image shown in the recent literature. In general, the CADe system developed by the researchers is divided into two classes with some kinds of variations of the class type, including: normal and abnormal [5]; mass and non-mass [6] and the finding of microcalcification and not [4][7].

There are some phases in developing the CADe system, including pre-treatment, the determination of RoI (Region of Interest), feature extraction, feature selection, and classification and the testing. The process of determining the RoI is under the direction and guidance of the Radiologist by conducting the cropping on the RoI to obtain some patches and the extraction and the feature selection. The use and selection of features are by viewing the purpose of the classification model development itself. Generally, previous researches, for the CADe system, use the mammography that is developed based on the three features on the mammogram image, those are the features of color, texture and shape. [8] use the color feature, while [9] use the shape feature and [6] combine the shape and texture on their research. Among both features, the last one is most widely used for mammography in previous researches [10][11][12][13] and [14]. The method of feature extraction widely used for the classification of RoI in statistic way is efficient and optimal and may describe the texture of the mammogram itself. Some previous researches that have developed the RoI classification system into two classes RoI (mass and a nonmass) mostly conduct the feature extraction statistically [6] and [12]. However, the determination of the part of RoI is still done manually; the system can only determine the class of RoI.

The next phase for the CADe system is a classification process that serves to classify the RoI predetermined by its feature similarity. Some classification methods commonly used in the field of pattern recognition are the artificial neural networks (ANN) [15], the support vector machine (SVM) [16], and the adaptive network-based fuzzy inference system [17]. [18] have successfully developed a classification method for

detecting the face image called the cascade classifier. After that, a lot of researches in the field of computer vision use and develop a cascade classifier for some purposes. Cascade classifier, previously drilled, has been proved to quickly detect the objects that have previously been drilled and successful in some multimedia images. Some problems often emerge in developing the CADe system with the supervised learning for the mammogram image. [7] states that there are two problems: the number of image pixels analysed in the large size and the vast areas of microcalcification (the positive area) that are not greater than the negative area commonly called the class imbalance; the limitations of RoI (positive) is also discussed by [19]. Besides the problem of the RoI limitation for positive samples (learning based cascade classifier), the rectangle of the cropping result of the Radiologist for the same area shows that there are inconsistencies on the part of the Radiologist in providing the markers on the RoI for the same area. Therefore, [19] propose three filtering strategies: sum, mean, and max. The testing result using the Jaccard coefficient has proved that a filter using the max operator has the best performance. The similar problem becomes the concern of [20] that is associated with the class imbalance, in which there are two proposed algorithm: the majority level uses the fuzzy membership function of Gaussian and alpha-cut types to reduce the size of data, while the minority class uses the diffusion membership function of mega- trend to generate some examples for the minority class. The two algorithms are proposed by [20] to cover the two classes that do not have the balanced amount. In this case, the data used is numerical and not in the image.

Therefore, this research develops the use of the cascade classifier concept with the aim to detect whether there are the lesions or not on the mammography. The mammography quality with multimedia images is very different; the mammography has a quality that is very far from the 'ideal' one, which results in difficulties for the Radiologist in identifying the abnormalities in the sought part. This research conducts the feature extraction on the mammography in wavelet by using the Haar feature and the integral image. The limited number of patches on the positive sample for the training data is one obstacle in developing the CADe system using the cascade classifier. Therefore, an algorithm is required to multiply the number of patches as a positive sample on the CADe system to improve its performance.

II. Research Method

The research is carried out in seven phases as follows.

A. Mammografic Image Acquisition and Ultrasonography

The image acquisition process is produced by the mammography and ultrasonography imaging technology from 60 Probands and patients in the Clinic of Oncology Kotabaru Yogyakarta. The breast imaging with mammography is conducted in two views: mediolateral-oblique (MLO) and craniocaudal (CC).

B. Annotations and Cropping Region of Interest (RoI)

After obtaining the image for each of these categories, the Radiologist gives an assessment on the mammography to provide an annotation on the part that is considered a disorder of the breast tissue. Furthermore, the given annotation ROI is classified into two categories: lesions and non-lesions. In determining an annotation on the mammogram image, the Radiologist also interprets the ultrasonography image to convince the truth of interpretations conducted visually on the mammography.

C. Multiplication of intersection (rectangle)

Based on the RoI cropping result by the Radiologist, there are some rectangles intersecting one another. Besides, the limited number of patches as the positive RoI samples (lesions) becomes an obstacle to the process of training using the Haar cascade classifier and may affect the recognition accuracy level. Therefore, to increase the RoI (the intersected rectangles), algorithm 1 and algorithm 2 are proposed, from now on called the algorithm for the multiplication of intersection rectangle. Algorithm 1 will work as long as rectangle 1 (X_1; Y_1; W_1; H_1) is not equal to rectangle 2 (X_2; Y_2; W_2; H_2).

```
Algorithm 1
WHILE (X1< X2) OR (Y1< Y2) OR (W1> W2) OR (H1> H2) //
as long as rectangle 1 ≠ rectangle 2
DO
     IF (X1< X2)
        X1 += 1;
     IF (Y1< Y2) AND (X1 = X2)
        Y1 += 1;
     IF (W1> W2)
        W1 -= 1;
     IF H1> H2 AND X1 = X2
        H1 -= 1;
     DRAW RECTANGLE( X1, Y1, W1, H1 );
ENDWHILE
```

```
Algorithm 2
IF ( R1 ∩ R2 ) //Check if the Rectangles intersect
between one another
BEGIN
     R2 = R1 ∩ R2 //newRect
ENDIF
     Algorithm 1; // use algorithm 1
```

D. Feature Extraction

Conducting the feature extraction of the disorders or RoI grouped into lesions and non-lesions by the Radiologists uses the Haar feature and the integral image to be able to represent the disorder feature on the mammography. The basis of classification on detecting the object lies in the use of some features of Haar-like. Some of these features are represented by the intensity values of the pixels by calculating the value difference between the light-colored pixel area and the dark one. Some of the Haar features can easily do the scaling either being raised or reduced in size to detect the image with various size (Viola and Jones, 2001). There are four basic types that can be used because it is easy to calculate the difference between the white area and the black one using the formula (1)

$$f_i = Sum(r_{i,white}) - Sum(r_{i,black}) \quad (1)$$

$$h_i(x) = \begin{cases} 1, \ jika\ f_i > threshold \\ -1, \ jika\ f_i < threshold \end{cases}$$

E. Feature Selection

Conducting the feature selection by modifying the procedure of Ada Boost further is stated in the points of discussion. The use of the appropriate features greatly affects

the accuracy of the system. The features are from the use of Haar-like feature and integral image. The next process is the selection of the best features that will be the basis for the classification in the next process. The algorithm used to select the best feature is the boosting algorithm. AdaBoost training algorithm is used to improve the performance of classification with the simple training algorithms. The feature selection process is by calculating the weight for each feature that is calculated using the formula (2) (Viola and Jones, 2001).

$$F = sign(w_1 h_1 + w_2 h_2 + \cdots + w_n h_{n)} \quad (2)$$

$$in\ which, h_i(x) = \begin{cases} 1, & if\ f_i > \theta_i \\ -1, & if\ f_i < \theta_i \end{cases}$$

F. Classification of lesions and non-lesions

Conducting the classification between lesions and non-lesions using a cascade classifier further is stated in the points of discussion. The performance scheme of the cascade classifier conducts the classification of RoI based on the features gradually used as shown in Figure 2. The calculation of the classification result has the greatest weight calculated based on the formula (3). The CADe system for the purpose of detecting the RoI consists of two classes: objects and undesired object. Viola and Jones (2001) have developed and tested the classification algorithm to detect whether the frame captured from the image is the form of a face object or not.

strong classifier =

$$(\alpha_1 h_1 + \alpha_2 h_2) + (\alpha_T h_T) < Threshold \quad (3)$$

G. Performance evaluation of CADe system

The testing scheme on the proposed phases of the algorithm in CADe system consists of two types, first, to test the results of training on the training data using the k-fold cross validation; second, to calculate the level of accuracy, sensitivity and specificity of the CADe systems with the results of the assessment and observations of the Radiologist on the mammogram and ultrasonography images. The assessment result of the Radiologist may become the preference in assessing the results of detecting the CADe system. A general description of each phase is shown in Figure 1.

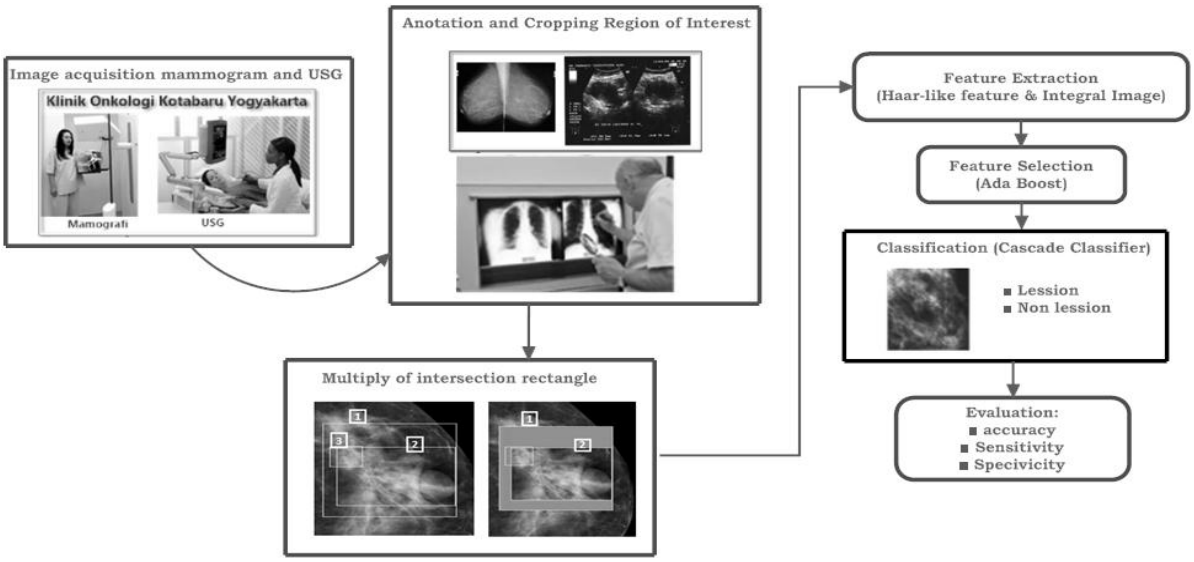


Fig. 1. General description of the research phase

III. Results and Analysis

The first phase conducted is the data acquisition of the mammography and ultrasonography as the result of breast imaging of the Probands and Patients in the Clinic of Oncology. After that, the annotation is conducted on the mammogram image on any part considered as the disorder by the Radiologist with the assessment validation using the ultrasonography image. From the cropping result conducted by the Radiologist, it may be inferred that there is only one rectangle/RoI on a single image, but on some mammogram images, it is found that the Radiologists give the annotations more than once at the close/intersected locations, and some are at different locations. Figure 2 shows examples of the mammogram image with annotation by the Radiologist more than one rectangle is shown in Fig. 2. In general, the results of annotations on more than one RoI may be grouped into two types: between a rectangle with the other one is apart from each other or not intersect to each other, while it is also found that a rectangle and the other one is intersecting to each other.

After the cropping of the RoI (lesion) by the Radiologist, the CADe system is also able to show the results of the cropping. If the Radiologist finds lesions in more than one location as shown in Fig. 3, it is necessary to give the special treatment so that the area as the intersected result of the both rectangles may have the pixel shifting gradually.

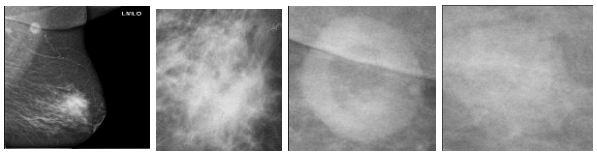


Fig. 2. Example of early image to do the cropping (a), on that image there are the lesions in 3 different locations (b,c,d)

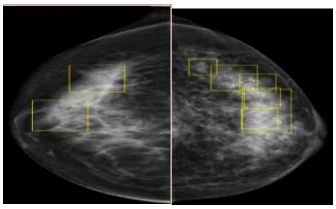


Fig. 3. The annotation result of the Radiologist: there is no intersection (left), and there is intersection (right)

Examples of the result of the assessment and the administration of annotations on the mammography with the view of MLO on the left and the right side with the RoI cropping results show that there are separated rectangles (not intersecting) as illustrated in Figure 4 and 5. Figure 6 shows the view of CC on the left and the right sides which rectangles are not intersecting. Figure 7 illustrates the example of an assessment result of mammography with the view of MLO on the left and the right sides and the RoI cropping result showing that there are intersecting rectangles. Figure 8 and 9 show the view of CC on the left and the right sides which rectangles are intersecting.

Therefore, this research proposes an algorithm for the multiplication of intersection rectangle to increase the intersecting rectangle patch by conducting the pixel shifting. Both algorithms are used before the RoI feature extraction. There are two algorithms, the first algorithm is used when there are two annotations (rectangles) that have different sizes, one is contained in the other. Figure 10 shows that rectangle 2 is in rectangle 1.

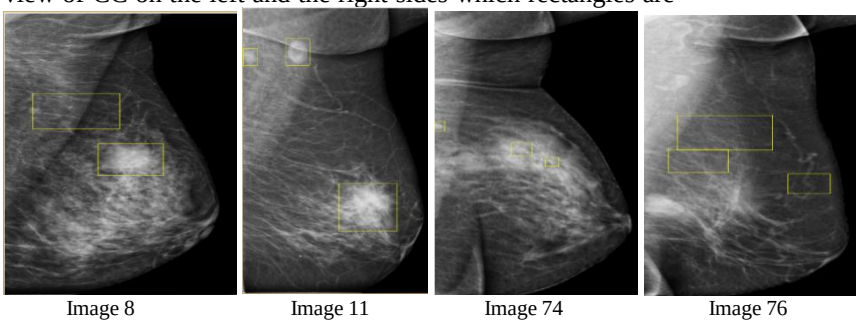


Image 8 Image 11 Image 74 Image 76

Fig. 4. The RoI *cropping* result with the view of MLO on the left side

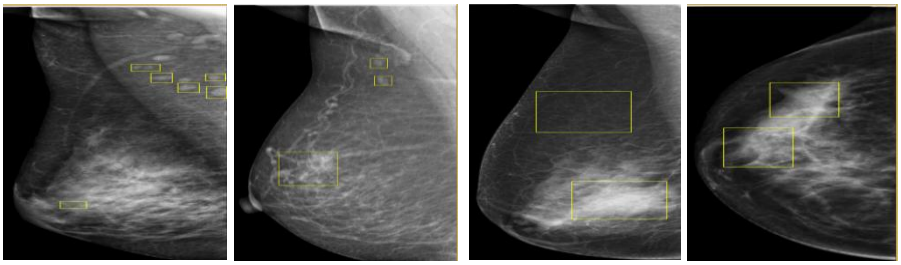


Image 20 Image 52 Image 38 Image 50

Fig. 5. The RoI cropping the view of MLO on the right side

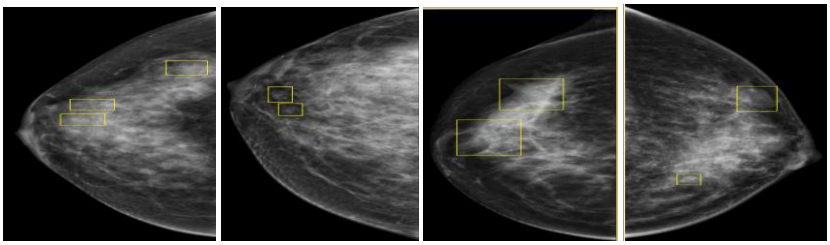


Image 19 Image 80 Image 50 Image 89

Fig. 6. The RoI cropping result with the view of CC on the right side (image 19, 80 and 50) and the left side (image 89)

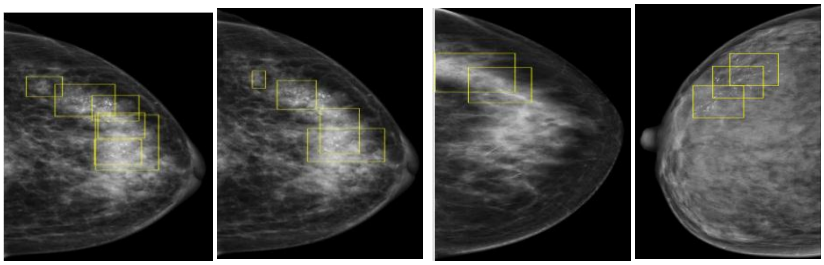


Image 15 Image 69 Image 48 Image 67

Fig. 7. The RoI cropping result is intersecting the view of CC on the left side (image 15, 69 and 48 and the right side (image 67)

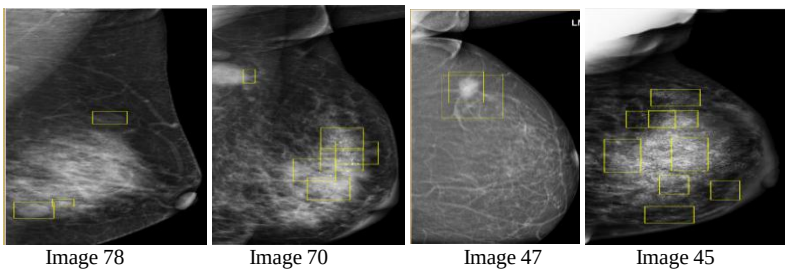


Image 78 Image 70 Image 47 Image 45

Fig. 8. The RoI cropping results with the view of MLO on the left side

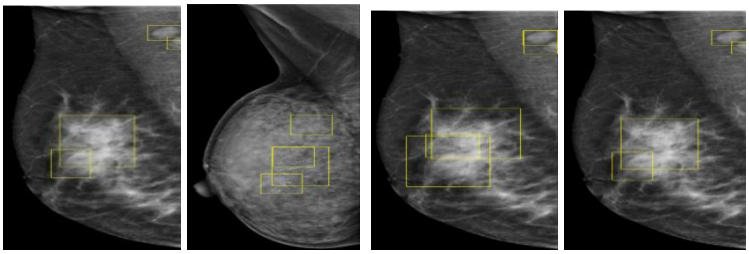


Image 27 Image 68 Image 51 Image 27

Fig. 9. The RoI cropping results with the view of MLO on the right side

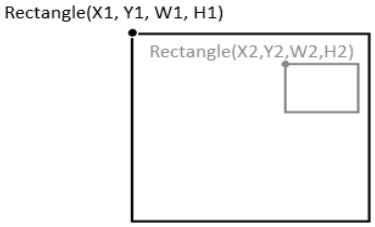


Fig. 10. Illustration of one rectangle is inside the other one

The aim is to make more RoI, or to make the difference between the two rectangles by making a positive sample from rectangle 1 to rectangle 2. The second condition, if there are two annotations (rectangles) one side of which (the line) intersects (tangent) with the side or the line of the other rectangle. The illustration can be shown as Fig. 11, where rectangle 1 intersects or has tangent with rectangle 2 (R1 ∩ R2), so the side that will have the pixel shifting (duplication) is the area included in the area of R1 and R2.

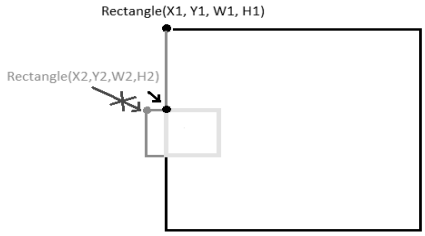
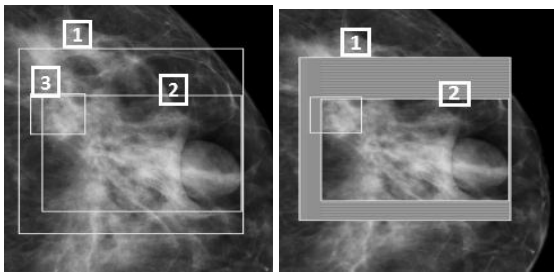


Fig. 11. Illustration of a rectangle intersecting or having tangent with another rectangle

The second algorithm is to create more rectangles. The multiplied area is the one resulting from the intersection of both rectangles. The workings of the algorithm proposed are as follows: Previously some processing on the image of the cropping results (RoI or abnormality or lesion) are conducted by the Radiologist. The image that will have the multiplication of RoI certainly is the mammogram image that has been given

more than one annotation by the Radiologist and is intersecting, which means that the Radiologist conducts the RoI cropping on one image with more than one rectangle. Furthermore, the RoI will have the image multiplication by processing the pixels sequentially and simultaneously starting from the largest RoI to the other intersection of the RoI. Figure 12 illustrates the RoI multiplication process with rectangle 1 and 2. The area on rectangle 2 will have the pixel shifting to rectangle 1 to make (X_1; Y_1; W_1; H_1) equal to (X_2; Y_2; W_2; H_2). The algorithm for the multiplication of intersection rectangle from the two rectangles will create a lot of rectangles with different positions and grayscale values at each pixel.



(a) (b)

Fig. 12. (a) The cropping result by the Radiologist (b) The multiplication result of RoI 1 and RoI 2.

There are some pre-treatments and trials before the training, including the process of creating the sample sized in 40 x 40 pixels is conducted before the image sample of the cropping result of the Radiologist whether there is an intersection of one another, or there is no intersection. Based on the experiment result with a few number of stages, finally, it obtains the best weight of the training experiment that is shown in the cascade classifier with eight stages. The bigger the stage value is, the more accurate the detection will be. However, the number of stages also should consider the number of the positive samples. In this research the determination of the number of stages is conducted by the trial error and the most optimal number of stages obtained is 8 stages. The maximum value of error (maximum false alarm) that may be received is 30%, which means that 30% of the negative sample may be detected as the positive ones. The greater this value is, the more inaccurate the detection will be. However, it cannot replace it with a value of 0% because the training process will not finish. There is an experiment on the neighbourhood value to obtain the best result in this research, and the best result is at a value between 30 and 40. The higher the neighborhood value is, the more accurate the detection will be, but if it is overdone it will not be able to detect anything on the mammogram images. While for the minimum scale used to find the pixels that is detectable, this research uses a scale of 170x170 after doing the trial and error to see data from the cropping result of the Radiologist on the file 'info.txt'. The larger the scale is, the less time it takes to process an image. The example of the result of the detection system using the algorithm for the intersecting rectangles to increase the RoI in the image, in which there is more than one intersecting rectangle, is the mammogram image as the annotation results by Radiologist, and the annotation resulted from the CADe system as shown in Figure 13.

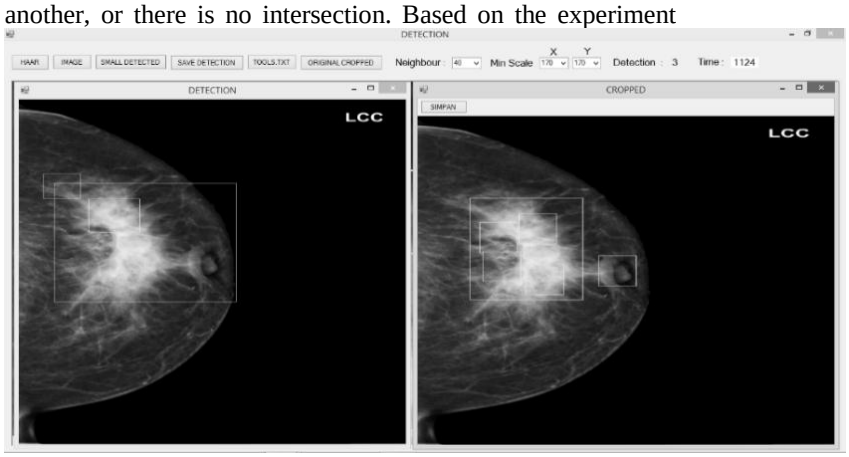


Fig. 13. The annotation is given by the Radiologist (right), the annotation as the detection result of the CADe system (left)

The testing result conducted in this research shows a significant difference between the CADe system that uses the algorithm for the multiplication of intersection rectangle with the one without using the algorithm with an accuracy of 44% and 76%, a sensitivity of 41% and 89% and a specificity of 48% and 63% as shown in Figure 14. The use of the algorithm for the multiplication of intersection rectangle may improve the accuracy of the CADe system on the mammogram image that has a number of positive rectangles for the positive class (there are lesions) that is very little, whereas the number of the training data largely determines the success of a system in the training process.

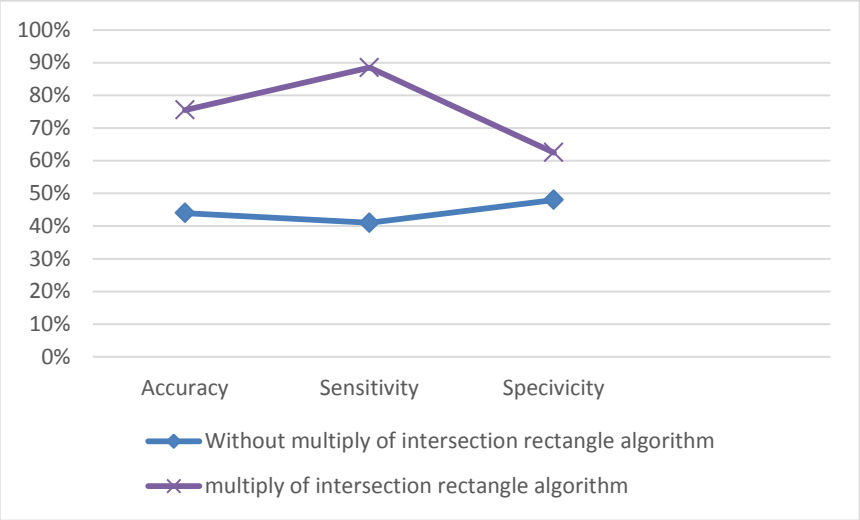


Fig. 14. The testing result with and without the multiplication of intersection rectangle algorithm

IV. Conclusion

The proposed algorithm for the improvement of selecting rectangle that aims to multiply the patch of RoI for a positive sample on the CADe system is proved to be able to improve the system performance. The testing result using the k-fold cross validation shows that the automatic detection of lesions using an approach based on the cascade classifier with the algorithm for the improvement selecting rectangle may detect the RoI much better with the level of accuracy, sensitivity, and specificity of 76%, 89%, and 63%, respectively. Meanwhile, without using the algorithm for the improvement of selecting rectangle it only has the level of accuracy, sensitivity and specificity of 44%, 41%, and 48%, respectively. However, it is required to develop other methods such as using the fuzzy systems to cope with the process of training on the CADe system with the very limited training data.

Acknowledgment

This research is supported by the Directorate Higher Education Islamic, the Directorate Geeral Islamic Education, the Minister of Religious Affairs of the Republic of Indonesia.

References

[1] A. Akhsan & T. Aryandono. Prognostic factors of locally advanced breast cancer patients receiving neoadjuvant and adjuvant chemotherapy. *Asian Pac J Cancer Prev*, 2010, Vol. 11, pp. 759-761.

[2] J.S. Drukteinis, E.C. Gombos, S. Raza, S.A. Chikarmane, A. Swami & R. L. Birdwell. MR imaging assessment of the breast after breast conservation therapy: distinguishing benign from malignant lesions. *Radiographics*, 2012, Vol. 32, No. 1, pp. 219-234.

[3] L.H. Eadie, P. Taylor & A. P. Gibson. A systematic review of computer-assisted diagnosis in diagnostic cancer imaging. *European journal of radiology*, 2012, Vo. 81, No. 1, pp. e70-e76.

[4] C. Marrocco, M. Molinara, C. D'Elia & F. Tortorella. A computer-aided detection system for clustered microcalcifications. *Artificial intelligence in medicine*, 2010, Vol. 50, No. 1, pp. 23-32.

[5] C. C. Jen & S. S. Yu. Automatic detection of abnormal mammograms in mammographic images. *Expert Systems with Applications*, 2015, Vol. 42, No. 6, pp. 3048-3055.

[6] M. L. de Oliveira, G. B. Junior, A. C. Silva, A. C. de Paiva & M. Gattass. Detection of masses in digital mammograms using K-means and support vector machine. *Electronic Letters on Computer Vision and Image Analysis*, 2009, Vol. 8, No. 2, pp. 39-50.

[7] A. Bria, N. Karssemeijer & F. Tortorella. Learning from unbalanced data: a cascade-based approach for detecting clustered microcalcifications, *Medical image analysis*, 2014, Vol. 18, No. 2, pp. 241-252.

[8] M. Langarizadeh & Mahmud. R. Breast Density Classification Using Histogram-Based Features. *Iranian Journal of Medical Informatics*, 2012, Vol. 1, No. 1, pp. 1-5.

[9] A. Vadivel & B. Surendiran. A fuzzy rule based approach for characterization of mammogram masses into BI-RADS shape categories. Computers in Biologyand Medicine, 2013, Vol. 43, No. 4, pp. 259-267.

[10] A. M. Khuzi, R. Besar, W. W. Zaki & N. N. Ahmad. Identification of masses in digital mammogram using gray level co-occurrence matrices. *Biomedical Imaging and Intervention Journal*, 2009, Vol. 5, No. 3.

[11] T. S. Subashini, V. Ramalingam & S. Palanivel. Automated assessment of breast tissue density in digital mammograms. *Computer Vision and Image Understanding*, 2010, Vol. 114, No. 1, pp. 33-43.

[12] I. K. Maitra, S. Nag & S. K. Bandyopadhyay. Identification of abnormal masses in digital mammography images. *International Journal of Computer Graphics*, 2011, Vol. 2, No. 1, pp. 17-29.

[13] M. A. Al Mutaz, S. Dress and N, Zaki. Detection of Masses in Digital Mammogram Using Second Order Statistics and Artificial Neural Network. *International Journal of Computer Science and Information Technology (IJCSIT)*, 2011, Vol. 3, pp. 176-186.

[14] S. 'Uyun, S. Hartati, A. Harjoko & Subanar, Selection mammogram texture descriptors based on statistics properties backpropagation structure, *International Journal of Computer Science and Information Security (IJCSIS)*, 2013, Vol. 11, No. 5, pp. 1-5.

[15] J. Jiang, P. Trundle & J. Ren. Medical image analysis with artificial neural networks. *Computerized Medical Imaging and Graphics*, 2010, Vol. 34, No. 8, pp. 617-631.

[16] S. W. Borges, D. E. Moraes, S. A. Correa, P. A. Cardoso & M. Gattass. Detection of masses in mammogram images using CNN, geostatistic functions and SVM. *Computers in Biology and Medicine*, 2011, Vol. 41, No. 8, pp. 653-664.

[17] F. C. Fernandes, L. M. Brasil, J. M. Lamas & R. Guadagnin. Breast cancer image assessment using an adaptative network-based fuzzy inference system. *Pattern Recognition and Image Analysis*, 2010, Vol. 20, No. 2, pp. 192-200.

[18] P. Viola, & M, Jones. Rapid object detection using a boosted cascade of simple features. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001, Proceedings of the 2001 IEEE Computer Society Conference*, pp. I-511.

[19] E. Cheng, H. Ling, P. R. Bakic, A. D. Maidment & V. Megalooikonomou. Automatic detection of regions of interest in mammographic images. In *SPIE Medical Imaging*, 2011, pp. 79623J-79623J. International Society for Optics and Photonics.

[20] D. C. Li, C. W. Liu, & S. C. Hu. A learning method for the class imbalance problem with medical data sets. *Computers in biology and medicine*, Vol. 40, No. 5, 2010, pp. 509-518.

Security Risk Scoring Incorporating Computers' Environment

Eli Weintraub
Department of Industrial Engineering and Management
Afeka Tel Aviv Academic College of Engineering
Tel Aviv, Israel

***Abstract*—A framework of a Continuous Monitoring System (CMS) is presented, having new improved capabilities. The system uses the actual real-time configuration of the system and environment characterized by a Configuration Management Data Base (CMDB) which includes detailed information of organizational database contents, security and privacy specifications. The Common Vulnerability Scoring Systems' (CVSS) algorithm produces risk scores incorporating information from the CMDB. By using the real updated environmental characteristics the system enables achieving accurate scores compared to existing practices. Framework presentation includes systems' design and an illustration of scoring computations.**

***Keywords*—CVSS; Security; Risk Management; Configuration Management; CMDB; Continuous Monitoring System; Vulnerability**

I. INTRODUCTION

Computing systems are subject to cyber-attacks which may cause damage to organizational and personal data, software and hardware [1]. Vulnerabilities are weaknesses or exposures stemming from bugs that are potential causes of security failures: loss of confidentiality, integrity or availability. Attackers are exploiting target systems making use of software vulnerabilities existing in systems' components. Attacks on users' computers cause them damages of many kinds such as stealing organizations' information or changing customers' data. The quality of knowledge an organization has of systems' weaknesses influences heavily on the success of organizations' defense activities. This work focuses on gaining accurate knowledge of computers' vulnerabilities, thus enabling improved organizational risk measures, which enable the development of efficient mitigation activities to defend computers from hostile attackers. Reference [2] states that Stuxnet worm included a process of checking hardware models and configuration details, and also downloads code from the controller to check if it was the "right" program before launching an attack. Both, attackers and security managers are interested in gaining accurate and detailed information of the target system, but for the opposite reasons. Organizations make decisions on actions they have to take, to limit their risks according to the amount of potential damage and vulnerability characteristics [3].

Risk has many definitions in research publications. We use the definition of [4]: "An event where the outcome in uncertain". According to this definition, this work is aimed at lessening risk uncertainty. The proposed model focuses on gaining accurate real-time information on systems' configuration, components and the environment which the system interfaces.

Several software products are aimed at defending computers from cyber attackers. Antivirus software, antispyware and firewalls are examples to some of these tools which usually perform periodic assessment of the target computer by comparing computers' software to the known published vulnerabilities. Antivirus software and firewalls use hash signatures to identify attacks on assets. In cases the defense software recognizes a hash signature in computers' software it reports to the computers' owner the existence of the attacking software. Those tools are aimed at identifying known threats but not new unpublished threats. Continuous Monitoring Systems (CMS) monitor computer systems in a near real time process aimed at detecting vulnerabilities and notifying organizations' security managers. Contemporary systems use vulnerabilities databases which are continually updated as new vulnerabilities are identified and a scoring algorithm which predicts potential business losses. CMS's are essential tools for limiting the time-frames organizations are exposed to risks, thus enabling organizations taking measures for risk mitigation.

Computers are defenseless to known threats as long as no patch exits to protect against the vulnerability. Preparing such a patch needs efforts of design, programming and testing activities that may last weeks or months. Only after the software vendor prepares a working patch, computers' owner has to load it to the operational system, which is the moment the computer ceases to be vulnerable. Loading patches to computer systems are usually performed as a periodical process, not continuously. The reason for this is avoiding too many interrupts required for uploading and activating the patch on the organizational computers. Other software tools usually use heuristic algorithms which are programmed to detect irregular suspicious activities of the software running on the computers. Those tools are programmed to detect deviations from the "normal" profile of computers' activities. In today's environment of zero-day exploits, conventional systems updating for security vulnerabilities has become a cumbersome process. There is an urgent need for a solution that can rapidly evaluate system vulnerabilities' potential damages and immediately fix them. [5].

Security Continuous Monitoring (SCM) tools are operating techniques for monitoring, detecting and alerting of security

threats on a regular basis. After identifying these risks, tools evaluate the potential impacts on the organization, sometimes suggesting risk mitigation activities to the organization to support organizational risk management decisions. Reference [6] states that SCM systems which are running on computers, continuously try to detect systems' vulnerabilities aim to close the gap between the zero-day of identifying the vulnerability, until the moment computers' owner loads the corresponding patch fixing the vulnerability. The time frame may be considerably long.

In this paper, we describe the mechanisms of a new SCM framework of a system that will produce better detection and prevention than current known SCM systems. Frameworks' capabilities makes use of two main resources: knowledge concerning the specific computers' organizational environment of the target system, and a prediction algorithm which runs continuously evaluating risk scores.

The rest of the paper includes the following sections: In section 2 a description of current known existing solutions. In section 3 a presentation of the proposed framework including systems' architecture. In section 4 a description of the risk scoring algorithm which computes risk scores. In section 5 results. In section 6 conclusions and future research directions.

II. Existing Solutions

SCM systems are using external vulnerabilities databases for evaluation of the target computers' risk. There are several owners of vulnerability databases [5]: The Sans Internet Storm Center services and The National Vulnerability Database (NVD). Vulnerability Identification Systems (VIS) aimed to identify vulnerabilities according to three categories: code, design, and architecture. Examples for VIS are the Common Vulnerabilities and Exposures (CVE), and The Common Weakness Enumeration (CWE).

In this work, we shall use NVD vulnerabilities database as an illustration of the proposed model.

Risk evaluation uses scoring systems for assessing the impacts of vulnerabilities on the organization. The Common Vulnerability Scoring System (CVSS) is a framework that enables user organizations to receive IT vulnerabilities characteristics [1].

CVSS uses three groups of parameters to score potential risks: basic parameters, temporal parameters, and environmental parameters. Each group is represented by several score compound parameters ordered as a vector, used to compute the score. Basic parameters represent the intrinsic specifications of the vulnerability. Temporal parameters represent the specifications of vulnerabilities that might change over time due to technical changes. Environmental parameters represent the specifications of vulnerabilities derived from the local IT environment used by the organization. CVSS enables omitting the environmental metrics from score calculations in cases they have no effect on the score and in cases the users' does not specify the detailed description of environments' structure, and it's components.

CVSS is a common framework for characterizing vulnerabilities and predicting risks, used by IT managers, risk managers, researchers and IT vendors, for several aspects of risk management.

CVSS is an open framework which enables managers to deal with organizations' risks and make decisions based on facts rather than evaluations. Organizations adopting CVSS framework may gain the following benefits:

- A standard scale for characterizing vulnerabilities and risks scoring.
- Normalizing vulnerabilities according to specific IT platforms. The computed scores enable users getting rational decisions in correlation to vulnerability risks.
- CVSS uses an open framework. Organizations can see the characteristics of vulnerabilities and the logical process of the scoring evaluation.
- Environmental scores. Organizations using the environmental parameters may benefit by considering changes in its IT configuration according to predicted risk scores. The specification of systems' configuration is defined using only high-level parameters such as system.

There are few other vulnerability scoring systems besides CVSS differing by what they measure. CERT/CC emphases internet infrastructure risks. SANS vulnerability system considers users' IT configuration and uses default parameter definitions. Microsoft's scoring system emphasizes attack vectors and the impacts of the vulnerability. Using CVSS scoring system, basic and temporal parameters are specified and published by products' vendors who have the best knowledge of their product. Users make estimates of environmental parameters since they have the best knowledge of their environments and vulnerability business impacts.

This paper focuses mainly on environmental metrics.

An exploit of a vulnerable component may cause major or minor damage to a system, depending on the technological and business characteristics of the configuration and of systems' users. CVSS environmental parameters specify the characteristics of vulnerabilities in correlation with systems' components. Environmental parameters are of three groups:

- Collateral Damage Potential (CDP).

A group of parameters which measure the economic potential loss caused by an exploit of a vulnerable component.

- Target Distribution (TD).

Parameters indicating the percentage of vulnerable components in users' environment. A large proportion might have more impacts on organizational potential damages.

- Security Requirements (CR, IR, AR).

Parameters indicating the security importance measures in users' organization. This group of parameters include parameters indicating the confidentiality (CR), integrity (IR), and availability (AR) of the vulnerable component. Higher security requirements might cause more security damages, thus causing more business losses.

Organizations' users should categorize all IT assets according to security requirement measures. Doing so raises the possibility to predict the organizational losses. Federal Information Processing Standards (FIPS) requirements demand implementation of a categorization [6] but does not require using any particular scale, thus risk comparison of users' systems is difficult.

III. The Proposed Framework

Federal organizations are moving from periodic to continuous monitoring implementing SCM's which will improve national cyber security posture [7]. The proposed framework includes two advantages over current practices. First, the environmental parameters are based on the components of the system as updated in the systems' Configuration Management Data Base (CMDB) [8]. This capability enables basing the scoring models to perform predictions of organizational damages on real IT environment rather than on user's evaluations. According to [9] it is impossible for organizations to make precise estimates of the economic losses caused by an attack without having full knowledge of users' IT environment. Reference [10] states that organizations should monitor their network continually, and analyze available vulnerabilities to provide the necessary security levels. Secondly, the information of the environmental components is described in this research is in resolution of data items rather than entire systems, thus enabling focused information in relevance to each data item. The proposed CMS examines a database of published asset vulnerabilities, compares in real time computers' assets for existing exposures and calculates computers' potential losses. Loss evaluation algorithm considers vulnerabilities at the moment they are identified even before software vendors prepares patches and before the organization loads the patches to the operational environment. The CMS's proposed architecture is described in fig. 1. Following, a description of systems' components and processes.

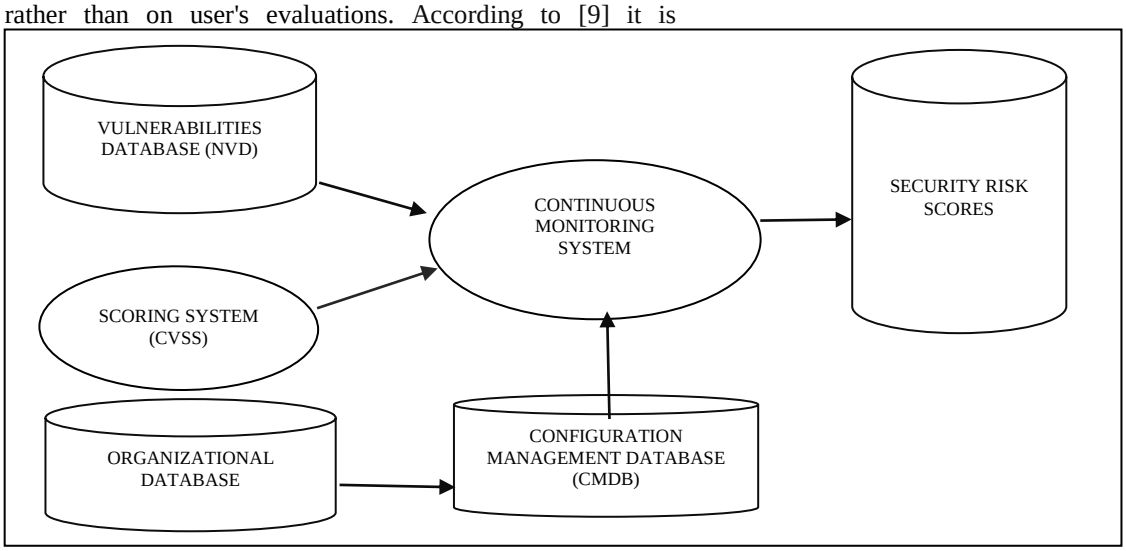


Fig. 1. Continuous Monitoring System architecture

Vulnerabilities Database includes all known vulnerabilities and their specification as published by Database owners. Examples of vulnerability specifications used by NVD are vulnerability category, vendor name, product name, published start and end dates, vulnerability update dates, vulnerability severity, access vector, and access complexity [6]. Scoring system (CVSS) is the algorithm this research uses for illustration of the proposed model, which computes security risk scores according to parameter groups: basic, temporal and environmental. As stated above there are other known scoring algorithms, some of them for public use other commercial. CMDB is a database which includes all hardware and software components of the target system. According to the proposed model the CMDB manages high resolution information of the organizational database. A frequently running process populates the CMDB, through reading the organizations' database contents. The CMDB contains information of all modules, components, and relationships among the components. The design of the CMDB includes software in the resolution of programs, services and parameters. The design of data is in the resolution of database, tables and data items. Input/output design includes screen-names and output messages in the resolution of data items. The target system might be one computer or a group of organizations' computers. The CMDB also includes all the components which interface with the system directly or indirectly up to external and end-users' interfaces. The CMDB also includes the security requirements (CR, IR, AR) of each component in the resolution of data items. Users define security requirement measures according to business security levels' definitions. The CMDB includes also all interfaces among components. For each interface are indicated the direction of data transfer between the components and probability of connections' occurrences.

The system runs continuously and starts computing losses in two cases: first is whenever a software vendor publishes in NVD a new vulnerability or a change in vulnerability status, second, is whenever systems' owner makes a change in a systems' component or the systems' environment or interface. The system performs evaluations of damage potential using NVD, CVSS, and CMDB. In each case the system identifies a new vulnerable component according to NVD, the system evaluates the new damage potential score and informs the organization. The system writes the computed risk scores on the risk scores database for risk management organizational usage.

In this work, the CMDB includes only a subgroup of all kinds of information of the target computer: high-resolution knowledge of the data entities included in the organizational database, and security requirements of the data entities of the organizational database.

IV. The Scoring Algorithm

CVSS's framework makes use of three kinds of parameters. Vendors who have the best knowledge of their products make estimates of the basic and temporal parameters. Users, who have the best knowledge of their IT configuration, interfaces, and vulnerabilities' business impacts specify the environmental parameters. This work deals with the environmental parameters. According to [6], in many organizations IT resources are labeled with criticality ratings based on network location, business function, and the potential for loss of revenue or life. For example, the U.S. government assigns every unclassified IT asset to a system which is a grouping of assets. Every governmental agency has to categorize systems according to "potential impact" ratings to show the potential impact of systems' compromises on the organization. The categorization should relate to three security objectives: confidentiality, integrity, and availability. Thus, every IT asset in the U.S. government has a potential impact rating of low, moderate, or high with respect to the three security objectives. The Federal Information Processing Standards (FIPS) 199.5 describes this security rating system [11]. CVSS follows this general model of FIPS 199 but does not require organizations to use any particular system for assigning the low, medium, and high impact ratings. Reference [12] states that organizations should define the specifications of security risks of their environment, but does not outline the ways organizations have to specify that information. The Department of State (State) has implemented an application called iPost and a risk scoring program that is intended to provide continuous monitoring capabilities of information security risk to elements of its information technology (IT) infrastructure. According to [13] the iPOST scoring model does not refine the base scores of CVSS to reflect the unique characteristics of its environment. Instead, it applied a mathematical formula to the base scores to provide greater separation between the scores for higher-risk vulnerabilities and the scores for lower-risk vulnerabilities. This work is targeted to close this gap.

The CMDB defined in this work handles the environmental information included in the organizational database in the highest resolution of data items to be able to assign scoring measures to all entities: data items, database tables, and the entire organizational database. The CMDB manages five kinds of environmental information for every data item in the organizational database. Table I describes the parameters defined for each data item. The values for each parameter are based on [11] definitions which categorize parameters according to security information types which are the following: privacy, medical, proprietary, financial, investigative, contractor sensitive, and security management. Reference [11] states that the potential impact is low if the loss of confidentiality, integrity, or availability could be expected to have a limited adverse effect on organizational operations, organizational assets, or individuals. The potential impact is moderate if the loss of confidentiality, integrity, or availability could be expected to have a serious adverse effect on organizational operations, organizational assets, or individuals. The potential impact is high if the loss of confidentiality, integrity, or availability could be expected to have a severe or catastrophic adverse effect on organizational operations, organizational assets, or individuals.

TABLE I. Configuration Management Table Columns

Column ID	Column Name	Column Description	Values (*)
CDP	Collateral	This metric measures the	N, L,

	Damage Potential	potential for loss of life or physical assets through damage or theft of property. The metric may also measure an economic loss of productivity or revenue.	M, MH, H
TD	Target Distribution	This metric measures the proportion of vulnerable systems. It is an environment-specific indicator to approximate the percentage of systems that could be affected by the vulnerability.	N,L,M, H
CR	Confidentiality Requirement	The importance of the affected IT asset to a user's organization, measured regarding confidentiality. "Preserving authorized restrictions on information access and disclosure, including means for protecting personal privacy and proprietary information…"	L,M,H
IR	Integrity Requirement	"Guarding against improper information modification or destruction, and includes ensuring information non-repudiation and authenticity…".	L,M,H
AR	Availability Requirement	"Ensuring timely and reliable access to and use of information…".	L,M,H

(*) N=none, L=low, LM=low medium, M=medium, MH=medium high, H=high

To present the proposed rating system, an introduction of a use case database follows, which will help explain and demonstrate the computations. Tables II and III describe the organizational database and contents. The use case consists of a bank accounts database containing two tables: customers table which contains customers' details and accounts table containing all the details of the loan accounts given to the customers. Each customer may have several loans. Customers' details include customer identification number, customers' name, customers' address, customers' telephone number, customers' salary, and customers' total amount of bank deposits. The details which are categorized as private information items according to FIPS categorization [11] are customer identification and customer name. Items categorized as financial are customers' salary and customers' total deposits. Customers' loan accounts details include account number for identification, account's balance, accounts' date which indicates accounts' opening date, accounts number of months until end loan, and accounts' customer identification. Table IV describes the structure and contents of the CMDB. Each row in the CMDB describes one data item and environmental parameters. Data items of the entire database are in sequential order.

TABLE II. CUSTOMERS TABLE COLUMNS

Cus ID	Cus name	Cus address	Cus telephone	Cus salary	Cus deposit
100	John	Washington	2031234	70000	200000
200	Dan	New York	2025688	90000	50000
300	Scott	Philadelphia	7059876	20000	40000
360	Ben	Boston	7027654	40000	70000
450	Gary	Yale	80175324	60000	8000

TABLE III. ACCOUNTS TABLE COLUMNS

Acc number	Acc balance	Acc date	Acc months	Acc cus ID
3	1000	16.07.15	36	100
5	2400	12.10.14	12	100
8	20	10.8.15	48	300
11	599	19.07.10	100	360
16	50	30.03.13	66	100
23	2000	18.07.12	8	450

TABLE IV. CONFIGURATION MANAGEMENT DATABASE COLUMNS

Item no'	Item name	CDP	TD	CR	IR	AR	Env' Score
1	Cus ID	H	H	H	H	H	5
2	Cus name	MH	H	H	M	M	4
3	Cus address	L	L	M	L	L	0.3
4	Cus telephone	LM	M	H	L	L	2.3
5	Cus salary	LM	L	M	L	L	0.8
6	Cus deposits	MH	L	M	L	L	1
7	Acc number	MH	M	H	H	H	3
8	Acc balance	MH	L	M	L	L	1
9	Acc date	L	L	L	L	L	0.3
10	Acc months	L	L	L	L	L	0.3
11	Acc cus ID	LM	L	L	H	L	0.8

Following, an illustration of data items information security types categorization rational. Security type categorization leads to rational estimations of the environmental parameters. In this use case the security categorization is as follows: The account number data item is private information since it is an important data which hackers often use for identifying customers' information. The account balance data item is clearly financial information. Data items customer ID and account ID are integrity information since their function in the database is primary and foreign keys used for records identification and for keeping on integrity constraints.

The CM Database Columns Table includes all the information of the data items. Table IV includes values of five environmental parameters for each data item in the database: CDP, TD, CR, IR, and AR. Each parameter value indicates the security risk environmental specifications for data items. For illustration, data item customer salary is specified as low-medium CDP, TD is specified as low, confidentiality requirement is medium since the data item is categorizes as financial, IR is specified as low, and AR is specified as low. Following, calculations of the environmental scores using CVSS calculator [14]. The calculator computes the environmental score producing a real number ranging in the interval [0, 5], whereas 0 indicates the minimal impact of the environment on organizational risk and 5 indicates the maximal risk. Using configuration management parameters values and CVSS calculator, the environmental scores are computed and presented in the rightmost column of Table IV. For each data item the calculator uses the environmental parameter values to compute the environmental risk score. For example for the data item customer name the evaluated score is 4 (out of 5), which is an indication of the high impact on organizations' risk. This score is a computed result of a medium-high value of the CDP, high values of TD and CR, and medium values of IR and AR. The rational for the high security score of customer name item follows the high and medium security environmental parameter values. Following, computations of the environmental scores of each table. Computations of all data item environmental scores are presented in the CM table.

Following a formalization of the environmental risk scores.

The symbol i denotes table number.

Column j (i) indicates column number j in table i.

SCORE-TABLE (i) indicates the total environmental risk score of table i. Evaluation of tables' environmental score involves computation of the maximal risk scores of all table columns. The underlying assumption is that in case the organization is facing a damage to the table, all columns cannot be used. Table scores evaluation formula follows in fig. 2.

$$\text{SCORE-TABLE}(i) = \max\left(\text{SCORE}(\text{COLUMN}(j(i)))\right) \quad \text{For all } j \in \text{Table}(i)$$

Fig. 2. Table scores formula

As an illustration SCORE-TABLE (1) indicates the maximal risk score of Customers table which are: 5, 4, 0.3, 2.3, 0.8, 1. Thus, tables' score is 5.

Calculating the environmental risk score of accounts table number 2 involves computing the maximal scores of the columns 3, 1, 0.3, 0.3, and 0.8 which yields a score of 3.

This result represents a higher risk in cases of damage to customers table (4) than to accounts table (3), which may lead to organizational decisions of implementing improved mitigation strategies for the defense of customers table. Assuming that the organization manages several tables in one database, the environmental score of the database will be the maximal score of all tables' scores. This score indicates the environmental impact on the organization in cases of damage to the database as one integral entity. Such cases are relevant when no possibility exists or no knowledge exists concerning which parts of the database were damaged.

Assuming an occurrence of a compromise event to the database, evaluation of the environmental score yields 5 which is the maximum of table environmental scores 3 and 5. Database score evaluation formula follows in fig. 3.

$$\text{SCORE}(\text{DATABASE}) = \max\left(\text{SCORE-TABLE}(i)\right) \quad \text{For all i Tables in the Database}$$

Fig. 3. Database scores formula

In case an organization manages several databases, the above formula may be applied to evaluate the environmental scores of all databases, yielding different risk scores, leading to varying mitigation actions for certain databases.

V. RESULTS

Fig. 4 illustrates the environmental scores computed by the scoring algorithm. The horizontal axis presents data items from 1 to 11 in our use case. The blue line shows the environmental score of the database as one entity. The red line shows the environmental tables' scores, which are 3 and 5. The green line shows data items environmental scores**.** In a case when a data item was stolen or damaged the score indicates the risk reflecting the damage to the organization caused by that data item. Data items' scoring calculations yields different scores. For example the computed score of data item number 1 is maximal (5) while the score of data item 3 is minimal (0.3) which leads to the conclusion that scoring all data items with one identical score is clearly far away from the specific characteristics as were specified by the users to the environmental parameters. In cases the organization can be specific about the damaged table then the scores evaluated will be 3 or 5 depending on the table. In cases the organization is unable to identify which data was damaged then a general score of 5 will be assigned, knowing that that score does not reflect the varying tables' risk scores and data items' risk scores included in the database.

As illustrated in this work, basing risks evaluations on high-level information yields higher risk scores which do not reflect the actual real environment and overestimate potential risk scores. The proposed framework enables organizations' risk managers getting improved risk scores based on high-resolution information of their configuration, thus allocating lower budgets to risk mitigation activities. Moreover, risk managers are now able to design risk management work plans based on accurate risk scores, enables being more efficient in risk management and allocating appropriate budgets to mitigate actual risks. Another benefit of the proposed model is using it by system managers and database administrators for database design. They can now use enhanced defense tools to protect their sensitive data from risks, for example, improved encryption techniques for higher security risk scored data items.

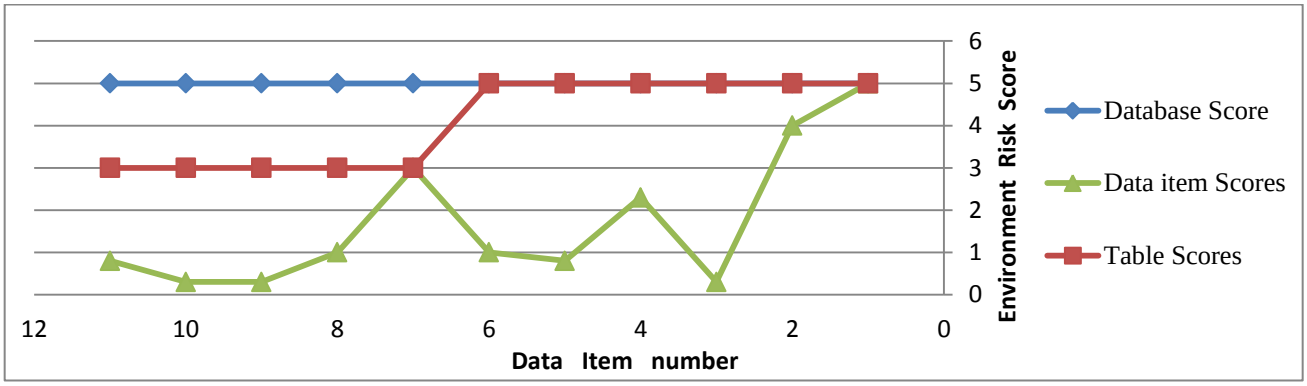


Fig. 4. Relative Environmental Scores

VI. Conclusions

In this work we described a framework of a Security Continuous Monitoring System, structure and mechanisms. The system introduces two new capabilities. First, the proposed model is basing risk scores on the actual environment and configuration of the target system, while existing models use users' estimations to environmental risk scores. Second, risk scoring is based on detailed information of the database in the resolution of data items. According to existing models and known practices, the resolution of the environmental information is defined by means of a whole database or a whole system, thus causing un-accuracies. Current practices produce high-level risk scores while the proposed model produces scores based on detailed information in the highest resolution levels. Accurate risk scoring enables efficient management of risk organizational budgets.

Future research direction may be the design of additional environmental parameters of the system such as hardware, software and communication components, for incorporation in the security scoring model. Another research direction is improving the CMS mechanisms by designing the interface between the organizational database and the CMDB, and automating the interfacing process.

References

[1] P. Mell, K. Scarfone, and S. Romanosky, “CVSS – A complete guide to the common vulnerability scoring system, version 2.0”, 2007.

[2] L. Langer, "Stuxnet: dissecting a cyber warfare weapon, security and privacy", IEEE, Volume 9 Issue 3, pages 49-51, NJ, USA, 2011.

[3] S. Tom and D. Berrett, "Recommended practice for patch management of control systems", DHS National Cyber Security Division Control Systems Security Program, 2008.

[4] A. Terje and R. Ortwin, "On risk defined as an event where the outcome is uncertain", Journal of Risk Research Vol. 12, 2009.

[5] Y. F. Nũez, "Maximizing an organizations' security posture by distributedly assessing and remeding system vulnerabilities", IEEE – International Conference on Networking, Sensing and Control, China, April 6-8, 2008.

[6] K. Dempsey, N. S. Chawia, A. Johnson, R. Johnson, A. C. Jones, A. Orebaugh, M. Scholl and K. Stine, "Information security continuous monitoring (ISCM) for federal information systems and organizations", NIST, 2011.

[7] M. G. Hardy, "Beyond continuous monitoring: threat modeling for real-time response", SANS Institute, 2012.

[8] A. Keller and S. Subramanianm, "Best practices for deploying a CMDB in large-scale environments", Proceedings of the IFIP/IEEE International conference and Symposium on Integrated Network Management, pages 732-745, NJ, IEEE Press Piscataway, 2009.

[9] M. R. Grimalia, L. W. Fortson and J. L. Sutton, "Design considerations for a cyber incident mission impact assessment process", Proceedings of the Intrnational Conference on Security and Management (SAM09), Las Vegas, 2009.

[10] I. Kotenko and A. Chechulin, "Fast network attack modeling and security evaluation based on attack graphs", Journal of Cyber Security and Mobility Vol. 3 No. 1 pp 27-46, 2014.

[11] FIPS Publication 199 - Federal Information proccessing standards publication, "Standards for security categorization of federal information and information systems", Department of Commerce, USA, February, 2004.

[12] E. Weintraub and Y. Cohen, "Continuous monitoring system based on systems' environment", ADFSL - Conference on Digital Forensics, Security and Law, May 19, 2015, Florida, USA.

[13] GAO – United States Government Accountability Office Report to Congressional Request, "Information security – state has taken steps to implement a continuous monitoring application but key challenges remain", July, 2011.

[14] National Vulnerability Database, "Common vulnerability scoring system version 2.0 calculator", https://nvd.nist.gov/CVSS/v2-calculator, retrieved March, 3, 2016.

Scalable Hybrid Speech Codec for Voice over Internet Protocol Applications

Manas Ray
Department of Electronics & Communication Engineering Birla Institute of Technology, Mesra Ranchi, India

Mahesh Chandra
Department of Electronics & Communication Engineering Birla Institute of Technology, Mesra Ranchi, India

B.P. Patil
Department of Electronics & Telecomm. Engineering, Army Institute of Technology, Dighi Pune, India

***Abstract*—With the advent of various web-based applications and the fourth generation (4G) access technology, there has been an exponential growth in the demand of multimedia service delivery along with speech signals in a voice over internet protocol (VoIP) setup. Need is felt to fine-tune the conventional speech codecs deployed to cater to the modern environment. This fine-tuning can be achieved by further compressing the speech signal and to utilize the available bandwidth to deliver other services. This paper presents a scalable -hybrid model of speech codec using ITU-T G.729 and db10 wavelet. The codec addresses the problem of compression of speech signal in VoIP setup. The performance comparison of the codec with the standard codec has been performed by statistical analysis of subjective, objective and quantifiable parameters of quality desirable from the codec deployed in VoIP platforms.**

Keywords*—VoIP; Speech Compression; Hybrid Speech Codec; ITU-T G.729 codec; db10 wavelet; Statistical Analysis

I. INTRODUCTION

In the recent years, there has been substantial standardization in the various codec deployment and access technologies for VoIP applications. These standardizations has laid the foundation of the modern 4G technology, which is a packet switched network based on internet protocol and has data rates up to 1 Gbps [1]. This compliments the general characteristics of VoIP setup of ease of launch of additional services along with the conventional telephony –voice services[2].The integration of different web-based applications in modern VoIP service had led to the various challenges of Quality of service(QoS) faced by the VoIP application. These challenges are [3], the requirement of high bandwidth, sensitivity to propagation delay, sensitivity to jitter, etc. Hence, to address these challenges, it is desired to build a highly efficient speech compression technique. The compression is desired to conserve the precious resource of bandwidth, so as to enable the amalgamation of other application in the final VoIP package. The activity of compression of the speech signal is carried out by the codecs deployed in the setup.

Currently, various ITU standard codecs are utilized in VoIP setup for seamless interconnectivity across different systems spanning continents. The data-rates provided by these codecs are the function of engineering adjustments between the quality of voice signals, complexity and bandwidth of the codec [4]. Most commercial VoIP application currently operates on the International Telecommunication Union – Telecommunication Standardization Sector (ITU-T) standardized codec known as ITU-T G.729 codec based on the principle of Conjugate Structure Algebraic Code Excited Linear Prediction (CS-ACELP)[5].The codec is a hybrid type codec which uses the algebraic sum of the fixed codebook gain vector and the adaptive gain to arrive at the actual codebook vector for performing the linear prediction[6].The codec operates at 8kbps and produces the highest quality of speech reproduction known as "Toll Quality" for most practical and real-time conditions and hence widely deployed in all voice-based operations[6]. This technology provides optimal complexity and is best suited for multimedia digital simultaneous voice and data communication, hence makes it best suitable for VoIP applications[7].The performance of this codec have been extensively tested on various environment, which resulted in concluding that the codec performs consistently with voice signals and narrowband control signals [8], international and regional speech coding standards, channel noise and over degraded transmission channels [9]. However, to accommodate additional services further compression of the speech signal is envisaged.

In the recent years, extensive research has been carried out on the topic of wavelet transforms and its applications in signal processing. The wavelet is defined as a short wave of finite duration, whose average value is zero. It is finite in nature [10].The Wavelet transform represents the signal with very high precision and limited storage requirements [11]. Here the signal is de-composed into component-signals resembling sine-waves, having compressed information in both frequency and time domains. The scaling function Φ (z) determines the resolution of the analysis and the actual analysis is performed by the mother wavelet function Ψ (z) [12].The definition of the Wavelet transform of the signal s (z) is [13]:

$$W_{\Psi}(p,q) = \int_{-\infty}^{\infty} s(z)\Psi_{pq}(z)dz$$
$$= \frac{1}{\sqrt{p}} \int_{-\infty}^{\infty} \left(s(z)\Psi\left(\frac{z-q}{p}\right)\right)dz \quad (1)$$

Where

s(z)= original signal

p = scaling parameter

q = translation parameter

The wavelet function is given by

$$\Psi_{p,q} = \frac{1}{\sqrt{p}}\Psi\left(\frac{z-q}{p}\right) \quad (2)$$

The Discrete wavelet transform (DWT) is obtained by sampling the Continuous wavelet transform (CWT) with dyadic grid parameters of translation q= t and the scale $p=2^j$ and the mother wavelet is defined by

$$\Psi(x) = 2^{j/2}\Psi(2^j x-t) \quad (3)$$

Similarly the scaling function is defined as:

$$\phi(x) = 2^{j/2}\phi(2^j x-t) \quad (4)$$

The original function s(x) can be reconstructed as $s'(x)$ from the scaling and the wavelet functions by the relation [14]:

$$s'(x) = \sum_{t=-\infty}^{\infty} c_t\phi_t(x) + \sum_{t=-\infty}^{\infty} d_{j,t}\Psi_{j,t}(x) \quad (5)$$

Where C_t are the average coefficients and $d_{j,t}$ are detail coefficients. The resultant signal thus generated from the wavelet coefficients as referred in equation 5 above provides a compressed representation of the original signal.

The codec we propose aims to enhance the performance of the CS-ACELP codec. The codec utilizes the concept of CS-ACELP as well as that of the wavelet transform, thereby providing an ideal combination of compression and standardization for deployment in VoIP setup.

The paper is organized into six sections. Section-I introduces the requirement of further compression of the speech signal in modern VoIP applications. Section –II provides a brief background of speech signal compression using wavelets. The concept of the proposed scalable hybrid speech codec is presented in Section-III. Section-IV defines the performance evaluation parameters. The results of the assessments are presented in Section – V. Section VI presents the conclusion.

II. Speech Signal Compression Using Wavelets

The wavelet transform is a transformation used to study the temporal and spectral properties of non-stationary signals like speech, audio etc. based on the frequency-time multi-resolution property of wavelets [15].The mother wavelet performs the analysis of the speech signal in wavelet domain.The general criteria for the selection of wavelet family for compression are [16]: a) Availability of minimal support in frequency as well as time domain and b) Availability of a relatively large number of vanishing moments. The greater the number of vanishing moments, quicker is the decay rate of co-efficient leading to compact signal representation [17]. Further, they also provide a higher quality of the reconstructed signal, less distortion and a high degree of compression with a trade-off of higher complexity of operation [18].

The selection process of mother wavelet for the analysis of the signal is governed by the consideration of the quality parameters required in the synthesized signal[19]. In VoIP application,the desirable quality requirements of the speech signal are naturalness of the speech, intelligibility, pleasantness, possibility of recognition of speaker[20]. The desirable parameters of the codecs deployed for such applications include [21-22]: Operation in low bit rate, low complexity, robustness across different speakers and languages, robustness in performance in the presence of channel errors, etc. The latest research in the field of signal processing has identified that Daubechies family of wavelets provides optimal results in the recognition of spoken digits as compared to other wavelets [15,23,24,25] and has the advantage of approximate shift –invariance and better edge representation as compared to real-valued discrete wavelet transform[26-27]. Hence, the Daubechies family of wavelets has been considered in the proposed codec.

The Daubechies wavelet family is extensively used for analysis of speech signals. They are orthogonal wavelets, with the highest quantity of vanishing moments per given support. Here the scaling and the wavelet functions are not defined[28]. The family of the wavelets is defined as *dbN* where *db* stands for family name Daubechies, *N* represents coefficients generated, similarly the number of vanishing moments being *N/2*[12,29].

Wavelet transform decomposes the signal into Average coefficients or scaling co-efficient indicating low-frequency components and Detail co-efficients or wavelet –coefficients containing high-frequency components. In the speech signal, high frequencies are present at the onset for a very brief period, while lower frequencies are present later for longer periods [30].Wavelet transforms resolve all these frequencies simultaneously localized in time to a level proportional to their wavelength, thereby obtaining localization in time as well as frequency. Recent studies of the wavelet decomposition [31] indicates that less than 5% of the maximum value is present in 90% of the wavelet co-efficient and hence can be treated as redundant for the purpose of signal analysis. Compression of the signal is hence achieved by truncating the redundant signals (low valued co-efficient) to zero and reconstructing the signal using the remaining co-efficient [17]. The speech compression technique using wavelet is graphically explained in Fig.1.

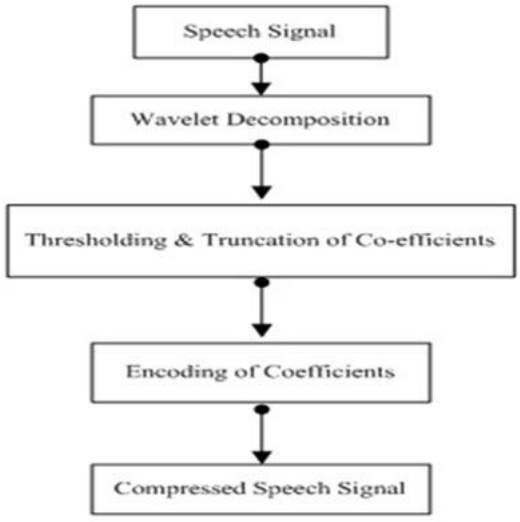


Figure 1. Speech Signal Compression Using Wavelets

The proposed codec utilizes the Daubechies family of the wavelet, specifically db10 wavelet for the wavelet analysis part of the codec.

III. Proposed Scalable Hybrid Codec

The work flow diagram of the proposed codec is presented in Fig. 2. In order to compress the bandwidth requirement for the operation of the standard G.729 codec, it is proposed to design a scalable hybrid codec, in which the output of the codec be cascaded to a wavelet-based compression technique to obtain the final compressed reconstructed signal.

The CS-ACELP codec acts as the core-layer for speech signal processing and the wavelet compression acts as the enhancement layer for the required compression. The original signal is fed to a conventional ITU G.729 codec to obtain 8kbps synthesized speech signal. This in-turn is subjected to 5-level recursive wavelet decomposition, since beyond the fifth level of wavelet decomposition, no added advantage is available in for signal processing [11].We have chosen Daubechies family of wavelets for the decomposition as they concentrate in their approximation coefficients more than 96% of the signal energy [31].Further, as described in the previous section, Daubechies family of wavelets provides better results in speech recognition, better edge representation for speaker identification and has approximate shift in-variance, which is best suitable for VoIP speech analysis.

The decomposed signal is subjected to thresholding in-order to truncate and de-noise the signal thereby improving the Signal to Noise ratio (SNR) [32]. Adaptive or soft thresholding is used in the proposed codec as this provides the efficient method of de-noising depending upon the signal under study[33]. The thresholding procedure determines the magnitude of compression of the signal[34]. The input signal is approximated from the truncated coefficients by applying inverse wavelet transform. The signal thus synthesized is the compressed version of the original speech signal.

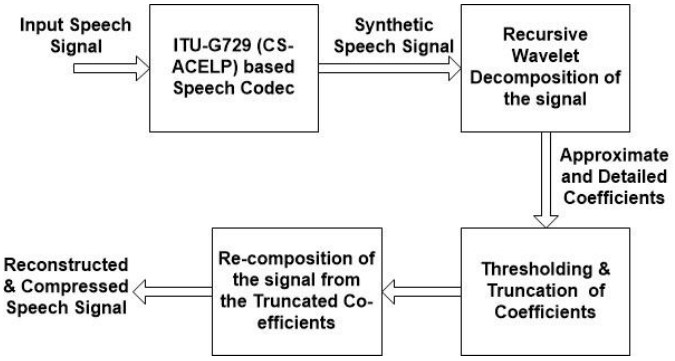


Figure 2. Proposed Speech Codec

As can be observed from the work flow diagram, the complexity of the proposed codec would be bit higher because of the requirement of wavelet transform calculations. The thresholding for compression employed is minimax principle thresholding as available in statistics.

IV. Performance Evaluation Parameters

Subjective and objective test are carried out on a standalone ITU –G.729 CS-ACELP based speech codec and the proposed codec to compare their performance. The test is performed using 08 samples of speech of both male and female speakers of English and Hindi languages. The age group of the speakers is 30-35 years and is of Indian ethnicity. Calculation of the Mean Opinion Score (MOS) is carried out to determine the subjective evaluation of the speech codec. In MOS test, the original signal and the reconstructed signals are presented to a user, who then provide an acceptability grading between 1 & 5, where 5 is excellent grade [35]. The general rating of CS-ACELP is in the range of 4.1-4.5[36].The objective performance of the codecs is assessed by measuring the parameters of Perceptual Objective Listening Quality Assessment (POLQA), Compression Ratio (CR), Normalized Root Mean Square Error (NRMSE), SNR and the total response time of the algorithm to process the speech signal sample. POLQA [37] is an ITU-T objective speech quality measurement recommendation described as the P.863 recommendation. It analyzes both the original and re-constructed signal sample by sample in the frequency domain with the temporal alignment of the original and the reconstructed signal to provide the MOS mapping [38].The mathematical expressions of the parameters viz. CR, SNR, NRMSE is given below[39-40]:

Compression Ratio is defined as:

$$CR = \frac{Bit\ length\ of\ (c(k))}{Bit\ length\ of\ (m(k))} \quad (6)$$

Where, c*(k)* is the original signal

m*(k)* is the reconstructed signal

Signal to Noise Ratio is defined as:

$$SNR = 10log_{10}\left(\frac{\sigma_x^2}{\sigma_e^2}\right) \quad (7)$$

Where σ_x^2 is the mean square of the original signal

σ_e^2 is the mean square difference between the original and synthetic signal.

Normalized Root Mean Square Error (NRMSE) is defined as:

$$NRMSE = \sqrt{\frac{(c(n)-m(n))^2}{(c(n)-\mu m(n))^2}} \quad (8)$$

Where, c (n) is the original signal,

m (n) is the synthetic signal

μm(n) is the mean of the original signal.

The bit rate in kbps is determined as [13]:

$$Bit\ rate = \frac{\text{Size of the reconstructed signal file in kilobits}}{Length\ of\ the\ reconstructed\ signal\ in\ seconds} \quad (9)$$

V. Results

ITU G-729 (CS-ACELP) based codec and the proposed scalable hybrid codec based on CS-ACELP and Daubechies family based wavelet is simulated in MATLAB. The test samples referred in Table-I are iterated against both the codecs and the individual performance is noted.

TABLE I. DETAILS OF SENTENCES USED IN THE EXPERIMENT

Sample No.	Speaker	Language	Sample Sentence
1	Male	English	Welcome to the internet my friend, How can I Help you?
2	Male	English	All systems are running as per predefined parameters
3	Male	Hindi	यहाँ से लगभग पाँच मिल दक्षिण पश्चिम में कटघर गाँव है ।
4	Male	Hindi	धोबिन जब सो कर उठती , तब देखती की चौंका साफ़ पड़ा हैं और बर्तन मजे हुए हैं ।
5	Female	English	Welcome to the internet my friend, How can I Help you?
6	Female	English	All systems are running as per predefined parameters
7	Female	Hindi	यहाँ से लगभग पाँच मिल दक्षिण पश्चिम में कटघर गाँव है ।
8	Female	Hindi	धोबिन जब सो कर उठती , तब देखती की चौंका साफ़ पड़ा हैं और बर्तन मजे हुए हैं ।

The quality of the speech codec is evaluated by measuring the MOS, POLQA, Compression Ratio, SNR, and NRMSE for the individual samples. The total response time to process the speech signals is also presented for analysis. These evaluation processes are carried out on Intel i-5 processor with 4GB RAM.

The operational bit-rate observed in the simulation of the proposed codec is 4kbps employing the minimax threshold in MATLAB simulation software. The details of the observation are presented in Table II. The individual results are provided in the following figures. Fig. 3 presents the comparison of the codecs in terms of the MOS, comparison in terms of the Compression Ratio is provided in Fig. 4, Fig. 5 compares the codecs in terms of SNR, analysis of the codecs in terms of NRMSE values are presented in Fig. 6, Fig. 7 compares the codecs in terms of the POLQA and finally the Fig. 8 provides the analysis of the delay response of the codecs. Fig. 9 presents the graphical representation of the original signal and the reconstructed signal output of the proposed codec. The results are summarized in Table III.

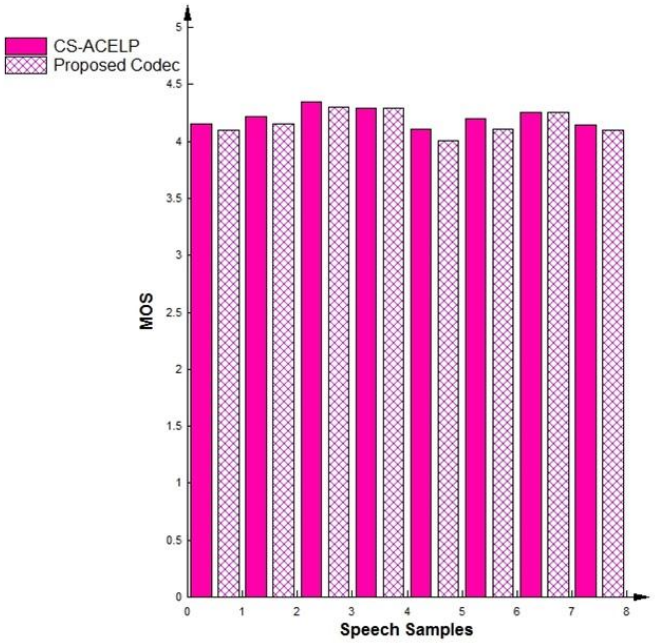


Figure 3. Comparison of MOS

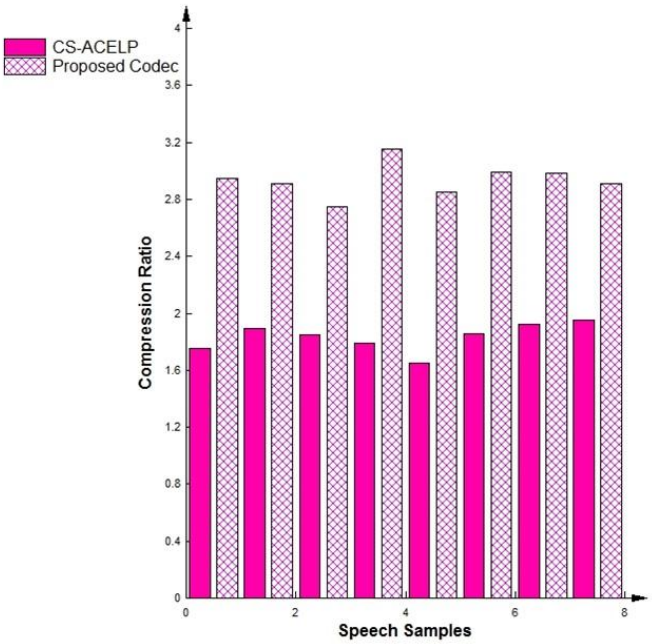
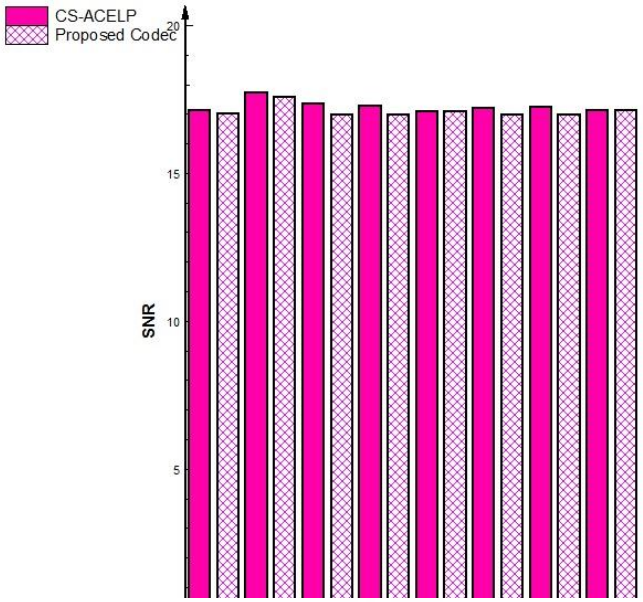
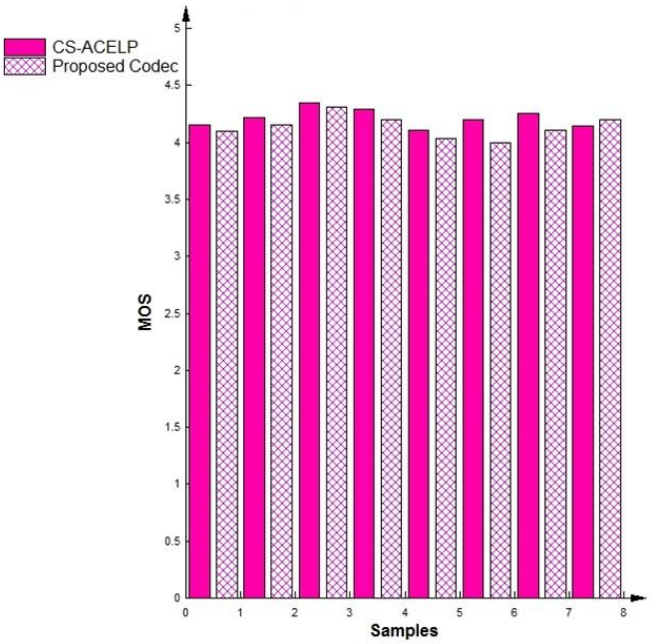


Figure 4. Comparison of Compression Ratio





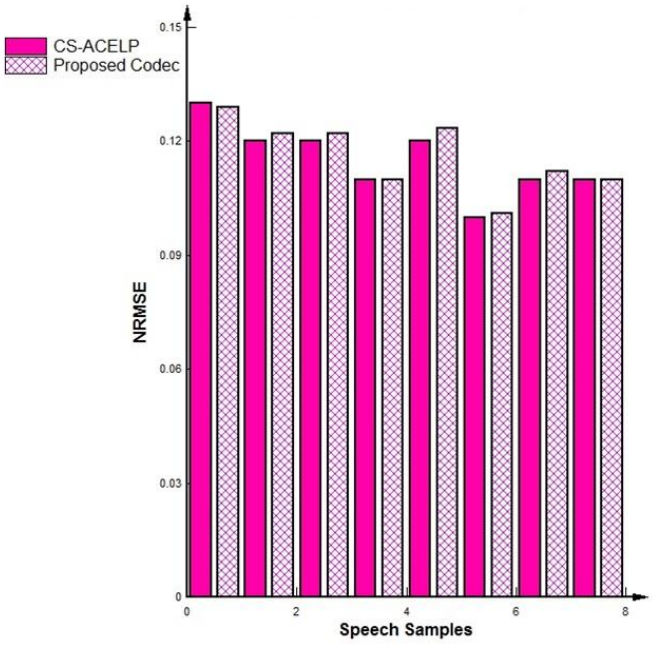


Figure 6. Comparison of NRMSE

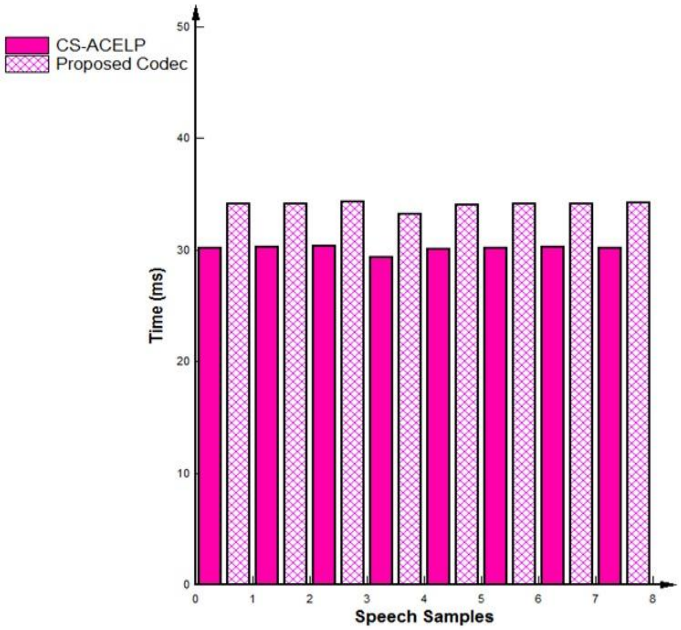


Figure 8. Comparison of Delay Response

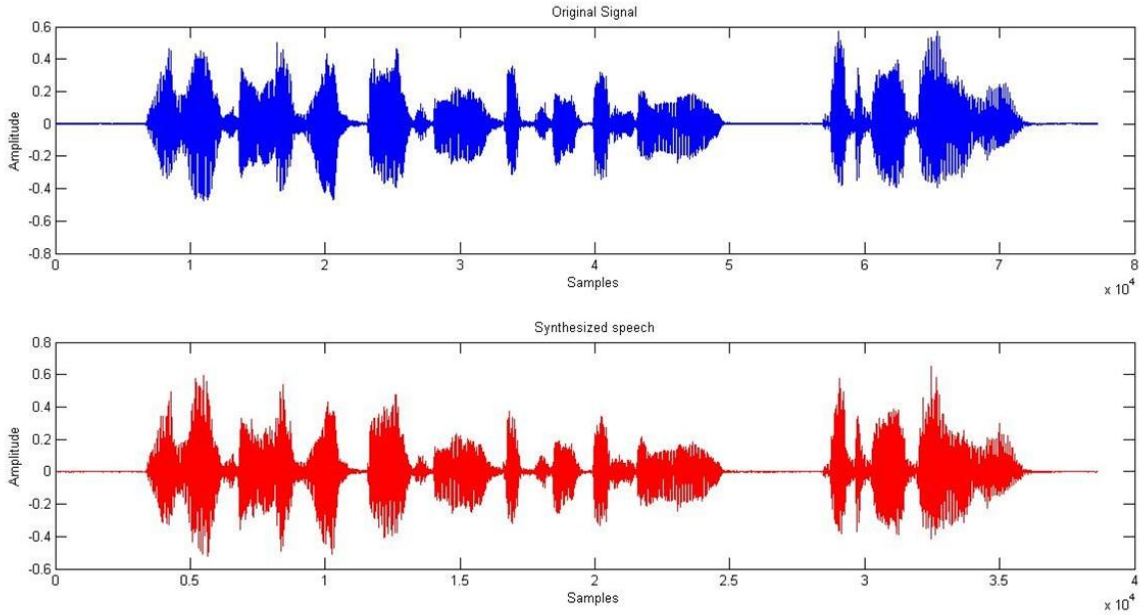


Figure 9. Speech Signal Compression Using Proposed Codec

TABLE II. SUMMARIZATION OF THE BIT RATES OBSERVED USING THE CODECS

	CS-ACELP Codec			**Proposed Codec**		
Sample	Length of Reconstructed Signal(seconds)	Size of Reconstructed Signal (Kb)	Bit Rate Observed (Kbps)	Length of Reconstructed Signal (seconds)	Size of Reconstructed Signal (Kb)	Bit Rate Observed (Kbps)
1	5	44	8.8	5	18	3.6
2	3	25	8.3	3	11	3.6
3	5	42	8.4	5	17	3.4
4	10	83	8.3	10	41	4.1
5	5	42	8.4	5	19	3.8
6	3	24	8.0	3	11	3.6
7	5	42	8.4	5	19	3.8
8	10	82	8.2	10	41	4.1

TABLE III. SUMMARIZATION OF THE RESULTS OF PERFORMANCE ANALYSIS OF THE PROPOSED CODEC WITH CS-ACELP CODEC

Sr. No	**Codec**	**Bit Rate (Kbps)**	**Average MOS**	**Average POLQA**	**Average SNR (dB)**	**Average NRMSE**	**Average CR**	**Average Delay (ms)**
1	CS-ACELP	8	4.33	4.25	17.35	0.13	1.82	30
2	Proposed Codec	4	4.30	4.22	17.30	0.13	2.95	34.75

VI. CONCLUSIONS

The performance of the CS-ACELP based standard codec (ITU-T G.729) is compared with the proposed scalable hybrid codec in this paper. It is observed that proposed codec operates at a bit rate of 4kbps and provides greater degree of compression than the standard codec. The compression can further be fine-tuned to incorporate additional application in the final service delivery by altering the threshold selected for truncating the wavelet co-efficients. It is observed that the proposed codec provides an additional delay to the tune of 5ms which is attributed to higher complexity of calculation of the wavelet transform. Further it is observed that the proposed codec provides comparable results in other parameters under test. It is also observed that the proposed codec provides robust performance across speakers and languages.

The additional delay of 5ms may be treated as acceptable on real-time applications, considering the fact that the codec requires lower bit rate of operation than the ITU-T G.729 standard codec while providing comparable performance in terms of MOS, SNR etc.

Based on the above it can be concluded that the proposed codec provides a viable alternative to the currently deployed codecs in the VoIP setup with an additional feature of high degree of compression as desired in the futuristic VoIP deployments.

REFERENCES

[1] Reference of 4G[Internet], WikiPedia,2014[updated 2016 February 14; cited 2016 March 10]. Available from: http://en.wikipedia.org/wiki/4G

[2] Sinaepourfard A, Hussain HM, "Comparison of VoIP and PSTN services by statistical analysis". Proceedings of Student Conference on Research and Development(SCOReD).Cyberjaya,pp. 459-461,December - 2011.

[3] Ray M, Chandra M, Patil BP, " Evaluation of CDMA Microwave Links at Different Environments for VoIP Applications". International Journal of Advance Research in Computer and Communication Engineering (IJARCCE) . vol. 1, Issue 8, pp. 508-512, October 2012.

[4] Ray M, Chandra M, Patil BP, "Speech Coding Techniques for VOIP Applications: A Technical Review", World Applied Sciences Journal(WASJ) , vol.33, Issue 5,pp.736-743, May 2015.

[5] Sat B., Wah B, "Speech and Network -Adaptive Layered G.729 Coder for Loss concealments of Real-Time Voice Over IP", 7th Workshop on Multimedia Signal Processing, pp.1-4, Shanghai, November2005..

[6] Salami R, Laflamme C, Adoul JP, Kataoka A, " Design and description of CS-ACELP: A toll quality 8kbs speech ceder", IEEE Transcations on Speech and Audio Processing, vol.6, Issue 2, pp.116-130, May 1998.

[7] Salami R, Laflamme C, Bessette B, Adoul JP. ITU-T G.729Annexure A :Reduced complexity 8kb/s CS-ACELP codec for digital simultenous voice and data. IEEE Communications Magazine , vol. 35, Issue 9,pp. 56-63, September 1997.

[8] Parkins ME,Evan K, Pascal D, Thorpe LA. Characterizing the Subjective Performance of ITU-T 8kb/s speech coding algorithm- ITU-T-G.729.IEEE Communications Magazine, IEEE Communications Magazine , vol. 35, Issue 9,pp. 74-81, September 1997.

[9] Ferraz deCampos Neto S, Karapetian W, "Performance of ITU-T G.729 8kb/s CS-ACELP speech codec with non-voice narrowband signals", IEEE Communications Magazine , vol. 35, Issue 9,pp. 82-91, September 1997.

[10] Chan YT. Wavelet Basics, Kluwer Academic Publishers,New York,1995.

[11] Agbinya JI, "Discrete Wavelet Transform Techniques in Speech Processing", IEEE TENCON-Digital Signal Processing Applications,pp.514-519, 1996..

[12] Somani KP, Ramchandran KI, Resmi NG, Insights into Wavelets from Theory to Practice, Prentice Hill International, New Delhi,2009.

[13] Kornsing S, Srinonchat J, "Enhancement of Speech Compression Technique Using Modern Wavelet Transforms", IEEE International Symposium on Computer ,Consumer and Control, pp. 393-396,Taichung, June 2012.

[14] Debdas S, Jagrit V, Chandrakar C, Quereshi MF, "Application of Wavelet Transform for Speech Processing", International Journal of Engineering Science and Technology(IJEST), vol.3, Issue 8,pp.6666-6670, August 2011.

[15] Elrgaby M, Amoura A, Ganoun A, "Spoken Arabic Digits Recognition Using Discrete Wavelet.", IEEE 16th International Conference on Computer Modelling and Simulation(UKSIM),pp. 275-279,Cambridge, March 2014.

[16] Vishwanath V, Anderson W, Rowlands J, Ali M, Tewfik A, " Real-time Implementation of a Wavelet Based Audio Coder on the T1 TMS320C31 DSP Chip",5th International Conference on Signal Processing Applications & Technology (ICSPAT), TX.,October 1994

[17] Strang G, Nguyen T, Wavelets and Filter Banks, Weslley-Cambridge Press ,USA,1996.

[18] Elaydi H, Jaber MI, Tanboura MB, "Speech Compression Using Wavelets", [Internet]. 2010[cited January 2016] Available from : http://site.iugaza.edu.ps/helaydi/files/2010/02/Elaydi.pdf

[19] Fahmy MF, El-Raheem GMA, Yaseen YM, Sommia AA., "On the Optimal Choice of wavelet bases and Signal Compression Using the wavelet lifting Scheme",IEEE Proceedings of the Nineteenth National Radio Science Conference (NRSC2002) , pp.434-441, 2002.

[20] Goode B, "Voice over Internet Protocol (VoIP)", Proceedings of IEEE, vol. 90, Issue 9, pp. 1510-1517, September 2002.

[21] Delprat M, Urie A, Envi C, "Speech Coding Requirements From The Prespective of The Future Mobile Systems",Proceedings of IEEE Workshop on Speech Coding for Teelecomunications,pp. 89-90, 1993.

[22] Benesty J.Springer , Handbook of Speech Processing, Springer-Verlag , Berlin,2008.

[23] Karam JR, Philips WJ, Robertson W, "Optimal feature Vector for Speech Recognition of Unequally Segmented Spoken Digits", IEEE Canadian Conference on Electrical & Computer Engineering, pp.327-330,Halifax, 2000.

[24] Krishnan VRV, Jayakumar A, Anto PB, "Speech Recognition of Isolated Malayalam Words Using Wavelet features and Artificial Neural Networks", 4th IEEE International Symposium on Electronic Design Testing and Applications, pp. 240-243, Hongkong, 2008.

[25] Joseph SM, "Spoken Digit Compression Using Wavelet", IEEE International Conference on Signal and Image Processing, pp. 255-259,Chennai, December 2010.

[26] Khare M, Prakash O, Srivastava RK, Khare A, "Daubechies Complex Wavelet Transform based approach for Multiclass Object Classification",IEEE International Conference on Contol, Automation and Information Sciences (ICCAIS); pp. 206-211,Gwanju, December 2014.

[27] Noureddine A, Bousselmi S, Cherif A, "Optimized Speech Compression Algorithm based on Wavelets Techniques in Realtime Implementation on DSP", International Journal of Information Technology & Computer Sciences (IJITCS),vol.7, Issue 3, pp. 33-41, February 2015.

[28] Reference on Daubechies Wavelet[Internet]. Wikipedia;2009 [updated 2016 March 09; cited 2016 March 10].Available from: http://en.wikipedia.org/wiki/Daubechies_wavelet

[29] Daubechies I., "Ortho-normal bases of compactly supported wavelets", IEEE Communication on Pure and Applied Mathematics, Volume 41, pp. 906-996, 1998.

[30] Rabiner L, Juang B,Fundamentals of Speech Recognition, Prentice Hall International Inc.,New Jersy,1993.

[31] Kinsner W, Langi A, "Speech and Image Signal Compression using Wavelets",IEEE Wescanex Conference Proceedings,pp. 368-375, Saskatoon, 1993.

[32] Sunny S, Peter SD, Jacob KP, "A New Algorithm for Adaptive Smoothing of Signals in Speech Enhancement". In Lee Garry Editor.

International Conference on Electronic Engineering and Computer Science., pp.339-343, Beijing: Elsevier Science Direct,May 2013.

[33] Ruwei L, Changchun B, Bingyin X, Moashen J., "Speech Enhancement Using the Combination of Adaptive Wavelet Threshold & Spectral Substraction Based on Wavelet Packet Decomposition" IEEE 11th International Conference on Signal Processing (ICSP),pp. 481-484, ,Beijing, 2012.

[34] Farouk MH. Application of wavelets in Speech Processing, Springer Science & Business Media, USA, 2014

[35] Ray AK, Acharya T. Information Technology: Principles and Applications, Prentice-Hill of India Private Limited, New Delhi ,2004.

[36] Nagireddi S, VOIP Voice and FAX Signal Processing, John Willey& Sons Inc. New Jersy 2008.

[37] Recommendation P.863: Perceptual Objective Listening Quality Assesment[Internet].ITU-2009[updated 2016 March 04;cited 2016 March 10].Available from: http://www.itu.int/rec/T-REC-P.863/en

[38] Reference on Perceptual Objective Listening Quality [Internet].Wikipedia;2014[updated 2015 October 26;cited 2016 February 24]. Available from: https://en.wikipedia.org/wiki/POLQA

[39] Ambika D, Radha V, "A Comparative Study between Discrete Wavelet Transform and Linear Predictive Coding", IEEE World Congress on Information and Communication Technologies, pp.965-969,Trivandum,November 2012.

[40] Najih AMMA, Ramili AR, Ibrahim A, Sayed AR, "Comparing Speech Compression Using Wavelets with other Speech Compression Schemes". IEEE Proceedings of Student Conference on Research and Development. pp. 55-58., 2003

Hyperspectral Image Classification Using Unsupervised Algorithms

Sahar A. El_Rahman[1,2]

[1]Electronics, Computers Systems and Communication, Electrical Department
Faculty of Engineering-Shoubra, Benha University
Cairo, Egypt

[2] Computer Science Department, College of Computer and Information Sciences
Princess Nourah Bint Abdulrahman University
Riyadh, Saudi Arabia

***Abstract*—Hyperspectral Imaging (HSI) is a process that results in collected and processed information of the electromagnetic spectrum by a specific sensor device. It's data provide a wealth of information. This data can be used to address a variety of problems in a number of applications. Hyperspectral Imaging classification assorts all pixels in a digital image into groups. In this paper, unsupervised hyperspectral image classification algorithms used to obtain a classified hyperspectral image. Iterative Self-Organizing Data Analysis Technique Algorithm (ISODATA) algorithm and K-Means algorithm are used. Applying two algorithms on Washington DC hyperspectral image, USA, using ENVI tool. In this paper, the performance was evaluated on the base of the accuracy assessment of the process after applying Principle Component Analysis (PCA) and K-Means or ISODATA algorithm. It is found that, ISODATA algorithm is more accurate than K-Means algorithm. Since The overall accuracy of classification process using K-Means algorithm is 78.3398% and The overall accuracy of classification process using ISODATA algorithm is 81.7696%. Also the processing time increased when the number of iterations increased to get the classified image.**

***Keywords*—hyperspectral imaging; unsupervised classification; K-Means algorithm; ISODATA algorithm; ENVI**

I. Introduction

Remote sensing is the art and science to obtain information about an object, area. It is viewed as the measurement and analysis of electromagnetic radiation transmitted through, reflected from, or absorbed and dissipated by the ambiance, the hydrosphere and by material at or near the land surface, for the purpose of interpreting and managing the Earth's resources and surroundings. Optical remote sensing makes use of visible, near infrared and short-wave infrared sensors to make pictures of the earth's surface by observing the solar radiation reflected from targets on the background as indicated in Fig. 1. Different materials reflect and absorb differently at different wavelengths. Thus, the targets can be differentiated by their spectral reflectance signatures in the remotely sensed images [1][2][3].

Hyperspectral sensors such as the Airborne Imaging Spectro-radiometer for applications (AISA) enabled the construction of an effective, continuous reflectance spectrum for every pixel in the scene. These schemes can be applied to discriminate among earth surface features [1][2][4].

By analogy to grasp the hyperspectral imaging concept better is that the human can sight visible light in the main three (RGB) bands, i.e. red, green, and blue, whereas spectral imaging divides the spectrum into many more bands which are infrared bands, RGB bands, and ultraviolet bands (see Fig. 2) [5][6].

The "hyper" in the word "hyperspectral" refer to "too many" as in "over" and indicate the massive number of wavelength bands. Hyperspectral imaging is spectrally specified, which indicates that it provides vast spectral information to distinguish and identify spectrally singular materials. Hyperspectral imaging supplies the possibility of further precise and exhaustive information extraction than potential with any other kind of remote sensing data [7]. It owns an increased capability that enhances the chance of detecting concerning materials and supplies more information needful for recognizing and classifying these materials [8][9]. The HSI pixels form spectral vectors demonstrate the spectral characteristics of these materials in the sight [10].

There are some limitations in hyperspectral images are images distortion that are resulting from the spherical of the earth, giving inaccuracies in the properties such as directions, distances, and scale. These distortions of images can be shadows; such as the shadow covers a specific area to be studied, and the brightness of the light on a specific area [11]. The size of the pixel where it is possible to be relatively large so that a pixel can contain a lot of properties and that is difficult to classify, or be very small in terms of not contains characteristics can be classified is one of the hyperspectral imaging limits [12].

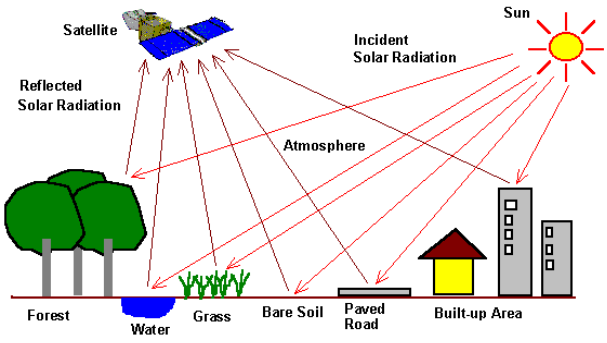


Fig. 1. Optical and Infrared Remote Sensing

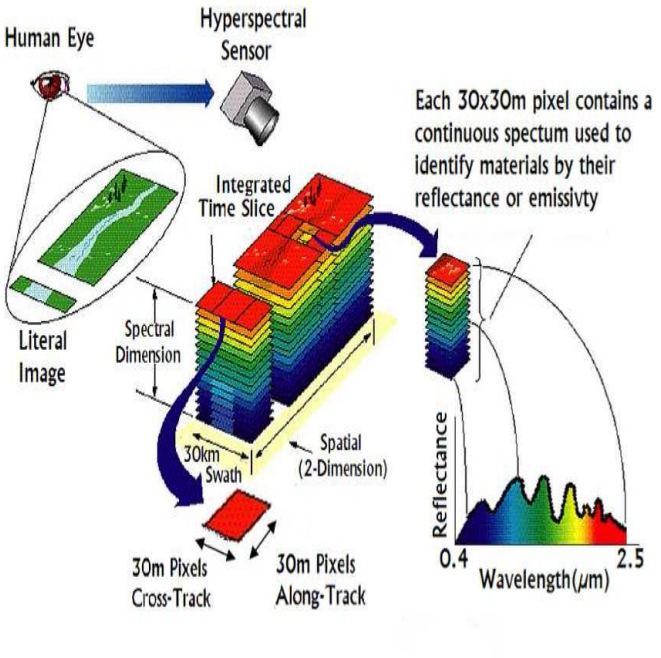


Fig. 2. HyperSpectral Imaging

II. CLASSIFICATION

The main purpose of classification of satellite imagery is to assess landscape properties accurately and extract required information [13]. Unsupervised and supervised classification algorithms are the two prime types of classification. Unsupervised classification is shown in Fig. 3 [14]. The classification chain is unsupervised, where the classification algorithms used are K-Means algorithm and ISODATA.

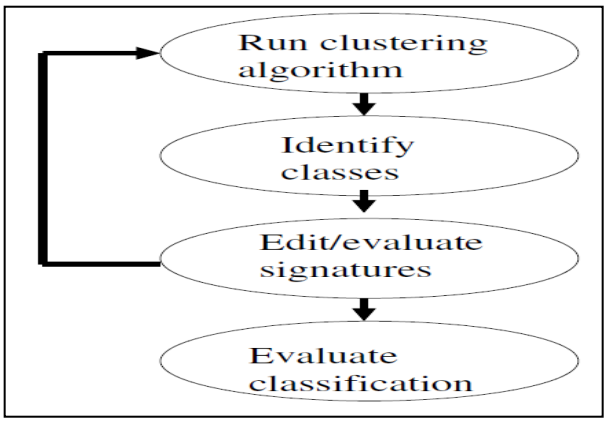


Fig. 3. Unsupervised Classification

A. K-Means Classifier

The K-means algorithm is a straightforward process for deriving the mean of a group of K-sets. The purpose of the K-Means algorithm is to reduce the cluster variability (see Fig. 4) [15][16][17]. The process of the K-means algorithm is described in the following pseudo-code [18].

1 Make initial guesses for the means $\mathbf{m}_1$, $\mathbf{m}_2$, ..., $\mathbf{m}_k$
2 Until there are no changes in any mean
3 Use the estimated means to classify the examples into clusters
4 For i from 1 to k
5 Replace $\mathbf{m}_i$ with the mean of all of the examples for Cluster i
6 end_for

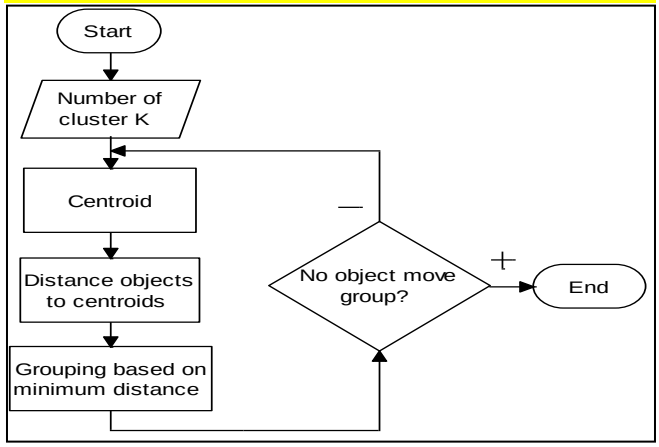


Fig. 4. K-means classifier

B. Iterative Self-Organizing Data Analysis Technique Algorithm (ISODATA)

The ISODATA algorithm is one of the most utilized methods in the unsupervised classification (see Fig. 5). In more particular, the steps in ISODATA clustering are as follows [19][20][21]:

1. Cluster centers are randomly placed.
2. Pixels are assigned based on the shortest distance to the center.
3. The standard deviation within each cluster, and the distance between cluster centers are calculated
   a. Clusters are split

If one or more standard deviation > the threshold.

   b. Clusters are merged

If the distance between them < the threshold.

4. A second iteration is performed with the new cluster centers.
5. Further iterations are performed until:
   a. The average inter-center distance falls below the user-defined threshold.
   b. The average change in the inter-center distance between iterations is less than a threshold, or the maximum number of iterations is reached.

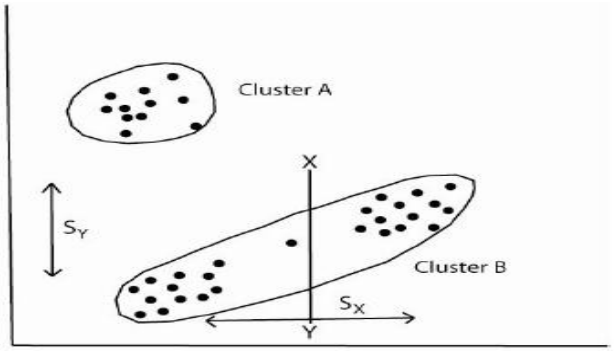


Fig. 5. ISODATA Classifier

III. Results and Discussion

The unsupervised classification was applied on a hyperspectral image using ENVI tool. The hyperspectral dataset, which has been applied to, is an image of Washington DC. The two steps that applied to the hyperspectral image are Principle Component Analysis (PCA) and K-Means or ISODATA algorithms. The result of applying K-Means algorithm and ISODATA algorithm is a classified image. Process time increased when the number of iterations increased to get the classified image. In this work, statistical information calculated from the classified image data and seen that K-Means algorithm and ISODATA algorithm are accurate since each pixel in the image is classified into a class that is not Unclassified Class. ISODATA algorithm is more accurate than K-Means algorithm since that overall accuracy of classification process using ISODATA algorithm is 81.7696% and the overall accuracy of classification process using K-Means algorithm is 78.3398%.

A. ENVI (Environment of Visualizing Images)

ENVI is an image processing system. It was designed to process remotely sensed data. It provides comprehensive data visualization and analysis for images. It has the ability of treating a broad set of scientific data formats [22][23].

B. Cases Studies

- Comparison between the results of applying different RGB bands on the same hyperspectral image.
- Studying the effects of changing the number of iterations in the process of classification on the accuracy of classification.
- Comparison between the result of applying K-Means algorithm in the first time with the result of applying the ISODATA algorithm at the second time.
- Comparison between the result of applying the PCA algorithm then applying K-Means algorithm in the first time with the result of applying only K-Means algorithm in the second time.

1) Case Study 1: It is using different RGB bands (see Table I) to the same hyperspectral image (Washington DC), and applying Principle Component Analysis (PCA) and K-means. The results of applying PCA on the image using Test values in Table I are shown in Fig. 6. The number of classes is 6 and the number of iterations is 3, where Fig. 7 shows the results of applying K-Means algorithm on the output images of PCA.

TABLE I. Case Studies Using Different RGB Bands

Test Values	R	G	B
1	180	95	35
2	176	88	23
3	29	59	129

2) Case Study 2: It is seen that when the number of iterations to classify a hyperspectral image is increased, the overall accuracy will increase while the overall accuracy will decrease when the number of iterations is decreased. This is approved through the above experiments. Table II is a summary of the results of the experiments.

3) Case Study 3: It is about using different algorithms on the same hyperspectral image (Washington DC), these algorithms are K-means and Iterative Self-Organizing Data Analysis Technique Algorithm (ISODATA). In both case as shown in Fig. 8, it is applied as a first step PCA then either K-Means algorithm or ISODATA as a second step. The selected bands are PC Band 172 for R, PC Band 86 for G, and PC Band 24 for B. Number of classes is 6 classes and the maximum iterations is 3 are chosen as K-Means parameters. Fig. 8-a shows the result of applying K-Means algorithm. Then, the number of classes range is from 4 to 6 and the maximum iterations is 3 are chosen as ISODATA parameters. Fig. 8-b shows the result of applying ISODATA.

4) Case Study 4: It It is about studying the effect of implementing PCA on classification result, as shown in Fig. 9-a, comparing to implementing just K-Means algorithm without applying PCA as shown in Fig. 9-b. This study is implemented on the hyperspectral image (Washington DC).

C. Class Statistics

1) Calculating Class statistics based on applying the K-means algorithm: Calculating statistics based on applying K-means algorithm on Washington DC image results, Fig. 10-a shows the Means for all classes and Fig. 10-b shows the standard deviation for all classes, those represent the relation between the band number and the value. Fig. 10-c shows the basic statistics include minimum, maximum and mean for each band for Tree class. Fig. 10-d shows the standard deviation for Tree class. Table III shows the class distribution summary and Table IV indicates confusion matrices using the ground truth image. A total class error is indicated in Fig. 11.

2) Calculating Class Statistics based on applying ISODATA algorithm: Calculate statistics based on applying the ISODATA algorithm on Washington DC hyperspectural image results, Fig. 12-a shows the Means for all classes and Fig. 12-b shows the Stdev for all classes, those represent the relation between the band number and the value. Fig. 12-c shows the basic statistics include minimum, maximum and mean for each band for Tree class. Fig. 12-d shows the standard deviation for Tree class. Table V shows the class distribution summary and Table VI indicates confusion

matrices using the ground truth image. A total class error is indicated in Fig. 13.

TABLE II. SUMMARY OF THE EXPERIMENTS OF CASE STUDY 2

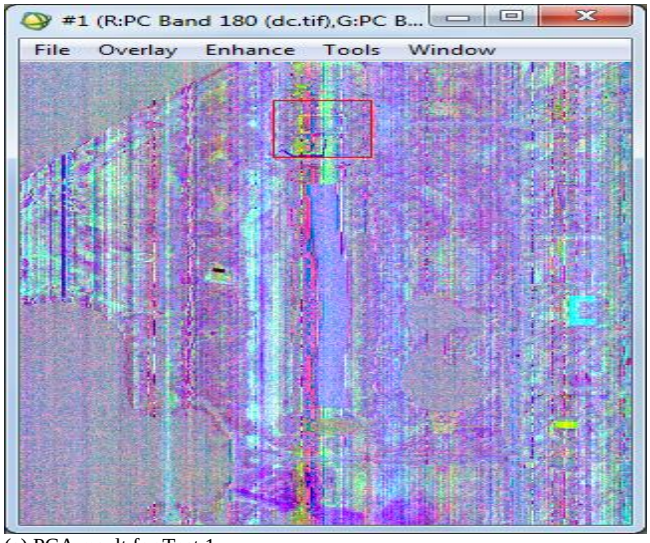
Experiment Description	Overall Accuracy	Class Percentage						
		Unclas-sified	Roof	Grass	Land	Trail	Road	Water
Comparison between the result of a classified image with only one iteration (Ground Truth) and a classified image with three iterations.	62.7079%	0.0	100	100	100	100	83.29	21.93
Comparison between the result of a classified image with three iterations (Ground Truth) and a classified image with ten iterations.	87.6110%	0.0	100	100	100	100	90.34	8.87
Comparison between the result of a classified image with only one iteration (Ground Truth) and a classified image with ten iterations.	57.8171%	0.0	100	100	100	100	89.59	27.06
Comparison between the result of a classified image with three iterations (Ground Truth) and a classified image with fifteen iterations.	87.6110%	0.0	100	100	100	100	90.34	8.87

TABLE III. CLASS DISTRIBUTION SUMMARY (K-MEANS ALGORITHM)

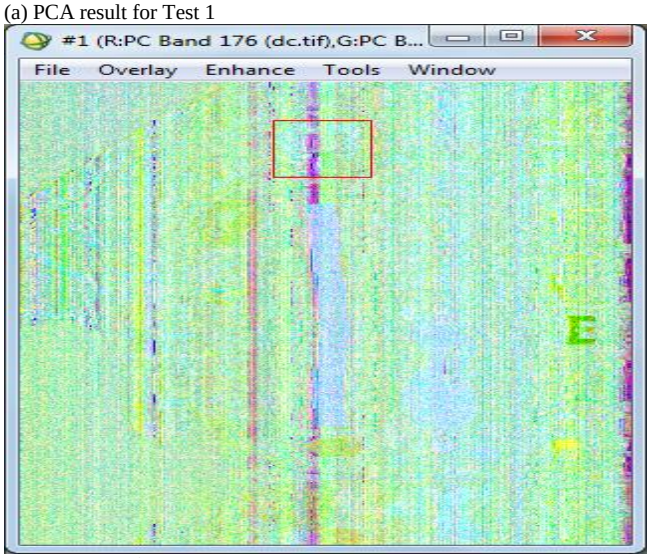
Unclassified	0 points (0.000%)
Roof	51,689 points (13.154%)
Grass	98,094 points (24.963%)
Tree	56,529 points (14.385%)
Trail	58,722 points (14.944%)
Road	69,678 points (17.732%)
Water	58,248 points (14.823%)

TABLE IV. CONFUSION MATRIX USING K-MEANS ALGORITHM

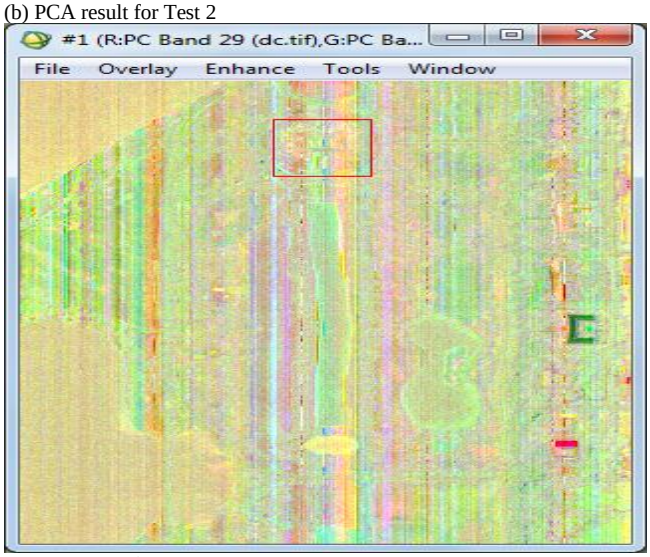
Overall Accuracy = (307844/392960) 78.3398%								
Kappa Coefficient = 0.7373								
Ground Truth (Percent)								
Class	**Unclas-sified**	**Roof**	**Grass**	**Tree**	**Trail**	**Road**	**Water**	**Total**
Unclassified	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Roof	0.00	100.00	1.59	0.00	14.92	0.00	0.00	13.15
Grass	0.00	0.00	93.34	17.02	0.15	0.00	0.00	24.96
Tree	0.00	0.00	5.07	46.06	32.74	0.00	0.00	14.39
Trail	0.00	0.00	0.00	33.14	51.08	6.26	0.00	14.94
Road	0.00	0.00	0.00	3.77	1.11	84.79	0.00	17.73
Water	0.00	0.00	0.00	0.00	0.00	8.95	100.00	14.82
Total	0.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00



(a) PCA result for Test 1



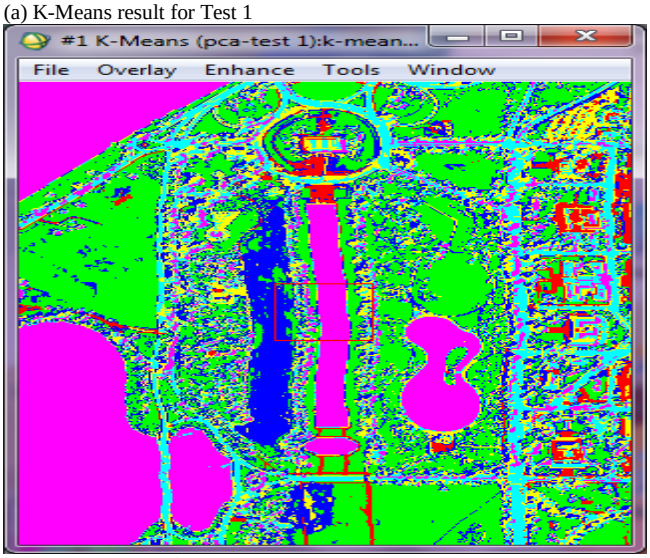
(b) PCA result for Test 2



(c) PCA result for Test 3

Fig. 6. Case Study 1: Using Different RGB Bands after Applying PCA

(a) K-Means result for Test 1



(b) K-Means result for Test 2

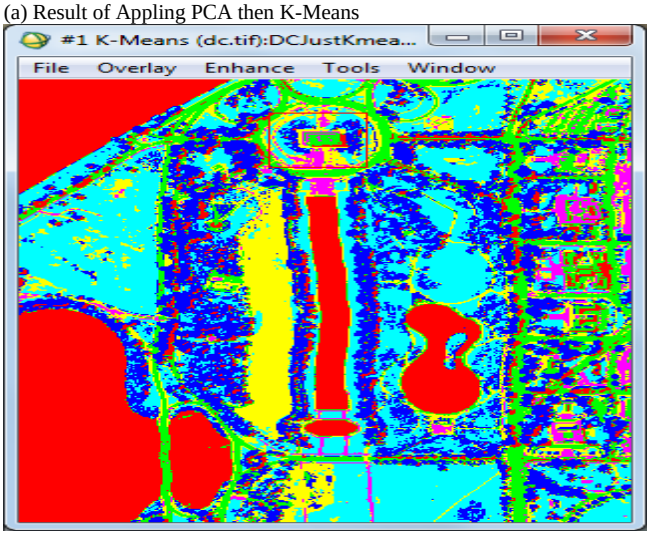
(c) K-Means Result for Test 3

Fig. 7. Case Study 1: Using Different RGB Bands after Applying PCA and K-Means

(a) Result of applying K-means algorithm.

Fig. 8. The Results of Case Study 3

(a) Result of Appling PCA then K-Means



(b) Result of Appling K-Means without PCA

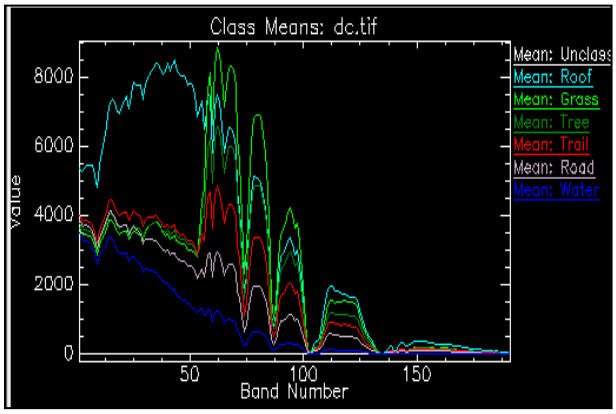
Fig. 9. The Results of Case Study 4

TABLE V. CLASS DISTRIBUTION SUMMARY (ISODATA ALGORITHM)

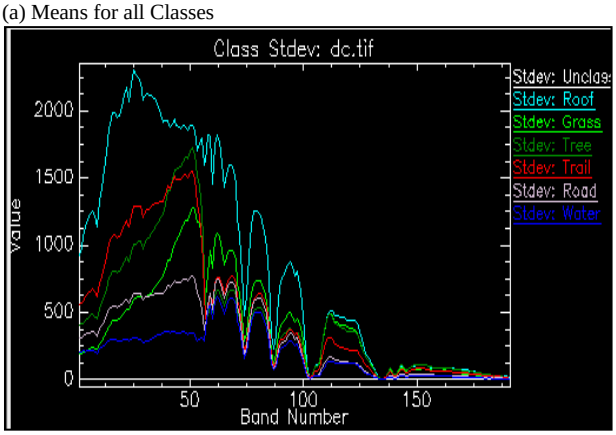
Unclassified	0 points (0.000%)
Trail	6.697 points (1.704%)
Grass	119.476 points (30.404%)
Tree	62.947 points (16.019%)
Road	76.234 points (19.400%)
Water	70.155 points (17.853%)
Roof	57.451 points (14.620%)

TABLE VI. CONFUSION MATRIX USING ISODATA ALGORITHM

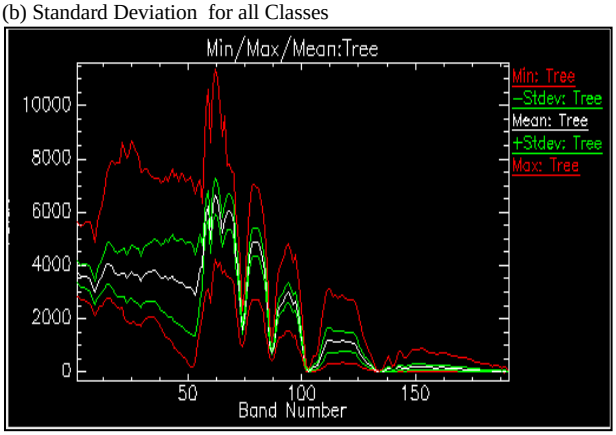
Overall Accuracy = (321322/392960) 81.7696%								
Kappa Coefficient = 0.7726								
Ground Truth (Percent)								
Class	**Unclas-sified**	**Trail**	**Grass**	**Tree**	**Road**	**Water**	**Roof**	**Total**
Unclassified	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Trail	0.00	36.54	0.10	0.00	0.00	0.00	2.32	1.70
Grass	0.00	0.00	99.81	21.07	0.00	0.00	10.95	30.40
Tree	0.00	0.00	0.09	60.90	14.40	0.00	8.88	16.02
Road	0.00	0.00	0.00	18.03	78.73	4.96	0.00	19.40
Water	0.00	0.00	0.00	0.00	1.53	95.04	0.00	17.85
Roof	0.00	63.46	0.00	0.00	5.34	0.00	77.85	14.62
Total	0.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00



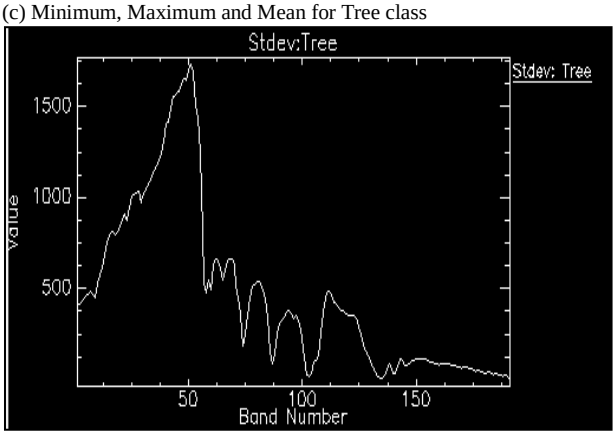
(a) Means for all Classes



(b) Standard Deviation for all Classes



(c) Minimum, Maximum and Mean for Tree class



(d) Standard Deviation for Tree Class

Fig. 10. Class statistics based on applying K-means algorithm

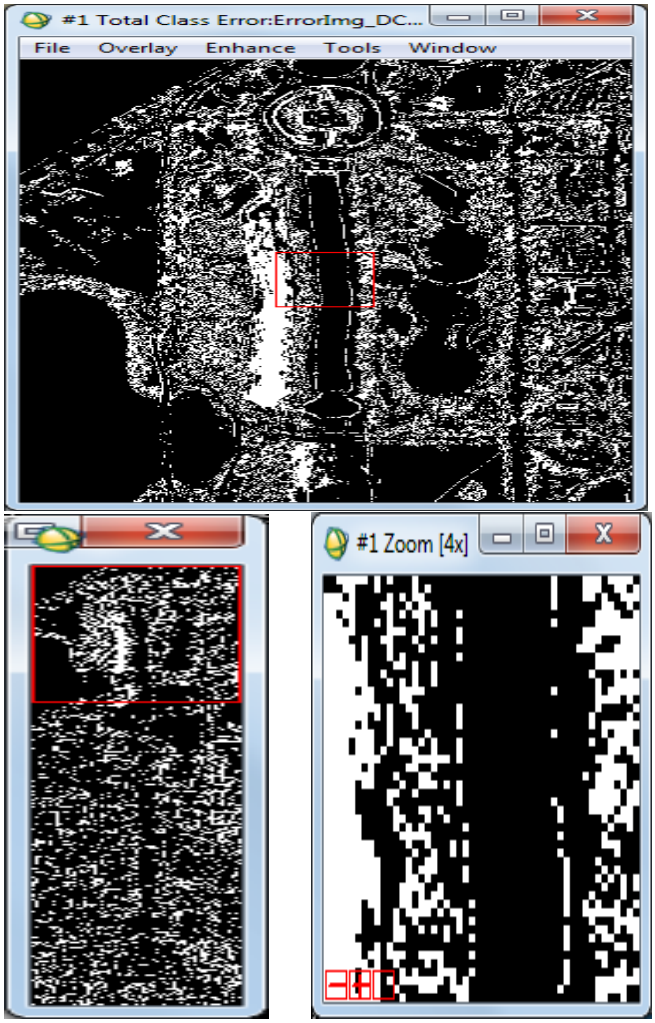
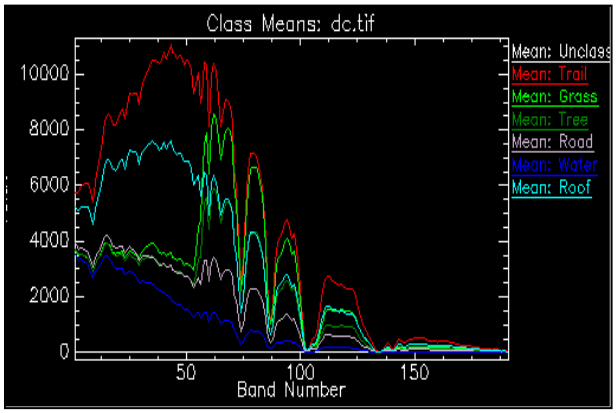


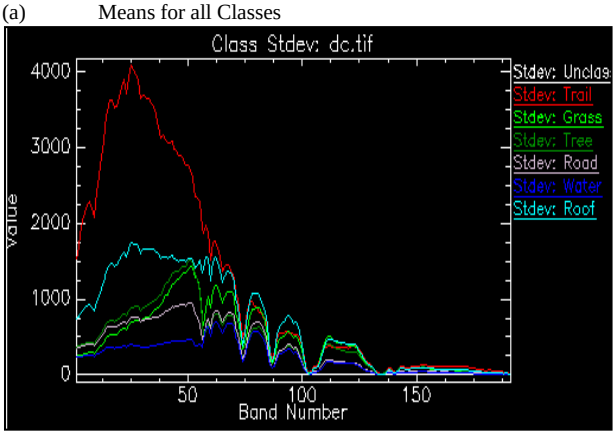
Fig. 11. Total Class Error (K-means)

IV. Conclusion

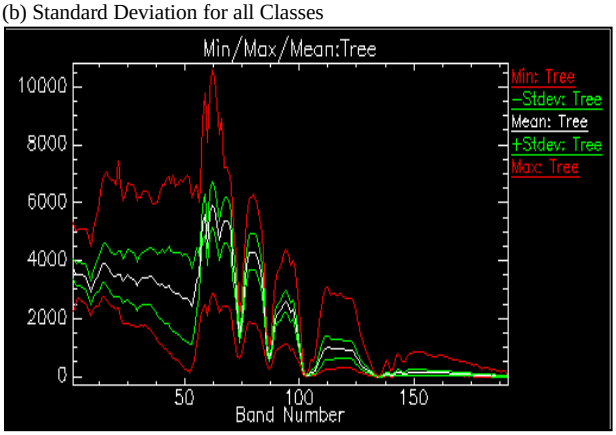
Hyperspectral images have broad spectral information to identify and distinguish materials spectrally unique. Classification of hyperspectral image means assigning objects with the same level of a class with homogeneous characteristics. In this work, unsupervised classification algorithms (K-Means algorithm and ISODATA algorithm) are used after applying Principle Component Analysis (PCA) using ENVI tool. PCA is used before the classification process as a technique in data analysis to reduce hyperspectral image dimensions. They are applied in a test site representative in the study area in Washington DC, USA. The overall accuracy was reported as 78.3398% for the K-Means classification approach, and 81.7696% for the ISODATA classification approach. It is found that K-Means algorithm and ISODATA algorithm give accurate results, but ISODATA algorithm has a better result on the study area image.



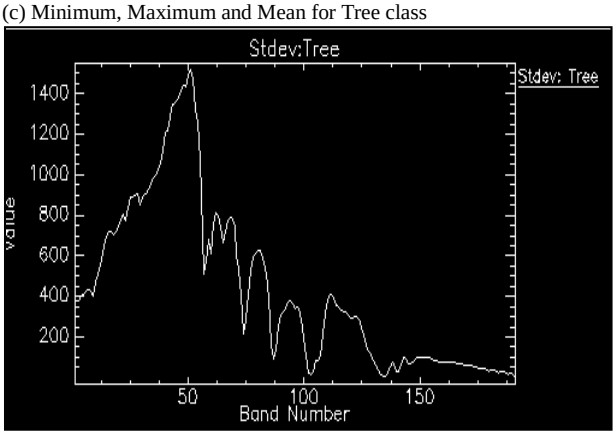
(a) Means for all Classes



(b) Standard Deviation for all Classes



(c) Minimum, Maximum and Mean for Tree class



(d) Standard Deviation for Tree Class

Fig. 12. Class statistics based on applying ISODATA algorithm

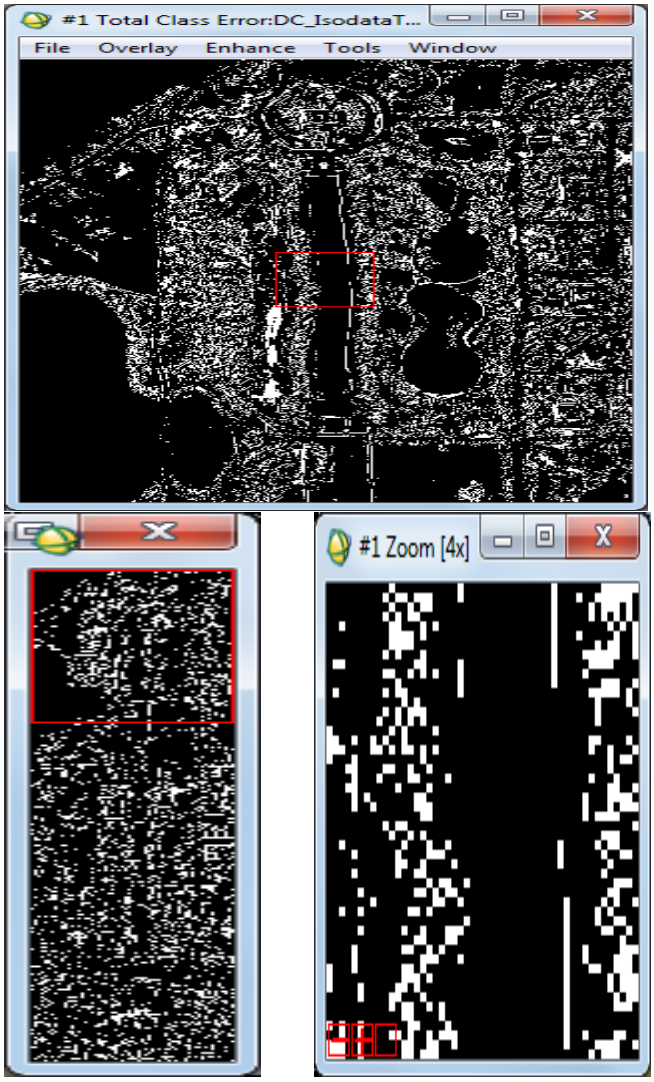


Fig. 13. Total Class Error (ISODATA)

V. Future Work

Further studies should be conducted in studying the underground water, and soil in different regions in the Arab region for starting new wheat pivots. Also, the regions are not suitable for growing wheat in order to save money and effort if it gets infected.

Acknowledgment

The author would like to thank all the participants involved in this work.

References

[1] Sachin T. Jadhav and Mandar M. Kolap, "A Review on Methodologies of Remote sensing, GPS And GIS", IOSR Journal of Electronics & Communication Engineering (IOSR-JECE), ISSN : 2278-2834, ISBN : 2278-8735, PP. 1-6.

[2] Foudan Salem, Tarek El-Ghazawi, and Menas Kafatos, "Remote Sensing And Image Analysis For Oil Spill Mitigation In The Red Sea", 2nd Biennial Coastal GeoTools Conference Charleston, SC, January 2001.

[3] Principles Of Remote Sensing, ITC Educational Textbook series, ©ITC 2001. [Online]. Available: http://www.gdmc.nl/oosterom/PoRSHyperlinked.pdf

[4] Foudan Salem and Menas Kafatos, "Hyperspectral Image Analysis For Oil Spill Mitigation", 22nd Asian Conference on Remote Sensing, Singapore, November 2001.

[5] "Hyperspectral imaging", Fleurus (Municipality, Province of Hainaut, Belgium): CRW Flags Inc., 2010.

[6] "HyperSpectral Image: How it happens". http://www2.brgm.fr/mineo/RapportFinal/Image16.gif

[7] S. Peg, "Introduction to Hyperspectral Image Analysis", Earth Science Applications Specialist Research Systems, Inc.

[8] P. Shipper, "Introduction to Hyperspectral Image Analysis," An International Electronic Journal, 2003. [Online]. Available: http://spacejournal.ohio.edu/pdf/shippert.pdf. [Accessed 10 october 2013].

[9] SpecTIR, "Hyperspectral Image," SpecTIR, 2012. [Online]. Available: http://www.spectir.com/technology/hyperspectral-imaging/. [Accessed 10 Nov. 2013].

[10] G. Shaw and H. Burke, "Spectral imaging for remote sensing," Lincoln Laboratory J, p. 3–28, 2003.

[11] S. Dr Ropert, "Introduction To Remote Sensing," New Mexico State University.

[12] J. B. Campbell, "Introduction to Remote Sensing" (3rd Edition), London: Taylor and Francis, 2002.

[13] [Online]. Available: http://disi.unitn.it/seminars/281. [Accessed 22 September 2013].

[14] Tan, "The K-Means algorithm," December 2006. [Online]. Available: http://www.cs.uvm.edu/~xwu/kdd/Slides/Kmeans-ICDM06.pdf. [Accessed 2013].

[15] A. Villa, J. Chanussot, J.A. Benediktsson, C. Jutten c, R.Dambreville, "Unsupervised methods for the classification of hyperspectral images with low spatial resolution", Journal of Pattern Recognition, Volume 46 Issue 6, pp. 1556-1568, June 2013.

[16] S.Praveena S.P.Singh I.V.Murali Krishna, "An Approach for the Segmentation of Satellite Images using K-means", KFCM, Moving KFCM and Naive Bayes Classifier. International Journal of Computer Applications, Volume 65, No.20, 2013.

[17] M.Priyadharshini,R.Karthi,S.N.Sangeethaa, Premalatha, k.S.Tamilselvan, "Implementation of Fuzzy Logic for the High-Resolution Remote Sensing Images with Improved Accuracy", IOSR Journal of Electrical and Electronics Engineering (IOSR-JEEE) e-ISSN: 2278-1676,p-ISSN: 2320-3331, Volume 5, Issue 3, 2013, PP 13-17.

[18] "K-Means Clustering," Home.deib.polimi.it, [Online]. Available: http://home.deib.polimi.it/matteucc/Clustering/tutorial_html/kmeans.html.

[19] "ERDAS Field Guide", 2005 Leica. Geosystems Geospatial Imaging, LLC, 2005.

[20] Mather, Paul, "Computer Processing of Remotely-Sensed Images", 2004 John Wiley & Sons, Ltd

[21] Memarsadeghi, N., Netanyahu, N.S., LeMoigne, J., "A Fast Implementation of the ISODATA Clustering Algorithm", International Journal of Computational Geometry and Applications, IJCGA, Vol. 17, No. 1, PP. 71-103, 2007.

[22] "Multispectral Processing using ENVI's Hyperspectral Tools," 10 December 1997. [Online]. Available: http://www.ltid.inpe.br/tutorial/tut14.htm. [Accessed 6 November 2013].

[23] ENVI, "ENVI User's Guide", Research Systems, Inc., 2004.

A Survey on Security for Smartphone Device

Syed Farhan Alam Zaidi
Department of Computer Science
COMSATS Institute of Information Technology,
Islamabad, Pakistan

Munam Ali Shah
Department of Computer Science
COMSATS Institute of Information Technology,
Islamabad, Pakistan

Muhammad Kamran
Department of Distance Continuing & Computer Education
University of Sindh, Hyderabad,
Pakistan

Qaisar Javaid
Department of Computer Science & Software Engineering,
International Islamic University,
Islamabad, Pakistan

Sijing Zhang
Department of Computer Science & Technology
University of Bedfordshire,
Luton, UK

***Abstract*—The technological advancements in mobile connectivity services such as GPRS, GSM, 3G, 4G, Blue-tooth, WiMAX, and Wi-Fi made mobile phones a necessary component of our daily lives. Also, mobile phones have become smart which let the users perform routine tasks on the go. However, this rapid increase in technology and tremendous usage of the smartphones make them vulnerable to malware and other security breaching attacks. This diverse range of mobile connectivity services, device software platforms, and standards make it critical to look at the holistic picture of the current developments in smartphone security research. In this paper, our contribution is twofold. Firstly, we review the threats, vulnerabilities, attacks and their solutions over the period of 2010-2015 with a special focus on smartphones. Attacks are categorized into two types, i.e., old attack and new attacks. With this categorization, we aim to provide an easy and concise view of different attacks and the possible solutions to improve smartphone security. Secondly, we critically analyze our findings and estimate the market growth of different operating systems for the smartphone in coming years. Furthermore, we estimate the malware growth and forecast the world in 2020.**

Keywords*—Smartphone Security; Vulnerabilities; Attacks; Malware

I. INTRODUCTION

The smartphone usage raised significantly in recent years, as smartphones provide users with several services like phone calls, Internet services, sharing data, keeping data, off-line games, online games, and some entertaining online/ off-line applications. As smartphone provides the vast services, thus are saddled with some challenges like security and privacy as well. Since most of the operations smartphones perform are on the Internet, so it is necessary to ensure security and safety of data and information. For smartphone authentication, a pattern like password, code password, PIN password, and face unlock can be used [1]. But these authentication methods are not secured at high ratio because with brute forcing and guessing such measures could be penetrated.

Critically, a lot of Malware, Viruses and Trojans have been developed which are based on smartphones APIs (application program interface) and most of them look like safe software; some reliable applications (Gmail, Facebook, etc.) collect user's information such as geolocation without user's knowledge with GPS service in smartphone [2]. There are many smartphone operating systems available, such as Android, iOS, Microsoft Window Phones, Symbian and BlackBerry [1]. Android is the widely used smartphone operating system with better performance as compared to other smartphone operating systems. Android OS is based on Linux operating system architecture. The desktop OS and the smartphone versions of such operating systems are very different, especially in user interfaces and system architecture. Using smartphones one can connect to the Internet and instantly communicate with friends, partners and browse data/information from the world wide web [3].

Now, smartphones pair mobile phones with other devices such as PDAs (personal data assistants), high definition camera, media player, GPS navigation units and other data storage and processing devices. Even the earlier mobile devices came with 3G and 4G compatibilities; but in the last decade, such devices transformed into mobile computers with the options of touch screen and laptop capabilities and can browse the Internet using wireless network and 3rd party applications. In the 3rd quarter of 2012, more than one million smartphones were in use [2]. According to Gartner Inc., the worldwide sales of mobile phones declined 3%, and smartphones sales were increased by 47 % in the 3rd quarter of the year 2012 [4]. In November 2012, 821 million smart devices were purchased in 2012 and 1.2 billons were sold in 2013 [5]. In August 2013, the smartphones sales were increased, and the growth was 46.5 % [6]. Some reports state that China with the highest number of smartphone users (519 million in 2014) [7]. The United States comes to the 2nd position with 165.3 million users and India to 3rd on the rank with 123.3 million users.

A. General Architecture of Smartphones

Smart devices are grouping of mobile phones and platform with rich connectivity and powerful computing proficiency. Therefore, a smartphone has the necessary modules of computing platforms, operating systems, third-party applications and smartphone hardware architectures, as shown in [8]. Unlike Android, the iOS operating system works only on iPad, iPhone, and iPod devices. To manage all operating

systems and devices, the OS provide necessary technology and interface and support to implement the new application to meet a variety of smartphone user needs.

The applications allow smartphone users to control their devices by interacting with the operating system, by such interaction users can access and control data communication interfaces and services. On another hand, the operating system can access user data and communicate directly with other services as well as devices. In general operating system can only access hardware directly, but the access to user's data might result in compromising user information and the information from the smartphone can be maltreated by attackers just like attacks on the computer such as viruses, Trojans, etc. The user data or information is the most valued property of smartphones. As discussed earlier, besides communication, smartphones connect to several other electronic devices such as computer and even servers through the Internet. The data without user's knowledge is usually retrieved through the applications infested by malicious codes or programs [8].

B. Structure of Smartphones Operating System

There are many operating systems for smartphones. In this part, we discuss Android, iOS, Windows Phone and Symbian operating system.

Android: Android is an open source mobile operating system which is based on Linux OS kernel and launched by Google. Android contains four layers including kernel, libraries, Android Runtime and Application framework. Application layer consists of all Android applications including email, SMS program, instant messaging, browsers, contacts and other various applications their names list is longer than few pages [9][10]. According to the authors in [11] and [12], application framework layer recognizes all Android applications. Libraries layer is divided into two parts: Android and Android runtime library. Android runtime combines the assets of the Java Virtual Machine and machine Dalvik. Android library consists of C / C ++ language.

iOS: The iOS is an operating system for Apple devices developed by Apple Inc. One obvious example is iPhone which was released in 2007. Now, iPhone is one of the larger competitors to the smartphone market shares. Application of Apple phone will need computer running MAC OS [13]. Like Android, new iOS has been developed for third party to overcome the capability limitations of platform [14].

Windows Phone: Microsoft Corporation has developed Windows phone operating system. In November 2011 [15], many devices has been built up for this OS including Nokia Lumia 800 and HTC Titan. After one year, Windows became the fourth most widely used operating system on the smartphone. Windows uses Android operating system like security model.

Symbian: PSION was established in 1980 before the Symbian. In 1990 [16], Symbian was created by Psion, Nokia, and Motorola. After that, some other vendors joined this corporation like Sony Erickson, Siemens in 2002. First, Symbian mobile platform was released in 2000 (EricksonR380) then Nokia announcement couple of versions (like Nokia N series). Symbian was developed with C++.

Almost all the smartphone OS provide mechanisms for users to enhance the security of their devices by certain login mechanisms. However, more than 30%, Mobile phone users do not use the PIN on their Phones. On the other hand, the amount of high valued contents stored on the phone is rapidly increasing, with mobile payment and money transfer application as well as enterprise data becoming available on mobile devices [17]. The statistical data obtained from sources [18] & [19] have been computed and represented in Table 1.

TABLE I. SMARTPHONE ESTIMATION BY OS 2014 SHIPMENT AND MARKET SHARE 2018

Vendor	2014 Shipment Values (Million)	2014 market % share	2018 Shipment values (Million)	2018 market % Share	Growth
Android	950.5	78.9 %	1321.1	76.0 %	10.7 %
iOS	179.5	14.9 %	249.6	14.4 %	10.2 %
Windows Phone	47.0	3.9 %	121.8	7.0 %	29.5 %
Black-Berry	11.9	1.0 %	5.3	0.3 %	-22%
Others	15.1	1.3 %	40.7	2.3 %	32.7 %
Total	1204.4	100.0 %	1738.5	100.0 %	11.5 %

To understand the existing security problems that distress smartphones, we examined the threats, vulnerabilities, targeted attacks on smartphone and study security solutions to protect them. Attention has also been paid to authentication issues, data protection and privacy issues. In this study, the review of related literature is made over the period of 2010-2015, by concentrating on smartphones vulnerabilities (issues that cause the attacks) and attacks (old and new attacks).

The paper is organized as, Section II introduces some background ideas and previous studies regarding the authentication problem, data protection and privacy, smartphone vulnerabilities, and attacks. The smartphone attacks are divided into two categories: Old attacks and new attacks. Section III evaluates the related works discussed in Section II. In this section, we summarize the old and new attacks, causes of attacks (vulnerabilities) their impact and solution to protect the smartphones. Section IV is a discussion; we discussed some open issues and possible future problems of smartphones in IoT (Internet of things). Finally, Section V draws some conclusion.

II. SMARTPHONE PROBLEMS

Powerful hardware, advanced operating system, latest applications, increasing capabilities of smartphone and functionality are enough, but an increase in present security threats in smartphones became a prominent issue. Other features of the smartphone such as broad bandwidth accelerators of the Internet, multiple peripheral interfaces also spreading viruses over the network. The multi-connectivity

gains high risk and make it easier to transmit viruses those may be aggravate threats [20][21]. The security challenges in the mobile environment are similar to the problems encountered in the personal computer world. Threat means possible destructions of smartphone security. Considering that the smartphone problems can be categorized into four categories: Authentication, Data Protection and Privacy, Vulnerabilities and Attacks. Fig.1 shows such categorization of smartphone security problems.

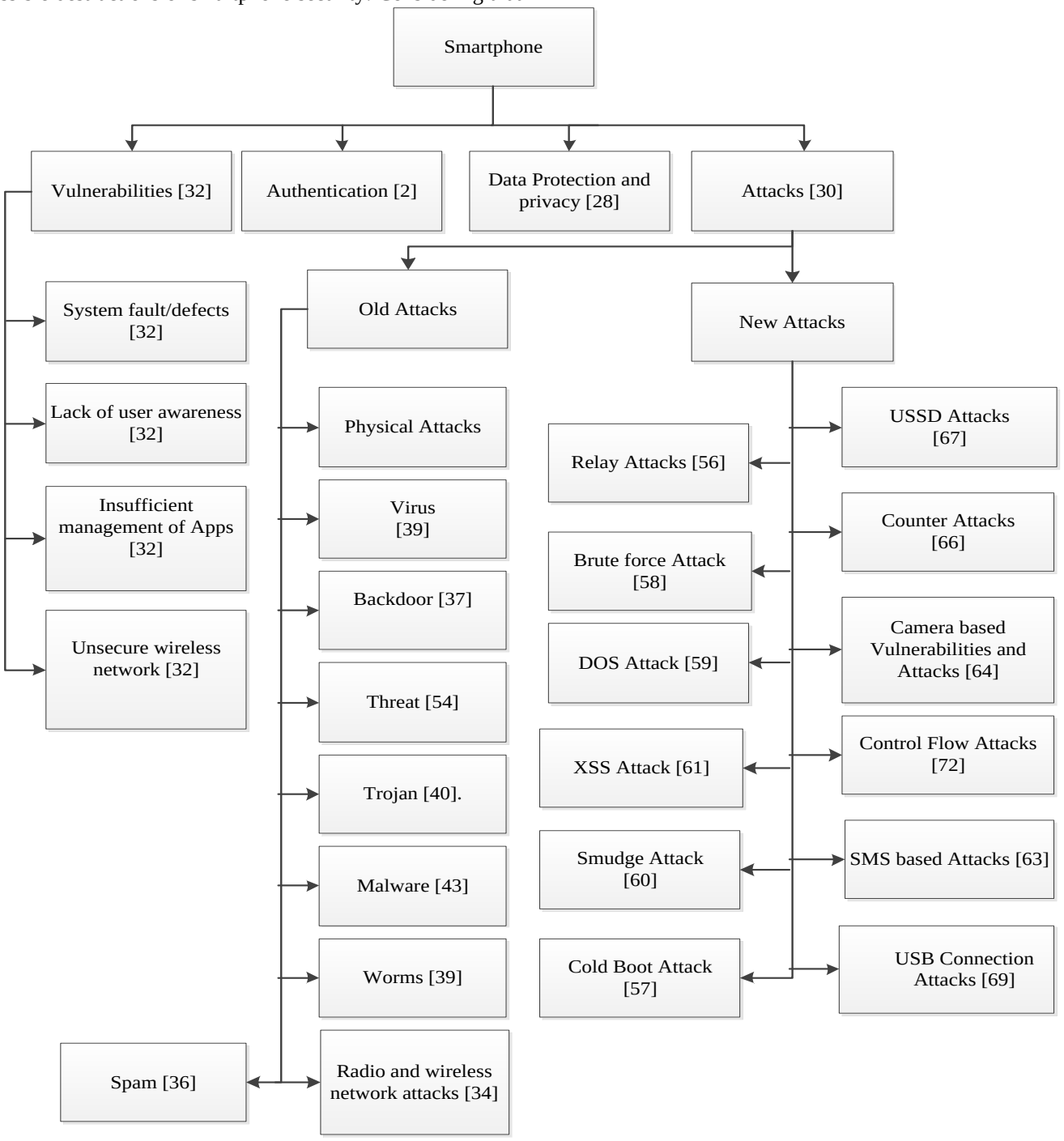


Fig. 1. Categorization of Smartphone Problems

A. Authentication in Smartphones

Authentication could be achieved using one of the following three methods. The first one is to use what users knows such as PIN or password. The second method is which users have certain code such as a token. The third method is commonly known as biometric. After introducing the general architecture of smartphone and its main parts or assets, we classify smartphone's security threats and vulnerabilities. In the study [2] the authors proposed a hierarchy of security

framework, consists of hardware, operating system application user data, communication as security.

Wei-Han proposed a multi-sensors-based system for smartphones to achieve the implicit authentication. The system incessantly learns the user's behavior patterns and setting by allowing the user to use a phone without disturbing the user's actions. This approach also has the capability to update user model. The experiment shows that the efficiency of this model only requires 10 seconds to run the model, 20 seconds to detect abnormal or fake request. In this model, the level of accuracy achieved can reach up to 90-95% [22].

Zahid *et al.* [23], proposed the user identification system to monitor the mobile phone key users to distinguish authentic consumers from quacks dynamically. The authors used custom data set of 25 users to point out the suggested system. That gives the fault rate lesser than 2% after detection mode, and the election of nearly zero after PIN authentication. They also connected their approach with five state-of-the-art procedures existing to identify basic user keystroke.

Authors in [24], suggested TAP *(Typing authentication and protection),* a virtual key based on a typing system for smart devices. There are two steps to improve the security of mobile by TAP, first is the login stage and second is the post login stage. In the first phase, TAP controls biometric information and hand morphology to secure the user's identity. In the second phase, TAP controls the dynamic behavior of the TAP Virtual user key. The experiments demonstrated that TAP preserves security and usability for the smartphone devices.

Chine-Cheng *et al.* have suggested the non-intrusive authentication method that uses the collected information from the orientation sensor of mobile devices. It is a new tactic that is operated by user's smartphone in its own unique way, and orientation sensor captures this type of behavioral biometrics. They use stepwise linear regression to select the feature set for each user. For classification, they used the k-nearest neighbor algorithm. The experimental results show equal error rate about 6.85% in method suggested. Their authentication model satisfies the performance that varies 3 to 8 different end users. The authors conclude that the non-intrusive mechanism can be used with the intrusive mechanism. For example, PIN or Password can be used with the biometric (finger-prints) to increase the security of smartphones [25].

Morris worked on combining traits from three different methods, i.e., biometrics, hand-written signature, face, and modalities of speech. He reported the authentication accuracy of a mobile device that would have been acceptable with a wide range of applications [26]. According to Gobbo *et al.* [27], SIM allows the user to access the network; make the user registration and authentication devices. Without a valid SIM module and a successful verification, mobile devices do not have access, so all the traffic on mobile infrastructure cannot inject. Firstly, SIM that enables the collection of resources is needed to launch an attack without disrupting users and risk found; secondly, the use of devices that are not in ownership of naive users can take part in the attacks as a botnet network nodes.

B. Data Protection and Privacy

Boshmaf *et al.* [28], address the problem of data protection from user-centered perspective and analyzed the user's need for data protection for smartphones systems. The authors outlined the types of data that users want to protect; they also investigated the practices of current users in the protection of such data and show how the security requirements vary across different types of data. They report the results of an exploratory study of the user in which 22 participants were interviewed. Overall, it was found that users want to protect the data on their smartphones, but find it inconvenient to do it in practice, by using the available solutions today.

Muslokhlove presents the problems of data protection against physical threats and possibility to overcome weak authentication. In that study, user's requirements to data protection are highlighted after interviews and survey studies. Finally, the author concludes that detection malicious data access approaches are not covering enough security although there remain several vulnerabilities but for data traffics these approaches are good. Upgrading the lock screen system for supporting authentication and user's accessibility and provide suitable security will increase the confidence of user and safety of smartphones [29].

Ghosh *et al.* worked in the context of privacy, protection and user data regarding semantic reasoning and user context modeling. In this work, the authors state that the privacy of users and smartphone under this framework are protected using embedded semantic policies that are based on the user's privacy and settings [30]. Kodeswaran *et al.* [31] have shown a framework to execute the privacy policies on smartphones, and to protect the enterprise data. The authors have defined their privacy policies of acceptable information flow on mobile devices. This flow of information depends on the object involved in conforming IPC (Inter-Process Communication) and its data. They have described their framework design which is based on policies for Android platform and have shown the results measuring executed by the framework.

C. Vulnerabilities

There are many several attacks and vulnerabilities in smartphones as shown in Fig. 1. According to [32] Smartphones have many vulnerabilities that can lead to insecurity or be victimized by malicious attackers to create attacks. Smartphones vulnerabilities include the following: System fault/defects, insufficient management of applications, insecure wireless network and lack of user awareness.

- System Fault / Defects

It is inconceivable for a smartphone to avoid both hardware and software defects. Such defects are only reveals after the device usage. Some defects can be observed / identified sooner and some later. The software defect can easily be corrected but the hardware faults may cause irregularities, and can be rectified by changing the hardware or by changing the device architecture. Such defects can be exploited by the attackers to initiate the attacks on smartphones.

- Insufficient Management of Apps

Most distinctive feature of the smart devices is their flexible APIs which are mostly used for application development. However, deficient API management is responsible for many malicious code infections. Thus, the API mismanagement is a main reason for malicious code attacks. APIs are classified into Open APIs, third party application development and control APIs; used to remote maintenance. Controlled APIs have particular higher privileges for updating system, file destruction, and information fetching. If attackers gain the APIs control, could easily initiate attacks and exploit the privileges of the APIs [32] [33].

- Unsecure Wireless Network

In wireless network, we use Wi-Fi technology, Bluetooth, cellular network and GPS to connect with any network or Internet. On any network device hacker can retrieve or fetch the packets on the network. So it is a vulnerability, and we can overcome it by using the encryption/decryption method in communication.

- Lack of User Awareness

User awareness to the security is important all the times especially when the smartphone is connected to the Internet for installing an unknown application or downloading data from insecure sources. There are many application available online that look like a legitimate source, but their save button is linked to some malicious codes. Also, activating wireless and Bluetooth interfaces can be executed secretly. Using protected access 2(WPA2) based on IEEE 802.11i is a new security protocol ensuring that only authorized users can access the network [32].

D. Attacks of Smartphone

Attacks are common in all computing devices and smart devices such as smartphones, tablets, etc., in the coming lines we will explain important attacks to the smartphones. The attacks are classified into two categories:

- *Old Attacks*

In this category, the most common attacks have been discussed. It includes physical attacks, viruses, worms, Trojans, malware, etc.

Physical Attacks: Smartphones and tablets are easily lost or stolen. Then, Sensitive data can be accessed and manipulated directly. Physical attacks also damage fallen or covering harmful disposals.

Radio and Wireless Network Attacks: Because the accessibility of wireless communication intruders can create wireless network attacks, they could be grouped into active attacks (Spoofing, corrupting, blocking and modifying) and passive attacks (sniffing and eavesdropping). Passive eavesdropping, the information is detected by listening to the message communication in the broadcasting wireless medium using malicious nodes. In wireless attackers create a fake Wi-Fi network to connect other users, thus, a common advice for smartphone users is to beware of what networks they are connecting to and using if it appears a fake wireless network; immediately disconnect and it is also a good practice to Switch off Wi-Fi sensors [34].

Jermyn and Zonunz,[35] studied the DoS attacks on the LTE and MAC uplink scheduler that cause several attacks. They state that such attacks depend on the QoSs (Quality of Services) requested by the clients. The authors proved the feasibility of suggested attacks on the Android-based simulator. C. Guo *et al.* [36], warm about the dangers of potential smartphone attacks to telecommunication infrastructure, the damage that can range from invasion of privacy and identity theft to emergency harassment centers that can result in a state of crisis. The authors outline various defensive strategies, many of which require a lot of research. It is also suggested to the system architect to concentrate on Internet insecurity in bringing new hardware to the Internet.

Backdoor: Backdoor accepts attackers to establish a connection with their network while evading detection [37]. Research has revealed many backdoor uses in target attacks. Backdoors result mainly from a system, bug, and revelation of controlled APIs. Some of smartphones come with insufficient authentications, based on these vulnerabilities. Backdoor bypass access to the attacker in a normal security [38]. Example Netcat and Virtual network security.

Virus: Virus infects executable files, boot sectors and normal files such as word processing documents, PDF, etc. The virus makes replication to the file with consuming the capacity of the system. Viruses also give a link to an unknown source like installation software without a request from a user [39]. Cheng and Lu [40], introduce a *virus detection system* and alert system for smartphones. This system detects viruses from the information of communication actions. They study the unusual behavior of the smart device, and develop a *SmartSiren* system and grab the result to show that the developed system avoids viruses effectively with reasonable overhead.

Worms: Worms are the programs that transport their copies from one device to another device with the help of different transport mechanism throughout the network without user interference [39] [41] [42].

Malware: Malware attacks harm smartphones by creating an application and provide it to a user to download that application, but that application is a malware. Malware constitutes a serious security threat that slows down the large scale wireless application development. Sometimes your data can crash once you accept or install malware software [43][44] [45].

Shabtai *et al.* suggested a framework (Host-based Malware detection system) that observes features and events acquired from smartphones and then apply a machine learning anomaly detectors to categorize the normal or abnormal data [46]. Peng [44] provides a study of malware, including the advancement of mobile malware, correlated concepts, and the risk of infection vectors. This article shows that the multiplicity and complication of mobile malware poses a major challenge in containing malicious software modeling.

In this paper, the authors suggested assessment criteria to evaluate the development of smart phone malware. They

provide a comparative analysis of case studies in which the progress malware detection and distribution concept of location data is attempted in the current smartphone platform [47]. E. Gelenbe and R. Lent [48], propose taxonomic malware attack vectors studies to better understand the Android malware; the attacker ways to infect smartphones, and a component of the project responsible for the detailed examination and finding of malware Android that NEMESYS structure. Infrastructure intended understanding and network attacks and smartphones detection.

To examine existing development of malware on smartphone platform and average programmer those have access functionary tools and library of smartphone, research [26][49] suggest specific evaluation criteria measuring the level of security of common OS such as Android, Apple iOS, BlackBerry, Windows phone and Symbian in the term of development of malware, and give comparative analysis and based on the proof of the study. However, this proof would not stop the easiness developing of malware attacks in all smartphone. Finally, they suggested solution against that malware, (a) users to be aware, (b) giving or using saves applications.

Trojan: Trojan is a program which is mostly useful, but it has hidden malicious functionality. The purpose is sneaked into the system without the knowledge of the administrations [43] [50]. Smartphones are becoming more complex and more dominant in providing more functions; growing concern about the opposite of smartphone users security threats. Same software architecture is used by smartphones just like a personal computer; they are susceptible to the same class of security hazards such as viruses, Trojans and worms [51]. Houmansadr *et al.* [51], suggest a *cloud-based smartphone-specific intrusion detection and response engine*, which unceasingly accomplishes a detailed forensics examination on the smartphone to notice any misbehavior. Misbehavior is detected; the suggested engine decides upon and takes optimal response actions to avoid the current attacks.

Spam: Spam is kind of malware attachments can be appended to electronic mail and MMS messages reach to smartphones. Sometime a user opens an attachment at this time smartphone can be infected by malicious codes such as Trojan, worms, etc. which appears as a normal attachment. Attackers manipulate smartphones zombies by sending junk messages and those message used as a door by the attackers to compromise smartphone [36] [52].

Xu and Zhu have studied the possibility of launching attacks and spam with Trojan applications installed by abuse customized notification service. The experimental results are presented and the fact of attacks in four major smartphone platform. Also, the authors present an approach to stealth spam content delivery that can help in identifying application Trojan that ignores the review process of the application in app stores. To maintain the proposed strikes propose design principles Semi-OS-controlled to see notifications, see a safe framework for public view and authentication services to log notifications review notification [53].

Threat: Delac *et al.* [54], show the threats and deeply study the threat mitigation mechanism. They show the attacker centric threat model for smartphones. They evaluated the vectors of attack and strategies and give a security model for two main smartphone Operating system; Android and iOS.

- *New Attacks*

In this category new types of network or system attacks have been discussed. It includes Brute force, DoS, smudge attacks, etc.

Relay Attacks: It involves only future applications on mobile phones. Elements and application access security relays APDU command interface / response network (GSM, UMTS, and Wi-Fi). Attackers can use victims' secured as if they have their physical possession. Relay application can access additional resources (address book, keyboard, etc.) [55]. In article [56], Peer-to-Peer communications in NFC (Near field communication) are being deliberated for a variety of applications with payment. Relay attacks are a threat and can circumvent security measures and encryption/decryption using temporary contracts. The author's contributions in this work include the implementation of practical demonstrations of the first relay attack using NFC mobile platform technology. They show that the attacker using NFC can create a proxy for the development and introduction of the software (without hardware change) of the MID let appropriate for mobile devices. The attack does not involve any code validation and software to be installed on the insurance program. It also uses ordinary, readily accessible APIs such *as JSR 257 and JSR 82*, need for action measures. Such attacks can be controlled intensely using location-based solutions discussed in [56].

Cold Boot Attack: Smartphones and tablets are easily stolen or lost. In paper [57], it is discussed that, this makes them vulnerable to low-grade memory attacks such as *cold-boot* attack using a bus, monitor to keep an eye on the memory bus and *DMA* attacks. The article further describes the *Sentry*, a system that permits applications and operating system modules to stock their code and data on the *System-on-Chip (SoC)* instead of DRAM. They propose the use a special mechanism of ARM-specific was specially intended for embedded systems, but it is still in existing mobile phones, to defend applications and OS in contradiction to a memory subsystem.

Brute Force Attack: Kim [58], proposed a keypad to make the brute force and smudge attacks difficult. This type of keypad increases the time that is required by both brute force and smudge attacks. Keypad time is increased by the formation of random buttons and display delay time.

Denial of Service Attack: Dondyk and Zou [59], proposed a new denial of service oriented attack for the smartphones used by ordinary operators who are not tech savvy. This type of attack which they call the DoS attack, does not prevent future technical perception to use the service through the operation of data management protocol connection to find your smartphone with Wi-Fi antenna. By creating a false eye Internet access Wi-Fi (using devices such as a laptop), the attacker can ask for a smartphone with a Wi-Fi enabled to dismiss the supply of mobile broadband connections that is authorized automatically and link to a bogus Wi-Fi

connection. As a result, it avoids the target smartphone to have any Internet link, unless the dupe can identify the attack and manually disable the Wi-Fi capabilities. They have shown that the most popular smartphones, with iPhone and Android mobiles susceptible to denial of accessibility. To counter these attacks, they propose a new enactment of Internet access authentication protocol to send secret passphrase from authentication server to Internet using a cellular network. Then you try to recover the secret key phrases via Wi-Fi channel that you created to verify the Wi-Fi access point. They have fully evaluated the attack, and defense prototype that runs on Android phones.

Smudge Attack: Gibson [60], explored smudge attacks using oil on the mobile touch screen and captured the smudges. They emphasized on the effect on password pattern of smartphone. They provide a primary study of applying the information learned in a smudge attack to predicting a pattern password.

Cross-Site Scripting (XSS) Attacks: Jin and Hu run the risk of systematic reviews in HTML5 - based mobile application, discovered a new injection code attack, which inherited a cross-site scripting (XSS) attacks (basic cause), but several channels used to insert XSS code. These channels exclusively for mobile devices, including contacts, SMS, bar codes, and MP3 to assess the occurrence of addition code susceptibility in mobile application based on HTML5. They developed a screening tool to analyze the weaknesses of 15,510 applications in Google Play, Phone Gap, 478 applications likely the rate of 2.30% error-positive rate and developed a model called No injection as a cover for the Android hone GAP to protect it from attack [61].

The problem is that HTML5-based malicious code can be inserted into any automated software or application and run. This is the cause of cross-site scripting (XSS) attacks are one of the most common attacks on Web-based applications or programs. Cross-site scripting can only target web application [62].

SMS Based Attack: Attacker can advertise and distribute phishing links via SMS attacks. Text messages can also be used by attackers to feat vulnerabilities [63][64]. Rieck and Stewin [62], study the security of SMS OTP (One-Time password) system architecture and attacks that present a hazard to service learning authentication through the Internet and authorization. They resolute two basic SMS OTP erected on wireless networks and mobile devices have totally dissimilar when SMS OTP is intended and introduced. During this exertion, which showed why SMS OTP system is not safe again? Their results based on proposed mechanisms to ensure SMS OTP against collective attacks and precisely against Trojan.

Hamandi *et al.* [65], examine some of the messaging design verdicts that cause a set of vulnerabilities in the Android operating system, and they show how applications can be built for malware detection to avoid abuse by this vulnerability. These applications appear as a normal application SMS messages and use them fundamental truths to send/receive short messages. Since many operators around the world offer a service that allows users to transfer credit/unit via SMS, cause the misuse of this service to transfer credit illegally. Subsystem “permission”, subsystem "broadcasting receiver," and ordering mechanism to contribute to the establishment of a haven for SMS malware, giving them total control over the sending, receiving and hiding SMS messages. Therefore, the application hides the malicious confirmation from telecom operators that can arise after the transaction for credit transfer. Such subsystems allow users to stream and balance malware attacks that have the potential to cause damage to a large number of users and telecommunications operators. The application has been shown in local control and successfully passed the standard inspection procedure aimed to catch malware. A set of possible solutions is also presented to decrease the risk of such attacks.

Counter Attacks / Escalation Attack: In [66], authors proposed a scheme for detection and prevention that protects Android with features like counter-attacks or escalation attack that attempt to gain full access to all data. These systems monitor the proposed scheme essentially used to call for the application process. If the call system is called by special components of the Android system in normal operation, the regime prevented it from performing. The scheme can detect and block new and unknown malware.

USSD Attacks: USSD (Unstructured Supplementary Service Data) is a protocol used by operators of www(world wide web) to run specific functionality between users and operators [67], examples such as functions including credit check and credit of USSD, USSD can send a prepaid callback, Mobile-Money services. The USSD contains following components: Main Activity, USSD interceptor Service, Boot service and Permission testing.

Hamdani, and Elhajj [68] identified and evaluated two types of Android smartphone based attacks. The first is done by sending an SMS in the background and push notifications network to steal customer credit. Also, they show how the SPM security structure in Android has grown, but they showed how the attack can still be performed. The second attack using the mobile dialler application using the USSD protocol on the target user background.

USB Connection Attacks: Decker and Zúquete [69], exposed serious weaknesses in some private provider’s Android operating system. They described the proof-of-concept to them, which can be used to explore the implications of vulnerability, such as root access. For advanced features are intended for use by suppliers of computer applications to configure and control the device, developed on purpose and with the intention stated. In their observation, the installation of such "features" must be at least possible released to the user, so they recognize the risks of an unprotected USB connection.

Camera based Vulnerabilities and Attacks: Currently, almost all smartphones have features like camera and touchscreen. These functions can lead to attacks on smartphones. Users change device through third party applications from the "app stores" or traditional websites. Source application is a problem, so users are constantly at risk of installing malicious programs that steal personal information or gain root access to their device [64][70].

In article [71], it is figured out the weaknesses associated with Android phones are also for those versatile and sound applications. The authors talked about pieces of spy cam (use of smartphone as spy cam), can play for their attack or gain customers. The authors argue that they found some spy camera forward attacks, including attacks related to continuous monitoring, remote control and two pass-code once led to the raid. Meanwhile, they suggested a plan to ensure a strong guard mobile phone spy cam all this aggression. They explore the possibility of conducting espionage attack (grab information used to launch a successful attack).

Control Flow Attacks: Runtime attacks and control flow (such as code injection or return-oriented programming) is one of the biggest threats to software programs [72] [73]. These attacks are common and have been recently applied to smartphone applications that are downloaded by many users. Davi *et al.* presents a mobile CFI (MoCFI) framework that provides a general countermeasure in contradiction of control flow attacks on smartphone platform by CFI. A typical smartphone that is involved because of two different architectures ARM and Intel. The authors prove that MoCFI is efficient for all smartphone OSes excluding iOS [74].

III. Performance Comparison

In this section, we review present solutions, settled to avoid different types of smartphone threats, attacks and vulnerabilities. To respond to the increasing number of attacks and malware with the vulnerabilities on smartphones, we have several solutions for the problems. So, we show all attacks and their solutions in tabular form. Table 2 shows the old attacks, causes of old attacks and their suggested solutions. Similarly, Table 3 shows a new form of attacks, the cause of these type of attacks and their solutions.

In article [28], 22 participants were interviewed and it was found that each participant wanted to protect data. For data protection and privacy many of the researchers proposed various solutions some of them are discussed here; In Musklokhlove *et al.* [29] authors gave solution for data protection and authentication. They purpose detection malicious data access approach for data protection and upgrade the lock screen system for smartphone authentication. The [30] and [31], articles provide a framework to execute privacy policies to protect user data and enterprise data.

In [18] and [19], it is discussed that the growth of selling smartphone is increasing gradually. In 2014, the shipment values were as the following with respect to Mobile Phone Operating System, Android phone: 950.5 million, iOS: 179.5 million, Windows Phone: 47 million, BlackBerry: 11.9 million, and other (Symbian, etc.): 15.1 million. And in 2018, their market share will increase 11.5%. Fig. 2 show the estimated market share and shipment values of 2018 with the help of 2014's data of shipment and market share.

According to report [75], they said that by the end of 2015, there will probably be more smartphones than people and in 2016 there could be 10 billion smartphones. So, it can be true if sale or shipment of smartphones could gradually increase. Because many peoples may has more than one device.

Table 4 shows the distribution of new mobile malware by their types (first is Installation program and the second is new mobile malicious programs) from the Quarter 4 (Q4) 2014 to Quarter 3 (Q3) 2015. The statistical data obtained from sources (Kaspersky Lab) [76] & [77] have been computed and represented in Table 4.

TABLE I. Distribution of New Mobile Malware (Q4 2014 to Q3 2015) [76] and [77]

Time Period	Mobile Malware Type	
	Installation Package	New mobile malicious program
Q4 2014	65443	30849
Q1 2015	147835	103072
Q2 2015	1048129	291887
Q3 2015	1583094	323374

The Q denotes quarter at x-axis in Fig. 3. Q4 2014 to Q3 2015, the mobile malware increase gradually. This shows in Q4 of 2015 the malware will increase. So, we can say that mobile malware will increase gradually till Q4 2020 shown in Fig. 3. But it is possible that the graph is stable or decrease if any control mechanism will introduce. This estimation is shown in Fig. 3. The middle line shows the stability of the malware and the bottom line is showing the decreasing in malware if a control mechanism is introduced.

These estimations show that due to increasing growth of selling smartphones, malware writers develop a lot of malware software that causes the security threats in smartphones.

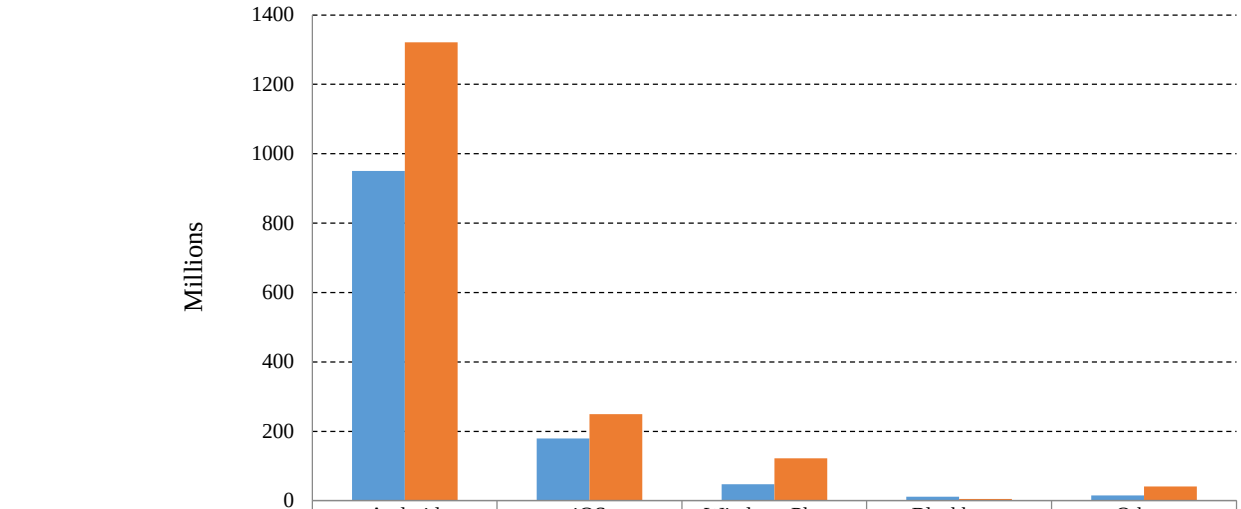
TABLE II. OLD ATTACKS, THEIR VULNERABILITIES AND SOLUTIONS

Attack Name	Vulnerabilities	Solution	Impact	Ref.
Physical Attack	System defects / fault.	Re-manufacturing whether is software or hardware.	Weak the security of mobile phone. Abnormal behavior.	
	Insufficient APIS Management.	Use trusted application from sources.	Malicious code can infect user's data or files.	[33]
Radio Wireless attack	Eavesdropping sniffing and spoof computing blocking.	Suddenly disconnect from the wireless network.	Data can be hacked easily. Weaken computer security.	[34]
	Insecure Wireless network.	Only use trusted network. Using encryption / decryption method to secure communication.	Information can be hacked during communication.	[34]
Backdoor	System bugs and disclosure.	Update your device and install strong antivirus.	Security of smartphone can weak. A backdoor for viruses can be made.	[38]
Virus	Target finding, replication file with unknown source.	Install update Antivirus in your system.	Abnormal behavior of application. Information or applications may be corrupted.	[39], [40]
Worms	Transferring information. Transfer malicious program.	Use updated Antivirus.	Can create the backdoor for hacker. Intertwined with the system files.	[39]
Malware	Downloading file from interested resources.	Use updated anti-virus, install malware prevent software. Use host-based malware detection system, use safe application.	Disturb computer operations. Gather sensitive information.	[43], [44], [46], [26]
Trojan	Downloading Apps from untrusted resources. Hidden malicious functionality.	Smart phone specific intrusion detection system. Use anti-virus.	Disturb computer operations. Gather sensitive information.	[78],[51]
Spam	Any attachment with malicious code transfer via E-mail or MMS. Attacker can advertise phishing links.	Avoid opening these types of emails and MMS. Only taking authentic services and using authentic application. Avoiding responding to any emails that you never asked for.	Fills your Inbox with number of ridiculous emails. Degrades your Internet speed to a great extent. Steals useful information like your details on you Contact list. Alters your search results on any search engine.	[53]
Threat	Spoofing, Information disclosure.	Use CTM (Cyber threat management) software.	Corrupt data. Weaken computer security. Provide back doors into protected networked.	[79]

TABLE III. NEW ATTACKS, THEIR VULNERABILITIES AND SOLUTIONS

Attack Name	Vulnerabilities	Solution	Impact	Ref.
Relay Attack	Insecure network environment. Use of unauthentic proxy service.	Use secure network and trusted proxy application.	Information hacked during communication.	[56]

Cold Boot Attack	Unauthorized access to RAM and encryption / decryption key of system	Use a system that store code and data on the SOC (System on chip) instead of RAM i.e. Sentry. Use powerful encryption decryption method	Encryption key may be hacked. Weak data security.	[57]
Brute Force Attack	Try again and again to unlock phone using many combination and no limit to prevent from hacking.	Set a limit for try again and again to unlock device and display time delay.	Password cracked. Slowing the CPU speed.	[58]
Smudge Attack	By keep touch screen dirty or using oily hands.	Keep clean and clear screen and use clean hand to operate device.	Easily guess the pattern password. Data unsecure.	[60]
Denial of Service Attack	By using other device dismiss the supply of mobile broadband connection. Link to bogus Wi-Fi connection	Use internet access authentication protocol.	Busy the network. Busy smartphone and block other services.	[59]
XSS Attack	HTML 5 based malicious code inserted into an application or software.	Use popular and authentic apps. Use screening tool to check the weakness of the apps.	Smartphone infected by inject malicious code via HTML page or any other untrusted script. Cause of hacking information or provide backdoor.	[62], [61]
SMS based Attack	Attacker can advertise phishing links.	Device can protect by setting up the Message settings, or to dis-allow auto receiving MMS or text.	Sensitive information can be fetch.	[63], [64], [80]
USSD Attack	Blue Jacking, Bluesnarfing and unknown coming calls.	Use Anomaly based Intrusion detection system	Personal data can be fetched. Cause the damage on the cell phones.	[68]
USB Connection Attack	Root access, enable ADB(open command tool and avail both developer and attacker)	Use apparently inoffensive smartphone charging station	Sensitive information can be fetch easily. Any malware can be injected easily.	[69]
ABD Attack	Open command tool and avail both developer and attacker	Backward slicing. Static analyzer and string analyzer.	Sensitive information can be fetch easily.	.[2]
Camera based attacks	Malicious program, unauthentic source and etc. Use camera of smart phone as spy cam by malicious program	Spy camera could support. Implement effective fine Grained access	Weak the smartphone security. Can fetch data or information.	[64], [71]
Control Flow attacks	Code injection, data over flow in Memory	Use Mobile control flow integrity framework	Can be exploited to snip the user's SMS or contacts database, to open a remote reverse shell. Exploiting memory corruption.	[74]



	Android	iOS	Windows Phone	Blackberry	Others
2014 Shipment Values (Million)	950.5	179.5	47	11.9	15.1
2018 Shipment values (Million)	1321.1	249.6	121.8	5.3	40.7

Fig. 2. Smartphone market share and shipment by OS in 2014 and its estimation for 2018 [18] & [19]

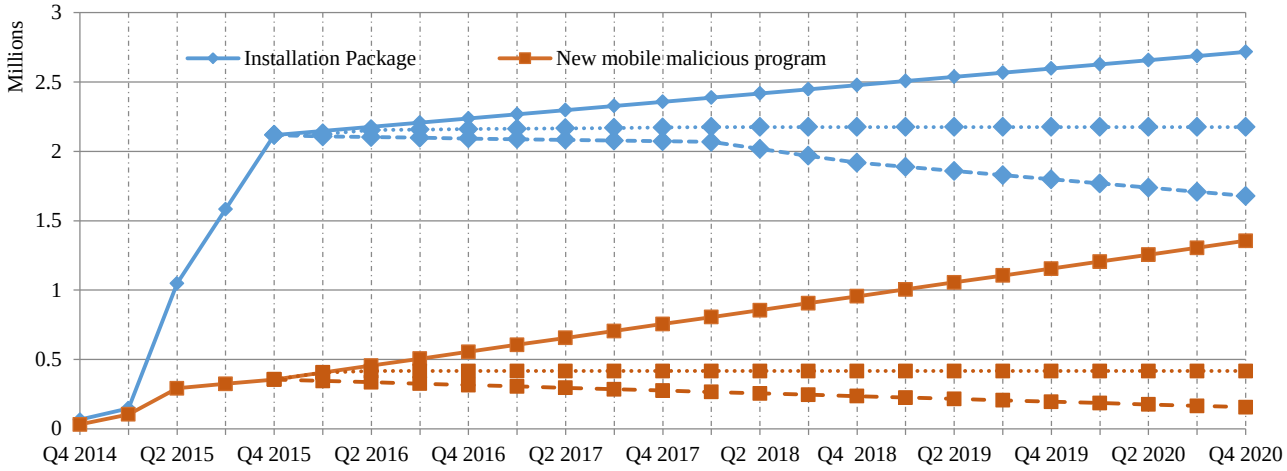


Fig. 3. Estimated Mobile malware from Q4 2014 - Q4 2020

IV. Discussion

Authentication, data privacy, vulnerabilities (which cause the attacks) and attacks are the major open issues of security. Authentication problem is the major part of security breach. An appropriate solution to authentication problem can overcome many authentication problems and saves smartphones from security breach. All users want to protect their data. So, Data Privacy is one of the biggest concerns to the smart phone users. Thus, the data privacy issues, can most of times could be addressed by using trusted sites and applications. Most of the smartphone attacks occur due to vulnerabilities. If the vulnerabilities are minimized, it can save the smartphones from most of the attacks. But in rapidly growing field where development occurs at large scale it is hard to achive100% security, but the careful design and development processes lead to more secure smartphones.

Number of smartphones is increasing rapidly. The reason behind this increment is the frequent technological changes and evolution. But with this technological evolution, more malware attacks are being launched. If we look upon the Kaspersky Laboratories reports regarding these attacks, we come to know that number of malware attacks is increasing every year, which is also included in this review paper. So, we should neither be satisfied upon the increasing in number of smartphone sales, nor it should be merely lookup for solution by the developer against the malware attacks being launched. But manufacturers as well as developers have to look around the reason behind these attacks, launched for smartphones which are obviously because of the loophole present in the architecture and software of the smartphone being provided by the manufacturer and the software developer.

The future belongs to IoT (Internet of Things); technology where all the devices remain online and interconnected. So, almost each routine gadget would be controlled by smartphone

via IoT. Which includes electronic devices, machines, vehicles, security based entrances, etc. So, this will cause a lot of issues regarding smartphones such as battery drainage issue, performance issue and security issues regarding not only data privacy but also illegal access to the personal devices via IoT. So, it is required to have a smartphone that used for IoT, must have best battery consumption, efficient processing and maximum security. So that we would be able to achieve maximum benefits from IoT. As we know that we don't have a mechanism for complete security regarding smartphone. We can't say that our data privacy and access is completely safe and sound. So that manufacturers as well as developers require building and presenting a mechanism that provides maximum security.

The purpose for writing this review is to provide a holistic account of smartphone vulnerabilities and problems and to look at various possible solutions suggested in the literature. These solutions and problems have been collected from review of previous researches.

V. Conclusion

Smartphones are the multipurpose handheld devices that contain a lot of third-party applications that extend the functionality of the device. With the quick production of smartphones prepared with many features such as several connectivity links and sensors, the mobile malware are growing. The smartphone environment is different from the PC environment. Similarly, the solutions to prevent the infections and diffusion of malicious code in smartphone are different from PC or other computer devices. Smartphones have insufficient resources, including power (battery) and processing unit. Increasing the capabilities of the smartphone, these features can be misused by attackers, as different types of links, sensors, services and user's secrecy.

In this work, at first, we discussed the current authentication problems, data protection and privacy problems. We investigated the vulnerabilities in smartphones and attacks that can occur in smartphones. Secondly, we have characterized identified attacks in contradiction of smartphones, concentrating on why attacks occur and what are their effects on smartphones. Finally, we have studied existing security results to prevent smartphones from infections, malicious codes and intruder's attacks.

References

[1] N. Yildirim, R. Das, and A. Varol, "A Research on Software Security Vulnerabilities of New Generation Smart Mobile Phones," in 2nd International Symposium on Digital Forensics and Security (ISDFS'14), 2014, pp. 6–16.

[2] A. Agrawal and A. Patidar, "Smart Authentication for Smart Phones," Citeseer, vol. 5 (4), pp. 4839–4843, 2014.

[3] P. Schulz and D. Plohmann, "Android security-common attack vectors," in Rheinische Friedrich-Wilhelms-Universitat Bönn, Germany, Tech. Rep, 2012.

[4] C. Pettey and R. van der Meulen, "Gartner says worldwide sales of mobile phones declined 3 percent in third quarter of 2012; smartphone sales increased 47 percent," Gartner),[Online], 2012.

[5] R. van der Meulen, "Gartner says 821 million smart devices will be purchased worldwide in 2012; sales to rise to 1.2 billion in 2013," 2012.

[6] R. van der Meulen and J. Rivera, "Gartner Says Smartphone Sales Grew 46.5 Per Cent in Second Quarter of 2013 and Exceeded Feature Phone Sales for First Time," 2013.

[7] "Global Smartphone Use Continues to Climb, Studies Show." [Online]. Available: http://www.eweek.com/mobile/global-smartphone-use-continues-to-climb-studies-show.html. [Accessed: 30-Oct-2015].

[8] H. Luo, G. He, X. Lin, and X. Shen, "Towards hierarchical security framework for smartphones," in Communications in China (ICCC), 2012 1st IEEE International Conference on. IEEE, 2012, pp. 69–72.

[9] M. Ahmad and N. Musa, "Comparison between android and iOS Operating System in terms of security," in Information Technology in Asia (CITA), 2013 8th International Conference on. IEEE, 2013, pp. 1–4.

[10] M. Goadrich and M. Rogers, "Smart smartphone development: iOS versus Android," in Proceedings of the 42nd ACM technical symposium on Computer science education. ACM, 2011, pp. 607–612.

[11] L. Ma, L. Gu, and J. Wang, "Research and Development of Mobile Application for Android Platform," Int. J. Multimed. Ubiquitous Eng., vol. 9, no. 4, pp. 187–198, 2014.

[12] J. Liu and J. Yu, "Research on Development of Android Applications," in Fourth International Conference on Intelligent Networks and Intelligent Systems. IEEE, 2011, pp. 69–72.

[13] T. Grønli and J. Hansen, "Mobile application platform heterogeneity: Android vs Windows Phone vs iOS vs Firefox OS," in Advanced Information Networking and Applications (AINA), 2014 IEEE 28th International Conference on. IEEE, 2014, pp. 635–641.

[14] D. Tilson, C. Sørensen, and K. Lyytinen, "Change and control paradoxes in mobile infrastructure innovation: the Android and iOS mobile operating systems cases," in System Science (HICSS), 2012 45th Hawaii International Conference on. IEEE, 2012, pp. 1324–1333.

[15] V. Remenar, S. Husnjak, and D. Peraković, "Research of Security Threats in the Use of Modern Terminal Devices," in 23rd International DAAAM Symposium Intelligent Manufacturing & Automation: Focus on Sustainability, 2012.

[16] A. Maji and K. Hao, "Characterizing failures in mobile oses: A case study with android and symbian," in Software Reliability Engineering (ISSRE), 2010 IEEE 21st International Symposium on. IEEE, 2010, pp. 249–258.

[17] O. Riva and C. Qin, "Progressive Authentication: Deciding When to Authenticate on Mobile Phones.," in Presented as part of the 21st USENIX Security Symposium (USENIX Security 12)., 2012, pp. 301–316.

[18] F. Al-Qershi, "Android vs. iOS: The security battle," in Computer Applications and Information Systems (WCCAIS), 2014 World Congress on. IEEE, 2013, pp. 1–8.

[19] D. Padmavathi and M. Shanmugapriya, "A survey of attacks, security mechanisms and challenges in wireless sensor networks," arXiv Prepr. arXiv0909.0576, vol. 4, pp. 1–9, 2009.

[20] M. A. Dar and J. Parvez, "Smartphone operating systems: Evaluation & enhancements," in 2014 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT), 2014, pp. 734–738.

[21] J. H. Choi and H.-J. Lee, "Facets of simplicity for the smartphone interface: A structural model," Int. J. Hum. Comput. Stud., vol. 70, no. 2, pp. 129–142, Feb. 2012.

[22] W. Lee and R. Lee, "Multi-sensor authentication to improve smartphone security," in Conference on Information Systems Security and Privacy., 2015, pp. 1–11.

[23] S. Zahid, M. Shahzad, S. Khayam, and M. Farooq, "Keystroke-based user identification on smart phones," in Recent Advances in Intrusion Detection. Springer Berlin Heidelberg, 2009, pp. 224–243.

[24] T. Feng, X. Zhao, B. Carbunar, and W. Shi, "Continuous Mobile Authentication Using Virtual Key Typing Biometrics," in 2013 12th IEEE International Conference on Trust, Security and Privacy in Computing and Communications, 2013, pp. 1547–1552.

[25] C.-C. Lin, C.-C. Chang, D. Liang, and C.-H. Yang, "A New Non-Intrusive Authentication Method Based on the Orientation Sensor for

Smartphone Users," in 2012 IEEE Sixth International Conference on Software Security and Reliability, 2012, pp. 245–252.

[26] A. Morris, "Multimodal person authentication on a smartphone under realistic conditions," in Defense and Security Symposium. International Society for Optics and Photonics, 2006, p. 62500D–62500D.

[27] N. Gobbo, A. Merlo, and M. Migliardi, "A denial of service attack to GSM networks via attach procedure," Secur. Eng. Intell. Informatics. Springer Berlin Heidelb., vol. 8128, pp. 361–376, 2013.

[28] I. Muslukhov, Y. Boshmaf, C. Kuo, J. Lester, and K. Beznosov, "Understanding Users' Requirements for Data Protection in Smartphones," in 2012 IEEE 28th International Conference on Data Engineering Workshops, 2012, pp. 228–235.

[29] I. Muslukhov, "Survey: Data Protection in Smartphones Against Physical Threats," in Term Project Papers on Mobile Security. University of British Columbia., 2012.

[30] D. Ghosh, A. Joshi, T. Finin, and P. Jagtap, "Privacy Control in Smart Phones Using Semantically Rich Reasoning and Context Modeling," in 2012 IEEE Symposium on Security and Privacy Workshops, 2012, pp. 82–85.

[31] P. Kodeswaran, V. Nandakumar, S. Kapoor, P. Kamaraju, A. Joshi, and S. Mukherjea, "Securing Enterprise Data on Smartphones Using Run Time Information Flow Control," in 2012 IEEE 13th International Conference on Mobile Data Management, 2012, pp. 300–305.

[32] R. Prodanovic and D. Simic, "A Survey of Wireless Security," J. Comput. Inf. Technol., vol. 15, no. 3, p. 237, Sep. 2007.

[33] A. Kataria, T. Anjali, and R. Venkat, "Quantifying smartphone vulnerabilities," in 2014 International Conference on Signal Processing and Integrated Networks (SPIN), 2014, pp. 645–649.

[34] K. Mandke, H. Nam, and L. Yerramneni, "The evolution of ultra wide band radio for wireless personal area networks," Spectrum, vol. 3, pp. 22–32, 2003.

[35] J. Jermyn, G. Salles-Loustau, and S. Zonouz, "An Analysis of DoS Attack Strategies Against the LTE RAN," J. Cyber Secur., vol. 3, pp. 159–180, 2014.

[36] C. Guo, H. Wang, and W. Zhu, "Smart-Phone Attacks and Defenses," Citeseer HotNets III., 2004.

[37] O. Ugus, D. Westhoff, and H. Rajasekaran, "A leaky bucket called smartphone," in 2012 IEEE International Conference on Pervasive Computing and Communications Workshops, 2012, pp. 374–380.

[38] M. Durairaj and A. Manimaran, "A Study on Security Issues in Cloud Based E-Learning," Indian Journal of Science and Technology, vol. 8, no. 8. pp. 757–765, 01-Apr-2015.

[39] M. La Polla, F. Martinelli, and D. Sgandurra, "A Survey on Security for Mobile Devices," IEEE Commun. Surv. Tutorials, vol. 15, no. 1, pp. 446–471, 2013.

[40] J. Cheng, S. H. Y. Wong, H. Yang, and S. Lu, "SmartSiren," in Proceedings of the 5th international conference on Mobile systems, applications and services - MobiSys '07, 2007, p. 258.

[41] S. Peng, M. Wu, G. Wang, and S. Yu, "Propagation model of smartphone worms based on semi-Markov process and social relationship graph," Comput. Secur., vol. 44, pp. 92–103, Jul. 2014.

[42] D. Liu, N. Zhang, and K. Hu, "A Survey on Smartphone Security," Appl. Mech. Mater., vol. 347–350, pp. 3861–3865, Aug. 2013.

[43] M. H. R. Khouzani, S. Sarkar, and E. Altman, "Maximum Damage Malware Attack in Mobile Wireless Networks," IEEE/ACM Trans. Netw., vol. 20, no. 5, pp. 1347–1360, Oct. 2012.

[44] S. C. Peng, "A Survey on Malware Containment Models in Smartphones," Appl. Mech. Mater., vol. 263–266, pp. 3005–3011, Dec. 2012.

[45] I. Burguera, U. Zurutuza, and S. Nadjm-Tehrani, "Crowdroid: behavior-based malware detection system for Android," in Proceedings of the 1st ACM workshop on Security and privacy in smartphones and mobile devices - SPSM '11, 2011, p. 15.

[46] A. Shabtai, U. Kanonov, Y. Elovici, C. Glezer, and Y. Weiss, "'Andromaly': a behavioral malware detection framework for android devices," J. Intell. Inf. Syst., vol. 38, no. 1, pp. 161–190, Jan. 2011.

[47] A. Mylonas, S. Dritsas, B. Tsoumas, and D. Gritzalis, "Smartphone security evaluation The malware attack case," in Security and Cryptography (SECRYPT), 2011 Proceedings of the International Conference on, 2011, pp. 25–36.

[48] E. Gelenbe and R. Lent, Eds., Information Sciences and Systems 2013, vol. 264. Cham: Springer International Publishing, 2013.

[49] A. Mylonas and S. Dritsas, "Smartphone security evaluation The malware attack case," in Security and Cryptography (SECRYPT), 2011 Proceedings of the International Conference on. IEEE, 2011, pp. 25–36.

[50] Z. Xu, K. Bai, and S. Zhu, "inferring user inputs on smartphone touchscreens using on-board motion sensors," in Proceedings of the fifth ACM conference on Security and Privacy in Wireless and Mobile Networks - WISEC '12, 2012, p. 113.

[51] A. Houmansadr, S. A. Zonouz, and R. Berthier, "A cloud-based intrusion detection and response system for mobile phones," in 2011 IEEE/IFIP 41st International Conference on Dependable Systems and Networks Workshops (DSN-W), 2011, pp. 31–32.

[52] M. J. Smith and G. Salvendy, Eds., "A Practical Analysis of Smartphone Security," in Symposium on Human Interface 2011, Held as Part of HCI International 2011, Orlando, FL, USA, July 9-14, 2011, Proceedings, Part I, 2011, pp. 311–320.

[53] Z. Xu and S. Zhu, "Abusing notification services on smartphones for phishing and spamming," in Proceedings of the 6th USENIX conference on Offensive Technologies. USENIX Association, 2012, pp. 1–1.

[54] G. Delac, M. Silic, and J. Krolo, "Emerging security threats for mobile platforms," in MIPRO, 2011 Proceedings of the 34th International Convention. IEEE, 2011, pp. 1468–1473.

[55] M. Roland, J. Langer, and J. Scharinger, "Relay attacks on secure element-enabled mobile devices," in Information Security and Privacy Research. Springer Berlin Heidelberg, 2012, pp. 1–12.

[56] S. B. Ors Yalcin, Ed., Radio Frequency Identification: Security and Privacy Issues, vol. 6370. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010.

[57] P. Colp, J. Zhang, J. Gleeson, S. Suneja, E. de Lara, H. Raj, S. Saroiu, and A. Wolman, "Protecting Data on Smartphones and Tablets from Memory Attacks," ACM SIGARCH Comput. Archit. News, vol. 43, no. 1, pp. 177–189, Mar. 2015.

[58] I. Kim, "Keypad against brute force attacks on smartphones," IET Inf. Secur., vol. 6, no. 2, p. 71, Jun. 2012.

[59] E. Dondyk and C. C. Zou, "Denial of convenience attack to smartphones using a fake Wi-Fi access point," in 2013 IEEE 10th Consumer Communications and Networking Conference (CCNC), 2013, pp. 164–170.

[60] A. J. Aviv, K. Gibson, E. Mossop, M. Blaze, and J. M. Smith, "Smudge attacks on smartphone touch screens," pp. 1–7, Aug. 2010.

[61] X. Jin, X. Hu, K. Ying, W. Du, H. Yin, and G. Peri, "Code injection attacks on HTML5-based mobile apps: Characterization, detection and mitigation," in Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security. ACM, 2014, pp. 66–77.

[62] K. Rieck, P. Stewin, and J.-P. Seifert, Eds., Detection of Intrusions and Malware, and Vulnerability Assessment, vol. 7967. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013.

[63] "A Survey Of Mobile Device Security: Threats, Vulnerabilities and Defenses | A Few Guys Coding Blog," in University of Colorado at Colorado Springs, 2011.

[64] M. E. S. Amravati, "A Review on Camera Based Attacks on Android Smart Phones," Int. J. Comput. Sci. Technol., vol. 6, no. 1, pp. 88–92, 2015.

[65] K. Hamandi, A. Chehab, I. H. Elhajj, and A. Kayssi, "Android SMS Malware: Vulnerability and Mitigation," in 2013 27th International Conference on Advanced Information Networking and Applications Workshops, 2013, pp. 1004–1009.

[66] H. Lee, D. Kim, M. Park, and S. Cho, "Protecting data on android platform against privilege escalation attack," in International Journal of Computer Mathematics, 2014, pp. 1–14.

[67] S. Arzt, S. Huber, S. Rasthofer, and E. Bodden, "Denial-of-App Attack: Inhibiting the Installation of Android Apps on Stock Phones," in

Proceedings of the 4th ACM Workshop on Security and Privacy in Smartphones & Mobile Devices - SPSM '14, 2014, pp. 21–26.

[68] K. Hamandi, A. Salman, and I. Elhajj, "Messaging Attacks on Android: Vulnerabilities and Intrusion Detection," in Mobile Information Systems 2015, 2015.

[69] B. De Decker and A. Zúquete, Eds., Communications and Multimedia Security, vol. 8735. Berlin, Heidelberg: Springer Berlin Heidelberg, 2014.

[70] M. Bartere and M. Pore, "Preventions and Features of Camera Based Attacks on Smart Phones," Int. J. Innov. Res. Comput. Commun. Eng., vol. 3, no. 5, pp. 4846–4853, 2015.

[71] R. Mayavan and A. Selvan, "Security Dangers to Versatile Interactive Media Applications: Cam Based Assaults on Versatile Telephones," 2015.

[72] "Control-Flow Integrity." [Online]. Available: https://www.trust.informatik.tu-darmstadt.de/research/projects/current-projects/control-flow-integrity/. [Accessed: 20-Feb-2016].

[73] "Control Flow Integrity: Talks." [Online]. Available: https://talks.cs.umd.edu/talks/71. [Accessed: 20-Feb-2016].

[74] L. Davi, A. Dmitrienko, M. Egele, and T. Fischer, "MoCFI: A Framework to Mitigate Control-Flow Attacks on Smartphones.," in 19th Annual Network & Distributed System Security Symposium (NDSS), 2012.

[75] "2015 Mobile Threat Report - The Rise of Mobile Malware." [Online]. Available: https://securityintelligence.com/events/the-current-state-of-mobile-threats/. [Accessed: 11-Dec-2015].

[76] "IT threat evolution in Q1 2015 - Securelist." [Online]. Available: https://securelist.com/analysis/quarterly-malware-reports/69872/it-threat-evolution-in-q1-2015/. [Accessed: 11-Dec-2015].

[77] "IT threat evolution in Q3 2015." [Online]. Available: http://www.tsecurity.de/it-security-sicherheit/malware-trojaner-viren/28810-it-threat-evolution-in-q3-2015. [Accessed: 11-Dec-2015].

[78] J. Jamaluddin, N. Zotou, R. Edwards, and P. Coulton, "Mobile phone vulnerabilities: a new generation of malware," in IEEE International Symposium on Consumer Electronics, 2004, 2004, pp. 199–202.

[79] "Threats in computer security." [Online]. Available: https://en.wikipedia.org/wiki/Threat_(computer). [Accessed: 04-Dec-2015].

[80] A. Pore and M. Bartere, "A Review on Camera Based Attacks on Android Smart Phones Anushree Pore," vol. 6, no. 1, Feb. 2015.

Hex Symbols Algorithm for Anti-Forensic Artifacts on Android Devices

Somyia M. Abu Asbeh
Software Engineering Dept.
Princess Sumaya University for Technology
Amman, Jordan

Hamza A. Al-Sewadi
Computer Sience Dept.
Prince Sumaya University for Technology
Amman, Jordan

Sarah M. Hammoudeh
Faculty of Medical and Human Sciences
University of Manchester
Manchester, UK

Arab M. Hammoudeh
College of Medicine
University of Sharjah
Sharjah, UAE

***Abstract*—Mobile phones technology has become one of the most common and important technologies that started as a communication tool and then evolved into key reservoirs of personal information and smart applications. With this increased level of complications, increased dangers and increased levels of countermeasures and opposing countermeasures have emerged, such as Mobile Forensics and anti-forensics. One of these anti-forensics tools is steganography, which introduced higher levels of complexity and security against hackers' attacks but simultaneously create obstacles to forensic investigations.**

A new anti-forensics approach for embedding data in the steganography field is proposed in this paper. It is based on hiding secret data in hex symbols carrier files rather than the usually used file multimedia carrier including videos, image and sound files. Furthermore, this approach utilizes hexadecimal codes to embed the secret data on the contrary to the conventional steganography approaches which apply changes to binary codes. Accordingly, the resulting data in the carrier files will be unfathomable without the use of special keys yielding a high level of attacking and deciphering difficulty. Besides, embedding the secret data in the form of hex symbols, the agreed upon procedure between communicating parties follows a random embedding manner formulated using WinHex software. Files can be exchanged amongst android devices and/or computers. Experiments were conducted for applying the proposed algorithm on rooted android devices, which also are connected to computers. The proposed methods showed advantages over the currently existing steganography approaches, in terms of character frequency, capacity, security, and robustness.

Keywords*—Mobile Forensics; Anti-Forensics; Artifact Wiping; Data Hiding; Wicker; Steganography

I. INTRODUCTION

In our days, not only water, air and food are considered to be our basic living needs but a consensus have been established to add technology including computers and mobile phones to this list. As mobile phones rapidly evolved from communication means to reservoirs of personal information and smart applications [1], they allowed their users to be exposed to increasing dangers and complexities. Consequently, many fields and technologies have been developed as counter-measures to such dangers. One of these fields is the Mobile Forensics, which aims at collecting and analyzing digital evidence to resolve mobile issues. However, on the other side, opposing measures such as Anti-Forensics technologies have been developed to hinder the use of mobile forensics [2]. One of these anti-forensics tools is steganography. Steganography systems are utilized to embed secret data in image, audio and video files that can only be discovered by the parties informed of the secret key of the steganography chosen algorithm. Thus, steganography introduces a higher level of complexity that would protect against attacks but at the same time create an obstacle for forensic investigations [3]. Current steganography approaches have been observed to present some disadvantages that require the development of a new approach or a solution to be overcome. This paper will be introducing some of the currently existing anti-forensics approaches and techniques. Thenceforth, the paper will be proposing a new steganography approach that utilizes Hex Symbols to hide data. The proposed approach in this paper has advantages over the currently existing steganography approaches in its capacity, security and robustness. Capacity indicates the quantity of the information that could be embedded in the stego-medium. Security is essential to keep confidential communication, a secret that can't be detected by intruders. Robustness refers to the ability of the stego-medium to handle alterations while maintaining the integrity of the embedded information [4].

The paper will be presenting background information and related work on anti-forensic techniques, artifact wiping, data hiding, and steganography tools and approaches in sections 2 and 3. Then the paper will be elaborating further on anti-forensics steganography in section 4. The description of the newly proposed steganography approach using hex symbols is presented in section 5 accompanied by the explanation of the implementation process in section 6. Finally the new approach is analyzed and discussed in section 7.

II. RELATED WORK

Several steganography tools have been developed, aiming at achieving the best method to embed secret messages, including the least significant bit (LSB) encoding technique,

the hash based LSB Techniques, and the Neighborhood Pixel Information.

The use of the LSB substitution technique in video steganography was introduced by Swathi and Jilani [5]. The principle of the technique revolved around finding and replacing the least significant bits in the image frames of the cover videos. all the color image components (i.e. red, green and blue) may be utilized for the same purpose by replacing the LSB in them by bits of the secret message. Thus the message to be hidden undergoes two conversions; the conversion to ASCII code and the conversion to binary representation.

In their work, Dasgupta el al. [6] presented a hash based LSB Techniques in spatial domain, utilizing an algorithm portrayed with AVI (Audio Video Interleave) file as a cover medium. A video stream (AVI) is composed of collection of frames in which the secret data can be concealed as payload. 8 bits of secret data would be concealed at a time in LSB of RGB pixel value of the carrier frames in 3, 3, 2 order respectively. This technique increased the payload and the difficulty of detection by human eyes due to the small variations in colours.

Another steganography tool was described by Hossain et al. [7] in which neighbourhood information are used to calculate the quantity of data that can be embedded in a cover image without causing a noticeable change. The complication and density of the different areas in the cover image are determined. Thence, small quantities of secret data are hidden in the smooth areas, while larger quantities are hidden in the complicated ones. This whole concept is built on psycho visual repetition in grey scale digital images; the edged parts can withstand more change in comparison to the smooth ones.

III. Anti-Forensics

Anti-forensics (AF) techniques are used to avoid and eliminate the possibility of evidence detection by the mobile forensics tools [1]. AF techniques and tools are continuously and rapidly evolving. By understanding the basics and principles of these techniques, more complex approaches or opposing forensic tools can be developed. Two major types of Anti-Forensic techniques, artifact wiping and data hiding, will be briefly presented next.

A. Artifact Wiping

Artifact wiping, also known as sanitation, overwrites data files from digital devices permanently erasing them. Some artifact wiping tools, including Binary Code (BC) wipe, Eraser, and Pretty Good Privacy (PGP) wipe, target empty and unallocated spaces [8].

One of the applications that function against mobile forensics is Wicker, a free application that allows users to send self-destructing files and messages [9]. According to its developers, Wicker has the ability to function without leaving any evidences or traces behind for forensic investigators. The unique characteristic of this application is that it allows the user to set a self-destruct timer to anything they send. To investigate the efficacy of Wicker, an experiment was conducted using an Android smart phone. A newly created account was used to initiate an instant massaging conversation with a Wicker a certain friend. A timer for self-destruction was set for a sample of messages (self-destruction duration: 5 days). Upon testing and searching for traces left from Wicker's conversations after the set time, none were found.

B. Data Hiding

Data hiding tools have been developed to secretly embed and hide undiscoverable data through multiple approaches. These approaches include transferring data to other portable storage devices and then wiping the data from the phone; making data "invisible" and concealing their existence; embedding data in multimedia (image, audio and video) files; and altering file extensions. Although some of these approaches, such as altering file extensions, are relatively old, evidence have shown that they can still bypass some forensic analysis methods. For instance, an experiment has shown that changing the extension of an .mp4 file to .pdf allowed for hiding it from the evidence tree generator, FTK imager, without applying any changes to its location [10].

IV. The Anti-Forensic Steganography

According to [3], "Steganography is the art and science of hiding information in plain sight". Thus, through steganography, a stego-system unknown to third, uninvolved parties can be created to allow for data exchange under extremely secure conditions. Digitally, data hiding techniques are important tools for the utilization of steganography. Through these tools, image, audio and video steganography can be applied.

Stego algorithms delete the repeated bits in the cover multimedia files and replace these bits with the secret data. Video and audio files with higher qualities contain larger quantities of the repeated bits required for steganography. The perk of relying on video files in hiding information is their relative immunity to hacker attack due to the relative intricacy of their structure in comparison to that of image files. The idea of video steganography is to hide the data in compressed or uncompressed domains that combine sets of frames and audio files [11]. Therefore, the complexity and efficacy of the process is increased in comparison with the simpler forms of steganography, the image and audio steganography. Such complexity allows for higher security against infiltration and hacker attacks. Video based steganography techniques are generally categorized into Spatial domain and frequency domain.

A spatial domain technique embeds the information to be concealed in the intensity pixels of the carrier multimedia file. The advantage of this category of techniques is their use of the Least Significant Bit (LSB) algorithms to embed the load of data. However, the drawback is that the majority of the LSB techniques are susceptible to attacks. In frequency domain techniques, on the other hand, images are transformed to frequency components by using some techniques, such as Fast Fourier Transform (FFT), Discrete Cosine Transformation (DCT) or Discrete Wavelet Transform (DWT). Thenceforth, the messages are planted and hidden in some or all of the transformed coefficients [6].

In brief, the process of steganography is commenced through an agreement of two parties on a stego-system and a secret key for the embedding algorithm. The accordingly

chosen embedding algorithm would be responsible for allocating the carrier files according to their bits content. The redundant bits are modified and replaced with the bits of the secret message to be exchanged by parties involved. This process prevents any third party lacking the knowledge of the secret key and the chosen embedding algorithm from discovering the embedded data or breaching the carrier file contents [3]. The general steganography process is summarized in Fig.1.

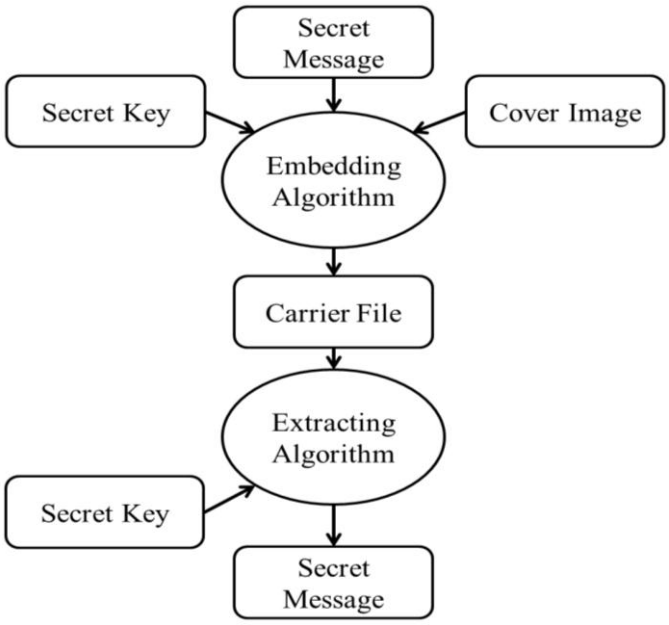


Fig. 1. The general steganography process

V. The Proposed Hex Symbol (HSA) Scheme

A. HSA Design

In this paper, we present a new anti-forensics approach for ciphering data in the steganography field. This approach hides the target data in hex symbols rather than the usually used multimedia files. Accordingly, the resulting data files will be unfathomable without the use of special keys; hence extremely hard to breach and decipher. The approach embeds the target data in the form of hex symbols in a random manner. Winhex facility is used to view the resulting file. The files can then be exchanged through android devices and computers. In this work, experiments were conducted on rooted android devices in combination with computer networks.

The new approach is expected to have several advantages over the traditional steganography approaches. One of these advantages is the impossible detection of these hidden data by the human eyes, as the content will be presented in the form of hex symbols. This problem was constantly observed in the other approaches in the form of disruptions in the sound and video files and changes in image frames and colors. One of the underlying causes of this problem is the limited capacity to conceal data in the traditional multimedia files. Exceeding this capacity leads to the leak of traces of the hidden data that are observable by external parties [4]. The nature of the carrier file used in the new approach allows for surmounting this problem.

Additionally, the traditional steganography methods apply supplementary tools to embed the secret data resulting in the appearance of additional errors and detectable traces during the data hiding process. These tools have been found to cause the concentrated addition of data to a single location in the file, the consistent addition of certain strings to multiple files, or the addition of signatures linking the file to the embedment tool. These flows brought about by the supplementary tools are eliminated in the HSA approach.

The randomness in hiding the data further upraises the level of security as it increases the difficulty of locating the hidden data while the hiding approach increases the difficulty of deciphering the hidden data. However, although this approach would provide a higher degree of security against attacks, it adds a new complication to forensic investigations.

B. Hex Symbols Algorithm (HSA)

To start with, certain patterns would be agreed upon amongst the communicating parties. These patterns are created by converting a chosen file into hexadecimal symbols using WinHex software, segmenting the file's content into 16x16 matrices, and numbering the matrices sequentially as shown in fig. 2. Some of these segments will be selected and used for embedding the secret messages. Embedding the messages would be guided by the previously agreed upon patterns representing the hiding keys shared by the communicating parties, as shown in fig. 3. However, any shape or number of shapes can be designed and used generally. These patterns are arranged randomly in a table, as shown in table I, and shared between the communicating parties to serve as the secret key codebook. Each table entry represents a string of pattern numbers denoted by a letter that is going to be the key.

After the hiding keys and the secret codebook are prepared, the secret message can be embedded in the carrier file according to the following steps:

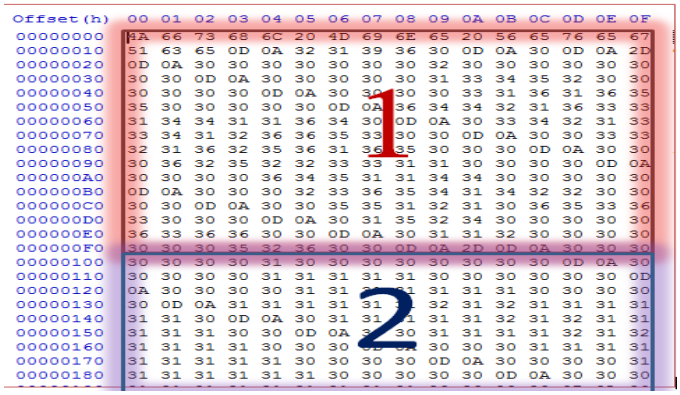


Fig. 2. The segmented and numbered hex code of the carrier file

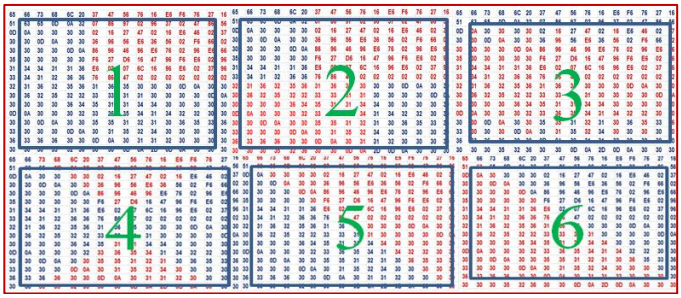


Fig. 3. Specific pattern design for matrices

1) The input secret message is converted into hex symbol.

2) The hex symbol decimals are then invered. For example, if the letter 'n' hex symbol is 63, then it is inverted into 36.

3) the resulting hex codes are then embedded into the carrier file which contains a randomly chosen sequence of matrix segments.

4) The contents of these matrix segments are relocated by exchanging rows with columns in order to increase the difficulty for hackers.

5) Once all the contents of the secret message are hidden, these segments (stego-file) are concatenated with the randomly chosen sequence of the matrix segments (or the key), and sent to the receiving party.

Fig. 4 shows an example of the matrix segment sequence when the key symbol 'S' was selected from table I.

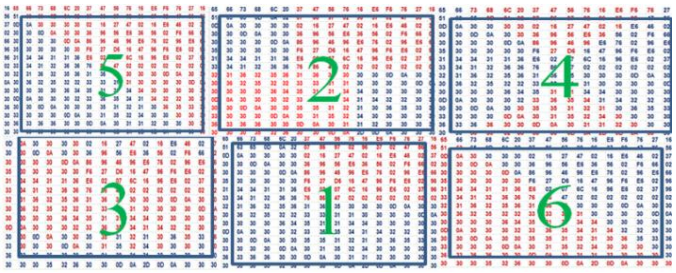


Fig. 4. Key symbol "S"matrix segment sequence "524316"

TABLE I. EXAMPLE OF THE SHARED KEY SYMBOL CODEBOOK

Key symbol	Selected random sequence
S	524316
O	643125
M	365123
Y	513642

A summary of the proposed HSA steganography algorithm is presented in fig. 5-a. The receiving party would be get the stego-file carrying the secret message as well as a key indicating the chosen pattern in a numeric representation. Accordingly, the receiving party would be able to comprehend the arrangement of the matrices by referring back to the secret codebook. Thereafter, the steganography steps can be executed in reverse with the guidance of the chosen secret keys and patterns to decode the hidden message (Fig. 5-b).

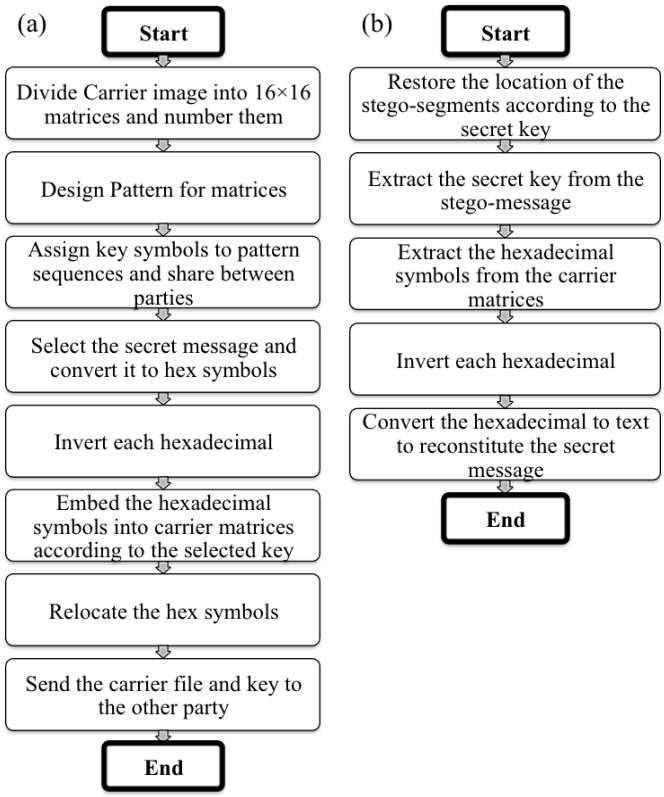


Fig. 5. The Flow Charts summarize the procedures of (a) embedding secret messages by the sending party and (b) extracting hidden messages by the receiving party according to the proposed HSA scheme

VI. IMPLEMENTATION

To implement the proposed HSA scheme, a message was embedded in a carrier file according to the following process. The secret message to be hidden was "Steganography is the art and science of hiding information in plain sight". First, the message was converted to Hexadecimal symbol representation. In this representation, each letter of the message is represented by two hexadecimal character components as shown in each first and second row of fig. 6. The two hexadecimal character components representing each letter are then inverted, resulting in the new hex symbols content shown in each third row of fig. 6.

s	t	e	g	a	n	o	g	r	a	p	h	y
73	74	65	67	61	6e	6f	67	72	61	70	68	79
37	47	56	76	16	e6	f6	76	27	16	07	86	97
	i	s		t	h	e		a	r	t		a
20	69	73	20	74	68	65	20	61	72	74	20	61
02	96	37	02	47	86	56	02	16	27	47	02	16
n	d		s	c	i	e	n	c	e		o	f
6e	64	20	73	63	69	65	6e	63	65	20	6f	66
e6	46	02	37	36	96	56	e6	36	56	02	f6	66
	h	i	d	i	n	g		i	n	f	o	r
20	68	69	64	69	6e	67	20	69	6e	66	6f	72
02	86	96	46	96	e6	76	02	96	e6	66	f6	27
m	a	t	i	o	n		i	n		p	l	a
6d	61	74	69	6f	6e	20	69	6e	20	70	6c	61
d6	16	47	96	f6	e6	02	96	e6	02	07	c6	16
i	n		s	i	g	h	t					
69	6e	20	73	69	67	68	74					
96	e6	02	37	96	76	86	47					

Fig. 6. Secret message hex symbols after inversion

The resulting hex symbols, representing of the secret message, were henceforth embedded into the carrier hex file according to the chosen pattern, as shown in Fig. 7.

65	66	73	68	6C	20	37	47	56	76	16	E6	F6	76	27	16
51	63	65	0D	0A	32	07	86	97	02	96	37	02	47	86	56
0D	0A	30	30	30	30	02	16	27	47	02	16	E6	46	02	37
30	30	0D	0A	30	30	36	96	56	E6	36	56	02	F6	66	02
30	30	30	30	0D	0A	86	96	46	96	E6	76	02	96	E6	66
35	30	30	30	30	30	F6	27	D6	16	47	96	F6	E6	02	96
31	34	34	31	31	36	E6	02	07	c6	16	96	E6	02	37	96
33	34	31	32	36	36	76	86	47	02	02	02	02	02	02	02
32	31	36	32	35	36	31	36	35	30	30	30	0D	0A	30	30
30	36	32	35	32	32	33	33	31	31	30	30	30	30	0D	0A
30	30	30	30	36	34	35	31	31	34	34	30	30	30	30	30
0D	0A	30	30	30	32	33	36	35	34	31	34	32	32	30	30
30	30	0D	0A	30	30	35	35	31	32	31	30	36	35	33	36
33	30	30	30	0D	0A	30	31	35	32	34	30	30	30	30	30
36	33	36	36	30	30	0D	0A	30	31	31	32	30	30	30	30
30	30	30	35	32	36	30	30	0D	0A	2D	0D	0A	30	30	30

Fig. 7. Example of embedded message into the carrier file segment (the embedded secret message represented by bold red color)

The contents of the carrier stego-file segment were then rearranged by interchanging the positions of rows and columns as shown in Fig. 8.

65	51	0D	30	30	35	31	33	32	30	30	0D	30	33	36	30
66	63	0A	30	30	30	34	34	31	36	30	0A	30	30	33	30
73	65	30	0D	30	30	34	31	36	32	30	30	0D	30	36	30
68	0D	30	0A	30	30	31	32	32	35	30	30	0A	30	36	35
6C	0A	30	30	0D	30	31	36	35	32	36	30	30	0D	30	32
20	32	30	30	0A	30	36	36	36	32	34	32	30	0A	30	36
37	07	02	36	86	F6	E6	76	31	33	35	33	35	30	0D	30
47	86	16	96	96	27	02	86	36	33	31	36	35	31	0A	30
56	97	27	56	46	D6	07	47	35	31	31	35	31	35	30	0D
76	02	47	E6	96	16	6C	02	30	31	34	34	32	32	31	0A
16	96	02	36	E6	47	16	02	30	30	34	31	31	34	31	2D
E6	37	16	56	76	96	96	02	30	30	30	34	30	30	32	0D
F6	02	E6	02	02	F6	E6	02	0D	30	30	32	36	30	30	0A
76	47	46	F6	96	E6	02	02	0A	30	30	32	35	30	30	30
27	86	02	66	E6	02	37	02	30	0D	30	30	33	30	30	30
16	56	37	02	66	96	96	02	30	0A	30	30	36	30	30	30

Fig. 8. Embedded message into the carrier file segment (the embedded secret message represented by bold red color)

For example, the whole segment elements could be flipping around the diagonal according to eq. 1:

$$x'_{ij} = x_{ji} \quad (1)$$

Where x'_{ij} are the new matrix elements of the stego-file and x_{ji} are the old matrix elements.

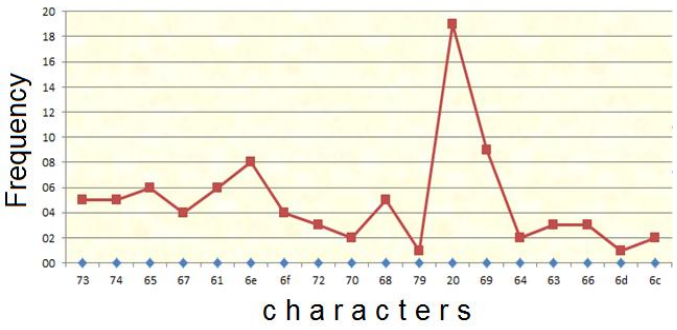
The resulting final form of the stego-file can be safely sent to the other parties. With the chosen secret patterns and keys, they can retrace the steps back and recover the embedded text.

VII. Analysis and Discussion

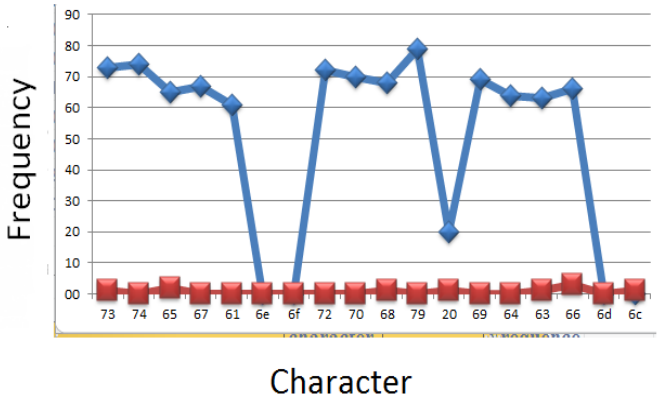
The frequency, capacity, using WinHex, the evidences and robustness of the proposed HSA scheme were analyzed as follows.

A. Frequency

In As the two hexadecimal character components of the hex symbol are inverted, the calculated frequency of occurrence for each character before inversion will differ from that after inversion (fig. 9-a and 9-b).



(a)



(b)

Fig. 9. Character frequency for embedded message (a) before inversion (b) after inversion

From the compiled frequencies analysis in table II and fig. 10, a clear change was observed between the frequency of characters before and after the inversion process. This is a positive indicator of the high level of security against third attacking parties. Furthermore, elongating the embedded sentence is expected to further increase the security factor of the process.

TABLE II. Character Frequency Before (F.B) and After (F.A) Inversion

Character	F.B	F.A
73	05	01
74	05	00
65	06	02
67	04	00
61	06	00
6e	08	00
6f	04	00
72	03	00
70	02	00
68	05	01
79	01	00
20	19	01
69	09	00
64	02	00
63	03	01
66	03	03
6d	01	00
6c	02	01

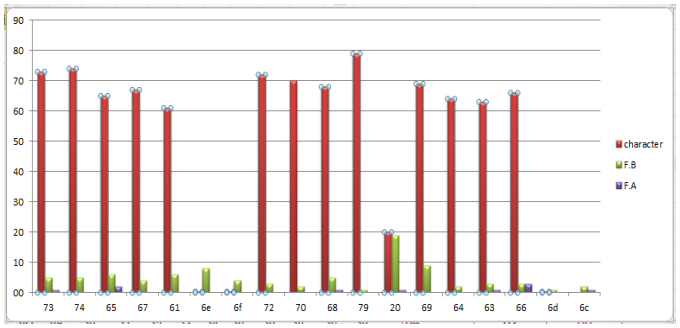


Fig. 10. Character frequency before (F.B) and after (F.A) inversion

B. The Capacity

TABLE III. THE CAPACITY COMPARISON

Steganography Approaches		**Comparison category**
Hash based LSB Techniques representation	Hex Symbols representation	Method for embedded
Binary	Hex	Type of ASCII Numerical representation
A= 01000001 R= 01010010 T= 01010100	A= 41 R= 52 T= 54	"ART" in ASCII representation
Select color value for example: G=10010100 R=10110111 B=11001001	Select hexadecimal for example: 61 62 64	Carrier Image
1001**010**0 1011**00**1**0** 11001**01** 1001**010**0 1011**00**1**1** 11001**10** 1001**010**0 1011**01**1**1** 11001**00**	**41 52 54**	Carrier Image containing the embedded word "ART"
• Binary numbers allow the use of only 2 symbols (0, 1) to symbolize any number and it becomes a tiresome job to express large numbers [12]. • When a large number is represented in binary system, it results into extremely difficult, lengthy and non- readable by human [12]. • More code but less capacity	• Hexadecimal numbers allow the use of 16 symbols 0 to 9 and additional symbols (A, B, C, D, E, F). Hexadecimal numbers were presented to fulfil the aim of symbolizing binary numbers in a more human readable form [12]. • When numbers are represented in hexadecimal system, they are considered to be easier and more human readable than binary number [12]. • Less code and more capacity	Output

Comparisons of some capacity and size characteristics between the proposed HSA scheme and the hash based LSB technique have been enlisted in table III. This comparison has shown the lower capacity requirement of the proposed steganography approach, which allows for the addition of larger extensions of messages without having a large effect on the carrier file.

C. Using WinHex

The use of WinHex to formulate the hex symbols during the hiding process is advantageous as the content will be difficult to trace and compare with previous versions. This advantages is further boosted by the frequent and continuous rearrangements of the hex symbols according to the chosen codebooks and patterns throughout the steganograohy procedure.

Moreover the alteration of the hex symbols by the inversion of each character element doesn't increase the original file size, leaving it stable and unchanging in terms of elements number.

Furthermore, the use of random numbers to select the segments provides an extra complication against deciphering the hidden text. A comparison between available steganalysis tools and hex symbols is presented in Table I V.

TABLE IV. A COMPARISON BETWEEN STEGANALYSIS AND HEX SYMBOLS

	STEGANALYSIS TOOLS	**HEX SYMBOLES**
OurSecret OmniHide BDV DataHider Max file encryption Masker StegoStick	These tools work by embedding information within videos by attaching it bluntly to the end of the file EOF [3].	Hex symbols substitutes the hexadecimal precisely on the same position.
OurSecret	This signature can be found after the last byte of the authentic unmodified file.	A valid signature similar to OurSecret does not apprear.
OmniHide Pro	White space characters tailing the initial sequence of bytes.	Hex symbols do not show the name of the embedded file.

D. The Evidences

To be able identify hidden data, investigators search for steganography tools, such as S-Tools, DPEnvelope, jpgx, on the suspect's personal devices (i.e. phones and computers). The presence of such tools would imply the highly possibility of finding hidden carrier files modified using these tools. Therefore, the investigators further expand their search to identify any possible multimedia or text files that could have been used to hide data [13]. However, with regards to the proposed HSA scheme, no specific tools are used to insert the embedded text; Common programming means found in widely used software, such as VBA found in Microsoft Excel, can be used in this approach to eliminate the shortcomings of the use of external tools. Moreover, the hex symbol file extension can be changed to thwart hackers and investigators.

E. Compression

The stego-files have been found to be resistant against changes in size and content when compressed to WinRAR or ZIP file format and when processed for message extraction. This resistance indicates the robustness of the proposed approach against processing procedure that could be applied to the carrier file such as compression.

VIII. Conclusion & Future Work

Hex symbol algorithm (HSA) scheme is a newly proposed steganography approach developed for hiding secret messages in hex symbols rather than the usually used multimedia files. The files can be exchanged with random keys through android devices or computers. In terms of character frequency, capacity, using WinHex, the evidences and robustness, HSA stagnography has shown improved outcomes in comparison with other steganography approaches. Hidden data using this approach are impossible to detect by the human eyes. Furthermore, traces such as changes in file size and clarity as well as additions of extra information at the end of files have been found to be eliminated using HAS steganography. In the future, this approach can be developed further to increase its complexity and utilize it in various applications. In the future, this approach can be developed further to increase its complexity and utilize it in various applications.

Acknowledgment

We thank prof. Nihad Yusuf – Dean of King Abdullah I Faculty of Graduate Studies and Scientific Research, prof. Arafat Awajan – Dean of King Hussein Faculty for Computing Sciences, and Dr. Saqer Abdel Rahim – Head of the computer science department in King Huessein Faculty for computing Sciences, for their continuous support of this work.

References

[1] A. Distefano, G. Me and F. Pace, "Android anti-forensics through a local paradigm," Digital Investigation, vol. 7, pp. S83-S94, August 2010.

[2] K. Dahbur and B. Mohammad "The Anti-Forensics Challenge," Proceedings of the 2011 International Conference on Intelligent Semantic Web-Services and Applications - ISWSA '11, ACM Press, April 2011.

[3] T. Sloan and J. Hernandez-Castro, "Forensic analysis of video steganography tools," PeerJ Computer Science, vol. 1, pp. e7, May 2015.

[4] S. Sirsikar and A. Deshpande, "Steganographic Tools for BMP Image Format," International Journal of Computer Science & Emerging Technologies (IJCSET), vol. 2, pp. 200-204, February 2011.

[5] A.Swathi, Dr. S.A.K Jilani "Video Steganography by LSB Substitution Using Different Polynomial Equations" International Journal Of Computational Engineering Research (ijceronline.com) Vol. 2 Issue. 5.2012.

[6] K.Dasgupta1, J.K. Mandal and P.Dutta "HASH BASED LEAST SIGNIFICANT BIT TECHNIQUE FOR VIDEO STEGANOGRAPHY(HLSB) " International Journal of Security, Privacy and Trust Management (IJSPTM), Vol. 1, No 2, April 2012

[7] M.Hossain, S. Al Haque, and F.Sharmin " Variable Rate Steganography in Gray Scale Digital Images Using Neighborhood Pixel Information " The International Arab Journal of Information Technology, Vol. 7, No. 1, January 2010.

[8] P. A. Kotsopoulos and Y. C. Stamatiou, "Uncovering Mobile Phone Users' Malicious Activities Using Open Source Tools," Advances in Social Networks Analysis and Mining (ASONAM), 2012 IEEE/ACM International Conference on, Istanbul, pp. 927-933, August 2012.

[9] T. Mehrotra and B. M. Mehtre, "Forensic Analysis of Wickr Application on Android Devices," 2013 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), Enathi, pp. 1-6, December 2013.

[10] A. Jain and G. S. Chhabra, "Anti-Forensics Techniques: An Analytical Review," 2014 Seventh International Conference on Contemporary Computing (IC3), Noida, pp. 412 – 418, August 2014.

[11] K. Dasguptaa, J. K. Mondalb, and P. Dutta, "Optimized Video Steganography using Genetic Algorithm (GA)," Procedia Technology, vol. 10, pp. 131-137, 2013. [International Conference on Computational Intelligence: Modeling, Techniques and Applications (CIMTA) 2013].

[12] "Difference between binary and hexadecimal", Schoolelectronic.com. [Online]. Available: http://www.schoolelectronic.com/2012/10/difference-between-binary-and-hexadecimal.html. [Accessed: Jan- 2016].

[13] B. Nelson, A. Phillips and C. Steuart, "Guide to computer forensics and investigations," 4th ed. Boston, MA: Course Technology Cengage Learning, 2010.

Urdu to Punjabi Machine Translation: An Incremental Training Approach

Umrinderpal Singh
Department of Computer Science,
Punjabi University
Patiala, Punjab, India

Vishal Goyal
Department of Computer Science,
Punjabi University
Patiala, Punjab, India

Gurpreet Singh Lehal
Department of Computer Science,
Punjabi University
Patiala, Punjab, India

***Abstract*—The statistical machine translation approach is highly popular in automatic translation research area and promising approach to yield good accuracy. Efforts have been made to develop Urdu to Punjabi statistical machine translation system. The system is based on an incremental training approach to train the statistical model. In place of the parallel sentences corpus has manually mapped phrases which were used to train the model. In preprocessing phase, various rules were used for tokenization and segmentation processes. Along with these rules, text classification system was implemented to classify input text to predefined classes and decoder translates given text according to selected domain by the text classifier. The system used Hidden Markov Model(HMM) for the learning process and Viterbi algorithm has been used for decoding. Experiment and evaluation have shown that simple statistical model like HMM yields good accuracy for a closely related language pair like Urdu-Punjabi. The system has achieved 0.86 BLEU score and in manual testing and got more than 85% accuracy.**

Keywords*—Machine Translation; Urdu to Punjabi Machine Translation; NLP; Urdu; Punjabi; Indo-Aryan Languages

I. Introduction

The machine translation is a burning topic in the area of artificial intelligence. In this digital era where across the world different communities are connected to each other and sharing a vast amount of resources. In this kind of digital environment, different natural languages are the main obstacle to communicate. To remove this barrier researcher from different countries and big companies are putting efforts to develop machine transition system to resolve this barrier. Various kinds of approaches have been developed to decode natural languages like Rule based, Example-based, Statistical and various hybrid approaches. Among all these approaches, statistical based approach is a quite dominant and popular in the machine translation research community. The statistical systems yield good accuracy as compared to other approaches but statistical models need a huge amount of training data. In comparison to European languages Asian languages are resources poor languages therefore it is challenging task to collect parallel corpus for training these statistical model. There are many machine translation systems which have been developed for Indo-Aryan languages [Garje G V, 2013]. Most of the work have been done using rule-based or hybrid approaches because the non-availability of resources. The proposed system based on an incremental training process for training the machine learning algorithm. Efforts have been made to develop parallel phrase corpus in place of parallel sentence corpus. Collecting parallel phrases were more convenient as compared to the parallel sentences.

II. Urdu and Punjabi: A Closely Related Language Pair

Urdu[2] is the national language of Pakistan and has official language status in few states of India like New Delhi, Uttar Pradesh, Bihar, Telangana, Jammu and Kashmir where it is widely spoken and well understood throughout in the other states of India like Punjab, Rajasthan, Maharashtra, Jharkhand, Madhya Pradesh and many other[1] . The majority of Urdu speakers belong to India and Pakistan, 70 million native Urdu speakers are in India and around 10 million speakers in Pakistan[2] and thousands of Urdu speakers living in US, UK, Canada, Saudi Arabia and Bangladesh. The word Urdu is derived from Turkic word ordu which means army camp[2]. The Urdu language was developed in 6th to 13th century. Urdu vocabulary mainly derived from Arabic, Persian, and Sanskrit and it is very closely related to modern Hindi language. Urdu is written in Nastaliq style and script is written from right to left using heavily derided alphabets from Persian which is an extension of Arabic alphabets. [3]Punjabi is an Indo-Aryan language and 10th most widely spoken language in the world there are around 102 million native speakers of Punjabi language across worldwide[4]. Punjabi speaking people mainly lived in India's Punjab state and in Pakistan's Punjab. Punjabi is the official language of Indian states like Punjab, Haryana, and Delhi and well understood by many other northern Indian regions. Punjabi is also a popular language in Pakistani Punjab region but still did not get official language status. In India, Punjabi is written in Gurmukhi script means from Guru's mouth and in Pakistan Shahmukhi is used means from the king's mouth. Despite from the different scripts use to write Punjabi, both languages share all other linguistics features from grammar to vocabulary in common.

Urdu and Punjabi are closely related languages and both belong to same family tree and share many linguistic features like grammatical structure and vast amount of vocabulary etc. for example:

Urdu: - وہ پنجابی یونیورسٹی کا طالب علم ہے

Punjabi: ਉਹ ਪੰਜਾਬੀ ਯੂਨੀਵਰਸਿਟੀ ਦਾ ਵਿਦਿਆਰਥੀ ਹੈ ।

English: He is a student of Punjabi University.

Despite from script and writing order where Urdu is written in right to left using Arabic script and Punjabi from left to right using Gurumukhi script, every other linguistic feature is the same in both sentences. Both sentences shares same grammatical order and most of the vocabulary, this is also true in care of more complex sentences. By analysis of both languages, we found that both languages share many similarities and are used by a vast community of India and Pakistan. Therefore, we need a natural language processing system which can help these people to share and understand text and knowledge. The efforts have been made to develop a machine translation system for Urdu to Punjabi text to overcome this language barrier between both the communities. With the help of this machine translation system, native Punjabi reader can understand Urdu text by translating into Punjabi text.

III. Challenges to Develop Urdu to Punjabi MT System

A. ***Resource poor languages:*** Urdu and Punjabi languages are new in natural language processing area like any other Indo-Aryan language. Both languages are resource-poor language, very small or no annotated corpus is available for development of a full-fledged system.

To develop a machine translation system based on the statistical model, one should need a huge parallel corpus to training the model. For rule-based approach or hybrid machine translation system, one should need a good part of speech tagger or stemmer and large parallel dictionaries. To best of our knowledge, Urdu-Punjabi language pair does not have these resources in a vast amount to train or develop the system. Therefore, development of resources is one of the key challenges to work on this language pair.

B. **Spelling variation:** Due to lack of spelling standardization rules, there are many spelling variation for the same word. [Singh, UmrinderPal et.al 2012] Both languages use tons of loan words from English. Therefore, many variations come in existence, for example, word 'Hospital' can be written in two ways in Urdu اسپتال/ ہسپتال hasptaal/asptaal. It is always a challenging task to cover all variation of a word. There is no standardization in spelling. Therefore, it all depends on a writer which spelling he/she choose to write foreign language words.

C. **Free word order:** Urdu and Punjabi are free word order languages. Both languages have unrestricted word order or phrase structures to form the sentences that make the machine translation task more challenging. For example,

Urdu: رام نے ستا کو اپنی کتاب دی

Transliteration: raam ne satta ko apanee kitaab dee.

English: Ram gave his book to Sita.

This can be rewritten as following:

Urdu: رام نے دے دی ستا کو اپنی کتاب

Transliteration: raam ne dee sata ko apanee kitaab.

Urdu: رام نے دے دے اپنی کتاب ستا کو

Transliteration: raam ne de apanee kitaab sata ko.

Urdu: رام نے اپنی کتاب ستا کو دے دی

Transliteration: raam ne apanee kitaab sata de dee.

Above example shows that same sentence can be written in various ways due to free word order and all sentences give exactly the same meaning. Therefore, it is always difficult to form every possible rule to interpreter's source language text to do machine translation.

D. **Segmentation issues in Urdu:** Urdu word segmentation issue is a primary and most significant task [Lehal, G. 2009]. Urdu is effected with two kinds of segmentation issues, space insertion and space omission [Durrani, Nadir et.al. 2010]. Urdu is written in Nastaliq style which makes the white space completely an optional concept. For example,

Non-Segmented: ا قافلےکےصدراحمدشیرڈوگرانےکہ

Segmented Text: قافلے کے صدر احمد شیر ڈوگرا نے کہ

Urdu reader can read this non-segmented text easily but this is still difficult for computer algorithms to understand. In preprocessing phase, modules like tokenization need to identify individual words for further processing, without resolving the segmentation issue, no NLP system can process Urdu text efficiently and yield less accuracy.

E. **Morphological rich languages:** Urdu and Punjabi are morphological rich languages, where one word can be inflected in many ways. For example, word 'chair' کرسی/kursi can take any form like کرسیا/kursiya, کرسیو/kurseo, کرسئے/kurseye etc. One should need to incorporate all the inflation in our knowledge base to translate them into the target language. Adding all the inflation forms of all words in training data is a big challenge otherwise, it will effect on the accuracy of the system.

F. **Word without diacritical marks:** Urdu has derived various diacritical marks from Arabic to produce vowel sounds, like Zabar, Zer, Pesh , Shad , hamza , Khari-Zabar, do-Zabar and do-Zer [Sani, Tajinder Singh 2011]. In naturally written text diacritical marks are used very rarely. Due to missing of diacritical marks, an Urdu word can be mapped to many different target language translations, for example, word dil/دل often used without diacritical marks and can be interpreted as 'Heart' and 'DELL' without knowing the context of this word. Missing of diacritical marks is a key challenge to choose a proper translation in the target language and the system always needs to disambiguate these words. Along with this, the missing diacritical marks create various variations of the same word, for example, word 'Urdu' can be written in three ways(اردو)(اُردو)(اُردُو). Therefore, one should need to include all of these variations in the training examples.

1. https://en.wikipedia.org/wiki/States_of_India_by_Urdu_speakers
2. https://en.wikipedia.org/wiki/Urdu
3. https://en.wikipedia.org/wiki/List_of_languages_by_number_of_native_speakers
4. https://en.wikipedia.org/wiki/Punjabi_language

IV. METHODOLOGY

An Incremental machine learning process has been used, in place of manually developed parallel sentences corpus of source and target languages. Urdu and Punjabi languages are resource-poor language; the non-availability of the parallel corpus is a primary challenge to develop a statistical machine translation system. Efforts have been made to develop a corpus of manually mapped parallel phrases. Figure 1 shows the overall learning process of machine translation systems. The system takes Urdu text document as input and translates using initial uniformed distributed data. Initially, the system has phrase tables for most frequent 5000 Urdu words mapped with Punjabi translations. Due to insufficient data in phrase tables, many Urdu words returned without translation in parallel phrase file generated by decoding module shown in Appendix 1.

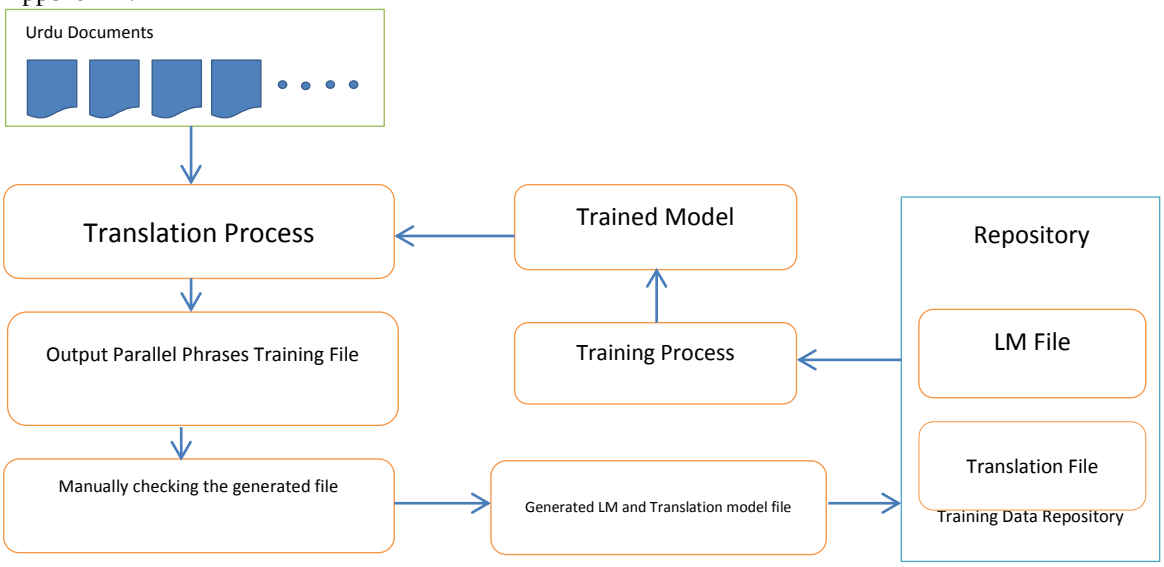


Fig. 1. Incremental MT training and decoding system

Then generated file manually corrected and updated with new translations by linguists. This updated file again submitted to the system to generate language model and translation model. The system learns new parameters from all the updated all files present in the repository of generated training files. Then system updates language model and phrase tables with a new vocabulary and update probabilities. With this incremental learning process, the system gets trained by each document it processes, learn and update language and translation model. The complete system is divided into five different processes or modules, Tokenization and segmentation, Text classification, Translation model learning, language model learning and decoding process.

A. **Tokenization and segmentation process:** Tokenization process is the primary and most significant task of any machine translation system. In preprocessing phase, the input text is divided into isolated tokens or words by tokenization process based on whitespace. Tokenization process is also a challenging task to identify valid tokens, when the system has noisy input data. Where tokens are often attached to neighboring tokens without any whitespace in-between them. This kind of writing trend is quite common in Urdu, where whitespace is an optional thing. The proposed tokenization process works on two levels, (1) isolates sentence boundary identification and (2) isolate word boundary identification.

1) **Tokenization into Sentences:** In sentence tokenization process, the system identifies sentence boundary based on few symbols used in Urdu to complete the sentence. For example, Urdu sentences often end with,{ ؟ ,۔ }, but symbol { ۔ } is an ambiguous one and not always used to identify the sentence boundary. This symbol { ۔ } also used as a separator in abbreviations. For example, آئی۔ سی۔ سی۔ , therefore, to tokenize text into sentences few rules were formed to check boundary conditions based on abbreviation. For example, the system always checks surrounding words of sentence termination symbols in abbreviation list.

2) **Tokenization into words:** The word tokenization process identifies individual tokens or words in the input text. To identify all the individual tokens first, one should need to separate all the words from symbols which are attached to words. For example, the system inserts whitespace in-between symbols and words and change them from آئے ہیں۔ to آئے ہیں ۔ .

ALGORITHM 1. Tokenization and Segmentation Process

```
Read Input Text in InputText
FinalList[]
Sentences[][]

        Insert space between word and symbols
        Tokenization InputText into Partial_Token_list[] form whitespace

        LOOP: Partial_Token_list[]
                IF: Current word is alphanumeric
                        Apply rules to word into split numeric and suffix part.
                        Add word in FinalList[]
                ELSE IF: Current word length > 3 and start with {کے , اور , سے}  and word not present in DB
                        Apply rules to split prefix and suffix parts
                        IF: suffix part is present in Phrase Table
                                Add prefix and suffix words in FinalList[].
                        END IF
                ELSE
                        Add word in FinalList[]
        END LOOP

        LOOP: FinalList[]
                IF: Current token is not a sentence separator
                        Sentence += token+" "

                ELSE IF:  Current token is a sentence separator AND previous and next are not abbreviation tokens
                        Add Sentence in Sentences[][]
        END LOOP
```

3) **Segmentation process:** The segmentation issue is a key challenge in Urdu text processing NLP applications. Segmentation issue can be handled on two levels, space insertion and space omission as discussed in MT challenges. In tokenization process, the system has handled only space insertion issue. Space omission problem is negligible in Unicode Urdu text but space insertion is quite frequent. To resolving the word segmentation problem in Urdu is quite a challenging task and need a full-fledged algorithm for this. Rather than handling all segmentation issues, the system has handled most frequent cases of segmentation. For example, in Urdu text, most of the time word attached with these prefixes {**کے, اور, سے**} which are ends with non-connecters and easily understood by Urdu reader but difficult for a computer algorithm to process. Few examples of segmentation words start with these prefixes are { **کےلیے, کےبعد, اورترک , اورنام , سےپہلے , سےکہیں** }. The analysis shows that these three words were 65% of all segmentation cases found in Urdu text and 5% cases of segmentation were related to alphanumeric words. Alphanumeric segmentation issue is also quite common in Urdu text, for example,{ **دسمبر21سے**،26}. Various rules have been developed to handle these types of tokens.

B. **Text Classification:** Most of the statistical machine translation system use single phrase table for translation. Instead of single phrase table for translation, the proposed system has used five different phrase tables for each domain. The system has trained on political, health, entertainment, tourism and sports domains. After tokenization process, text classifier needs to classify input text into most probable class, then translation module uses specific domain phrase table to translate input text. The text classifier returns a list of all domains with the higher probable domain on top followed by less probable domains. Other domains are used as a backoff model when the system did not find an Urdu phrase in the top domain then it searches in next less probable domain and so on.

$$C(punjabi\ phrases) = \begin{cases} phrase\ translation\ \ if\ domain = x1 \\ phrase\ translation\ if\ \ domain = x2 \\ phrase\ translation\ \ if\ domain = x3 \\ phrase\ translation\ if\ \ domain = x4 \\ phrase\ translation\ \ if\ domain = x5 \\ \quad else\ return\ original\ phrase \end{cases}$$

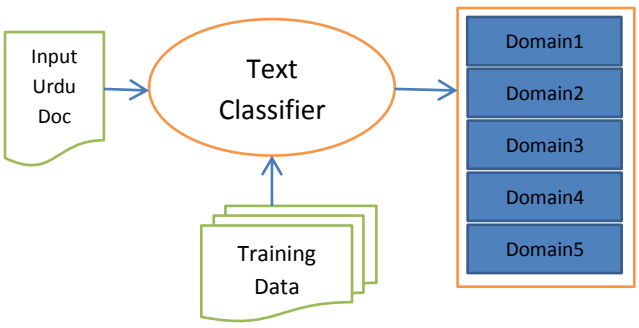


Fig. 2. Text classification system

Naïve Bayes model has been used to classify the input text, Naïve Bayes model considers document as bag of word where word positions are not important for classification, The Naïve Bayes approach based on Bayes rule defined as:

$$C = argmax_{c \in C} \frac{P(d|c)P(c)}{P(d)} \quad (1)$$

Rewriting by dropping the denominator because of constant factor:

$$= argmax_{c \in C} P(d|c)P(c) \quad (2)$$

To representing features of the documents for a class, equation can be written as:

$$= argmax_{c \in C} P(x_1, x_2, \dots, x_n|c)P(c) \quad (3)$$

Joint probability of whole set of independent features defined as:

$$P(x_1, x_2, \dots, x_n|c) = P(x_1|c) * P(x_2|c) * P(x_3|c) * \dots * P(x_n|c) \quad (4)$$

Simplified as:

$$C_{NB} = argmax_{c \in C} P(c_j) \prod_{x \in X} P(x|c) \quad (5)$$

To calculate maximum likelihood estimate and prior defined as:

$$P(w_i|c_j) = \frac{count(w_j, c_j)}{\sum_{w \in V} count(w, c_j)} \quad (6)$$

$$P(c_j) = \frac{Doc_{count}(C = C_j)}{N_{doc}}$$

To handle the unknown words, classifier has used Laplace smoothing defined as:

$$P(w_j|c) = \frac{Count(w_i, c) + \lambda}{\sum_{w \in V} count(w, c) + \lambda} \quad (8)$$

Rewritten as:

$$P(w_j|c) = \frac{Count(w_i, c) + \lambda}{\sum_{w \in V} count(w, c) + |V|} \quad (9)$$

Where $|V|$ is size of vocabulary and λ is constant value to add in frequency count of word in a document.

The system has used a list of 100 stop words to remove uninformative words which are common in training examples. Urdu is a morphologically rich language and one word can appear in the corpus with different suffixes, therefore, to transform all inflected words to root form in the training examples Urdu stemming rules has been used [Rohit Kansal et.al 2012].

C. **Translation and Language model Training:** The machine translation system's training process is divided into two main parts, Translation model, and Language model learning. The system used Hidden Markov Model (HMM) as learning process and Viterbi algorithm as a decoder.

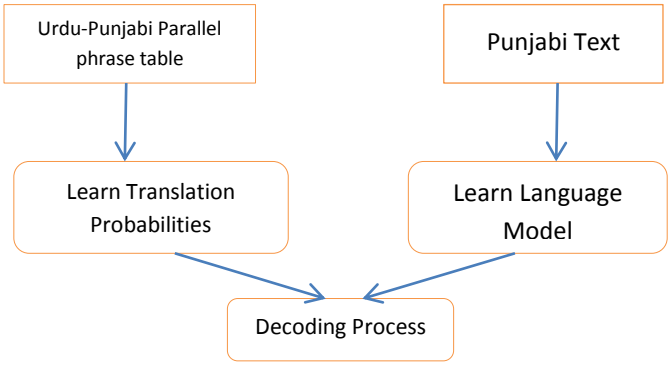


Fig. 3. Statistical machine translation model

HMM is a generative model defined as:

$$f(s_1 \dots s_n) = \underset{t \dots t_n}{\text{argmax}} P(s_1 \dots s_n, t_1 \dots t_n) \quad (10)$$

Where $s_1 \dots s_n$ are source language phrases and $t_1 \dots t_n$ target language phrases. By inputting the $s_1 \dots s_n$, we take the highest probability phrase sequence as output of target language. One should define bigram HMM model as below:

$$p(s_1 \dots s_n, t_1 \dots t_n) = \prod_i q(t_n|t_{n-1}) \prod_i e(s_i|t_i) \quad (11)$$

$$q(t_n|t_{n-1}) = \frac{Freq(t_{n-1}t_n)}{Freq(t_{n-1})} \quad (12)$$

$$e(s_i|t_i) = \frac{Freq(t_i \rightarrow s_i)}{Freq(t_i)} \quad (13)$$

1) **Translation model:** Urdu and Punjabi languages are closely related languages. Both languages share identical grammatical structure as well as same word order [Durrani, Nadir et.al 2010]. To learn the translation model we have manually mapped the phrases of source and target languages. Where IBM models provide an elegant solution to automatically mapped source and target language phrases, but for that, one should really need a large parallel corpus to train the model. Urdu and Punjabi are resource poor languages as we discussed in challenges. Therefore, the efforts have been made to find out a simple and effective solution for the training process.

The system takes manually mapped phrases as a training file and calculates translation probabilities. Sample of a training file is shown in appendix 1.

For example: word اتفاق can translate into four different ways.

TABLE I. POSSIBLE TRANSLATIONS

Urdu Word	Punjabi Word
اتفاق	ਸਹਿਮਤ
	ਸਹਿਮਤੀ
	ਸਹਿਯੋਗ
	ਹਮਾਇਤ

Maximum likelihood estimation of word اتفاق .

$$P_{Urdu}(punj)\begin{cases}0.19047619048 \ if\ punj = \text{ਸਹਿਮਤ}\\ 0.17460317460 \ if\ punj = \text{ਸਹਿਮਤੀ}\\ 0.49206349206 \ if\ punj = \text{ਸਹਿਯੋਗ}\\ 0.14285714286 \ if\ punj = \text{ਹਮਾਇਤ}\end{cases}$$

$$P(phrase) = \sum_i P_{Urdu}(punj)_i = 1 \quad (14)$$

TABLE II. POSSIBLE TRANSLATION WITH PROBABILITY VALUES

Urdu Words	P(punj|urdu)
اس	ਇਸ (0.53138492195)
	ਇਹ (0.4251 793756)
	ਉਸ (0.04350714049)
سفر	ਸਫਰ(1.0)
میں	ਮੈਂ (0.0193076817)
	ਵਿਚ (0.0013791201)
	ਵਿੱਚ (0.98055440629)
وہ	ਉਹ (1.0)
پہلا	ਪਹਿਲਾ (1.0)
میچ	ਮੈਚ (1.0)

$$p(p|u) = q(\text{ਇਸ}|\text{اس}) * q(\text{ਸਫਰ}|\text{سفر}) * q(\text{ਵਿੱਚ}|\text{میں}) * q(\text{ਉਹ}|\text{وہ}) * q(\text{ਪਹਿਲਾ}|\text{پہلا}) * q(\text{ਮੈਚ}|\text{میچ})$$

= 0.53138492195*1.0*0.98055440629*1.0*1.0*1.0

= 0.521051826

If training algorithm knows mapping in advance then it is quite straightforward to calculate translation probabilities from their occurrence in training data. In proposed method, the training algorithm already has alignments of all phrases, therefore; it can calculate parameters for the generative model.

$$P(phrase_i) = \frac{Count(phrase_i)}{\sum Count(*)} \quad (15)$$

Appendix 1 shows one-to-one, one-to-many, many-to-one, many-to-many word mapped phrases. In training data, we try to combine multiple words into a phrase which are frequent or combined words yield valid translation in target language. To compare with IBM models, we have used 50000 thousand parallel Urdu-Punjabi sentences to train the model using Moses toolkit which used Giza++ for phrase alignment. For 50000 sentences Moses generated over 3168873 phrases of size 503 MB. By examined generated phrase table manually and found many miss alignments and unnecessary long phrases those were increasing the size of phrase table and adding complexity to search space for decoding algorithm. As compared to an automatically generated phrase table, our manually mapped phrase table for the same set of sentences contains 56023 thousand phrases which are sufficient to translate given sentences accurately of that domain as shown in experiment section. In our phrase table, a maximum length of any phrase was four-gram and total four-gram phrases was 1093 compared to automatically generated phrase table contain several thousands of four-gram phrases.

Automatically find the alignment of words and phrases using parallel corpus is a graceful solution but when we deal with resource-poor languages we need to find out alternative ways. Development of machine learning resources like sentence-aligned parallel corpus is a time-consuming job. To train any machine translation model; one should require millions of parallel sentences. Therefore, if one do not have parallel corpus it is better idea to map phrases rather than writing parallel sentences. Mapping and checking phrases incrementally makes the job easier. Mapping the phrases gave you three advantages first you just need to write a short phrase in place of the whole sentence in the target language. During training processes system generate partial translation or nearly complete translation of an input document. We just need to check or mapping new words in generated files. Second is your phrase table size will be very small compared to automatically generated phrase table it will make a decoding process more efficient. Third, a linguistic person needs less time to generate parallel phrases then parallel sentences.

2) **Language model:** The language model is responsible for generating natural language. The system has been used Kneser-Ney smoothing algorithm to generate language model (Chen and Goodman 1998). Kneser-Ney is an extension of Absolute Discounting and provides state of the art solution for predicting next word. Absolute Discounting method is defined as:

$$P_{AbsDis}(w_i|w_{i-1}) = \frac{c(w_{i-1}, w_i) - d}{c(w_{i-1})} + \lambda(w_{i-1})P(w) \quad (16)$$

Kneser-Ney is a refined version of Absolute Discounting and gave a better prediction on lower order models when higher order modes have no count present. Following equation shows the second order Kneser-Ney model.

$$P_{KneserNey}(w_i|w_{i-1}) = \frac{\max(c(w_{i-1}, w_i) - d, 0)}{c(w_{i-1})} + \lambda(w_{i-1})P_{Conti}(w_i) \quad (17)$$

Where λ is normalized constant, defined as:

$$\lambda w_{i-1} = \frac{d}{c(w_{i-1})}|\{w: c(w_{i-1}, w) > 0\}| \quad (18)$$

Where $\{w: c(w_{i-1}, w) > 0\}$ is number of word types that can fallow, w_{i-1}.

$P_{Conti}(w_i)$ used as a replacement of maximum likelihood of unigram probabilities with continuation probability that estimate how likely the unigram is to continue in a new context. Continuation probability distribution defined as:

$$P_{Conti}(w) = \frac{|\{w: c(w_{i-1}, w) > 0\}|}{\sum_{w'}|\{w': c(w'_{i-1}, w') > 0\}|} \quad (19)$$

$|\{w: c(w_{i-1}, w) > 0\}|$: Where numerator equation is a count of different word types before the word w.

$\sum_{w'}|\{w': c(w'_{i-1}, w') > 0\}|$: Denominator equation is a normalized factor, total count of different words preceding the all words. Recursive formation of kneser-Ney for higher order model defined as:

$$P_{KneserNey}(w_i|w_{i-n+1}^{i-1}) = \frac{\max\left(C\left(w_{i-n+1}^{i-1}\right) - d, 0\right)}{C\left(w_{i-n+1}^{i-1}\right)} + \lambda(w_{i-n+1}^{i-1}) P_{Conti}(w_i|w_{i-n+1}^{i-1}) \quad (20)$$

To form the language model we have used a mixture of phrase and word-based language model. Generally, machine translation systems and other NLP applications used word-based language model. We have tried to develop phrase-based model along with word-based model which gives accurate predictions to choose correct phrases or word to generate target language. The system generates phrase separator training data files to generate phrase and word-based language model file shown in Appendix 2. Changes have been made in language model training data to reduce vocabulary size. For example, we have changed all numeric tokens with a unique token like 22.201 and 545.1 numeric values with 11.111 and 111.1 respectively. Changing the numeric token with unique tokens helped smoothing algorithm to efficiently predict phrase sequence with the same pattern with different numeric tokens for example.

He paid $50 to shopkeeper.

He paid $30 to shopkeeper.

Both these sentences changed to:

He paid $11 to shopkeeper.

Along with numeric patterns, we changed patterns like an email address to unique token [e@e] which helped us to decrease the size of a language model.

D. **Decoding:** Decoding problem find the most likely state sequence from given observation $O = o_1, o_2, o_3 \ldots . o_n$, to decoding the Hidden Markov Model and find the state sequence with the maximum likelihood the system had used Viterbi algorithm. The sequence of states is backtracked after decoding the whole sequences.

ALGORITHM 2. Viterbi

Input: a Sentence
$x_1 \ldots x_n$ *and Parameters* $q(t_n|t_{n-1}), e(s_i|t_i)$
Define K to set of all tags. $K_{-1=}K_0 = (start)$
$\pi(0, start, start)$=1
For $k = 1 \ldots . n$
For $a \in K_{k-1}, b \in K_k$
$\pi(k, a, b) = argmax(\pi(k - 1, a, b) * q(b|a) * e(x_k|b))$
Return $\arg\max(\pi(n, b) * q(stop|b))$

ALGORITHM 3. Complete Translation Process

Read input in UrduInputText
Tokenization and Segmentation UrduInputText in TokensList[]
Classify TokensList[] Text in Classes[]
Load DomainPhraseTables[] according to Classes[]
Load LanguageModel[]

For each Token in Tokens[]
Decode TranslationModel[] and LanguageModel[] using Veterbi

End For
Return Translation

V. Experiment and Evaluation

The system has been evaluated using BLEU score which is automatic evaluation metric (Papineni et. Al 2002) and evaluated by human evaluators which were a monolingual non-expert translators have knowledge of only target language. Where BLEU score range between 0 > 1 and for manually checking we have set four parameters as shown below.

TABLE III. Manually Evaluation Scores

Score	Cause
0	Very Poor
1	Partially Okay
2	Good with few errors
3	Excellent

For BLEU score based evaluation, one target translation reference has been used to calculate a score which was prepared by same linguistic experts those who prepared training data. For incremental training, all training data was collected from BBC Urdu website. The system has been evaluated after every 100 training documents. BLEU scores for per domain shown in chart 1 to chart 5.

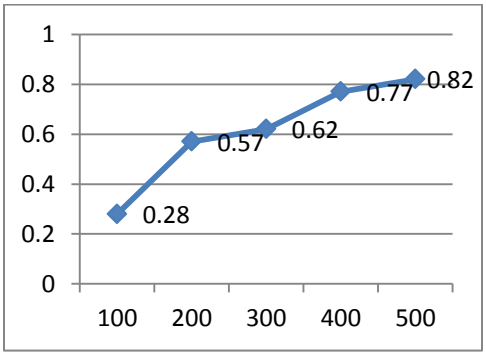


Chart 1: Political News Accuracy

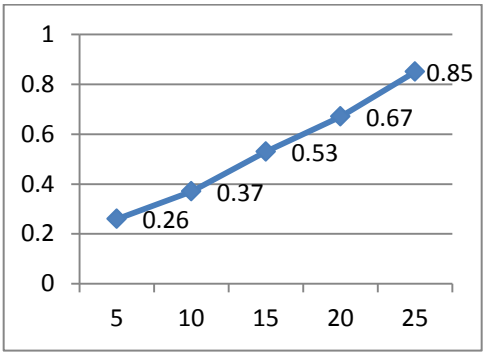


Chart 2: Tourism News Accuracy

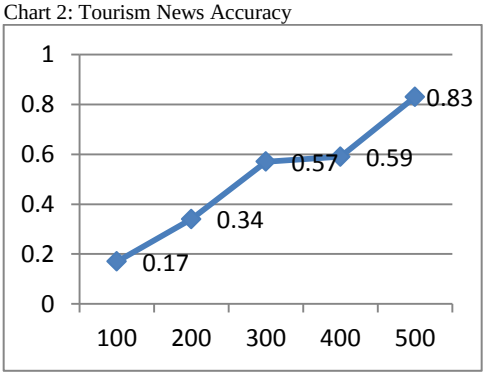


Chart 3: Entertainment News Accuracy

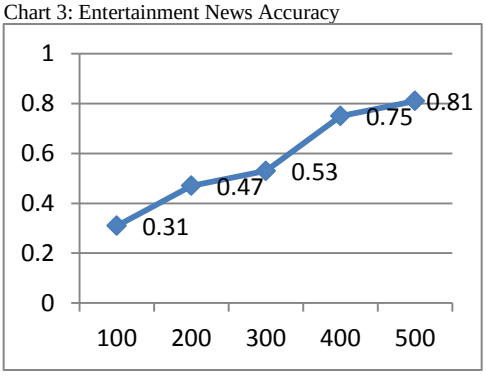


Chart 4: Sports News Accuracy

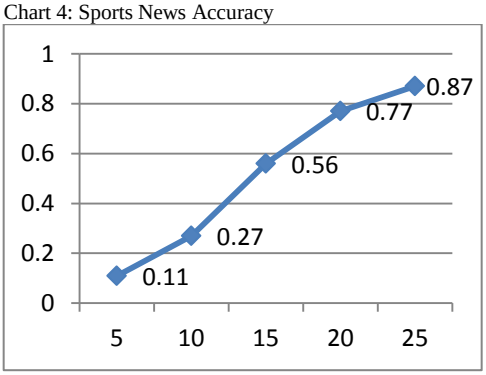


Chart 5: Health News Accuracy

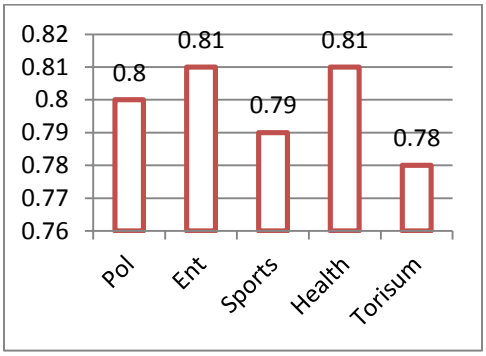


Chart 6: Overall Accuracy without Text Classifier

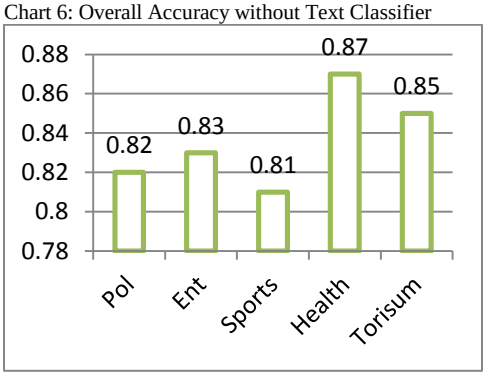


Chart 7: Overall Accuracy with Text Classifier

Manual testing was performed at the end of the training section. Test set contained 10 documents from each domain combined 1123 sentences. In manual testing 85% sentences got score 3 and 2 and 10% sentences got score 1 and remaining got score 0 which are new to the system and overall BLEU score was 0.86 for the same set of sentences. The text classifier before translation showed an increase in overall accuracy. The text classifier helped translation algorithm to pick correct translations phrases according to the domain of input text. The text classifier was evaluated using standard metrics as shown below.

Manual testing was performed at the end of the training section. Test set contained 10 documents from each domain combined 1123 sentences. In manual testing 85% sentences got score 3 and 2 and 10% sentences got score 1 and remaining got score 0 which are new to the system and overall BLEU score was 0.86 for the same set of sentences. The text classifier before translation showed an increase in overall accuracy. The text classifier helped translation algorithm to pick correct translations phrases according to the domain of input text. The text classifier was evaluated using standard metrics as shown below.

$$Recall = \frac{c_{ii}}{\sum_j c_{ij}} \quad (21)$$

$$Precision = \frac{c_{ii}}{\sum_j c_{ji}} \quad (22)$$

$$Accuracy = \frac{\sum_i c_{ii}}{\sum_j \sum_i c_{ij}} \quad (23)$$

TABLE IV. CONFUSION MATRIX OF TEXT CLASSIFIER

Documents	Assigned to Political	Assigned to Entertainment	Assigned to Sports	Assigned to Tourism	Assigned to Health
Political	471	8	3	0	0
Entertainment	13	482	7	0	0
Sports	14	6	487	0	0
Tourism	2	4	3	25	0
Health	0	0	0	0	25

TABLE V. PER CLASS RECALL AND PRECISION

	Recall	Precision
Political	0.977	0.942
Entertainment	0.960	0.964
Sports	0.960	0.974
Tourism	0.735	1
Health	1	1

The text classifier able to classify any given text document with overall accuracy 0.961. The text classifier was failed when document did not contain sufficient text to classify or text was very ambiguous for classifier like a political document which contains more sports related text than politics.

Our experiment shows that simple statistical model like HMM also yields good results for the closely related language pair. HMM based model quite popular in the field of part of speech (POS) tagging and Named Entity (NE) tagging and researcher showed really good results for sequence tagging NLP applications. Various researchers [Thorsten Brants, 200] had been shown that with a good amount of training tokens even simple statistical model also perform well compared to MaxEnt etc.

Appendix 3 shows that sample output and comparison of Google translator and our machine translation system. The proposed system generates nearly perfect or perfect translation of given text compared to Google translator which generates grammatical incorrect, meaningless and partial output in all cases. The system's output was compared with all five domains. Urdu inputs examples were quite simple without any ambiguous words.

The comparison is difficult between both systems because both systems used different training data sets, but we had checked the entire words list manually on Google translator and nearly all words were in its translation database, but decoder was not able to translate the input text by using its knowledge base. Google translator has very rich phrase translation database but the translation is still quite poor for Urdu-Punjabi language pair.

VI. CONCLUSION

The Paper has presented incremental learning based Urdu to Punjabi machine translation system. In place of parallel corpus, where system learns parameters from parallel sentences of source and target language. The proposed system used manually mapped parallel phrases training data and learned the parameters for translation model and language model rather than using parallel sentences corpus. In preprocessing phase, the system has used rules for segmentation, tokenization and text classification system to translate given text according to a preferred domain which also helped translation system to improve overall accuracy. The system has been trained and tested for Urdu Punjabi language pair which is closely related languages and share grammatical structure and vocabulary. Urdu and Punjabi languages are resources-poor languages and one should need a huge amount of parallel corpus to train statistical machine translation model to get decent accuracy. In our learning method, the system has able to achieve 0.86 BLEU score which is relatively good compared to other statistical translation systems. Like Urdu and Punjabi, many other Asian languages are resource poor languages and this approach can be applied straight away for other closely related language pairs.

ACKNOWLEDGEMENT

We are thankful to Technology Development for Indian Languages (TDIL) for supporting us and providing the Urdu Punjabi parallel corpus of Health and Tourism domain which is used for the evaluation and comparison.

REFERENCES

[1] Brants, Thorsten, TnT -- A Statistical Part-of-Speech Tagger, In proceeding of the 6th Applied NLP Conference, pages 224-231 2000

[2] Chen and Goodman 1998, "An Empirical Study of Smoothing Techniques for Language Modeling

[3] Durrani, Nadir et.al. "Urdu word segmentation." Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Association for Computational Linguistics. 2010

[4] Durrani, Nadir et.al "Hindi-to-Urdu Machine Tarnation through transliteration" Processing of the 48th Annual Meeting of the Association for Computational Linguistics, pages 465-474 2010

[5] Goyal, Vishal et.al, Evaluation of Hindi Punjabi Machine Translation System, IJCSI International Journal of Computer Science Issues, Vol. 4 No. 1, pages: 36-39 2009

[6] Garje G V, Survey of Machine Translation Systems in India, International Journal on Natural language Computing Vol 2, No4, pages: 47-67 Oct 2013

[7] Josan, Gurpreet Singh, A Punjabi To Hindi Machine Translation System, Companion volume – Posters and Demonstration, pages: 157-160 2008

[8] Kansal, Rohit et.al, "Rule Based Urdu Stemmer", processing of Colling 2012, Demonstration paper, pages 267-276 2012

[9] Lehal, Gurpreet Singh. A word segmentation system for handling space omission problem in Urdu script. 23rd International Conference on Computational Linguistics. 2010

[10] Lehal, G. A Two Stage Word Segmentation System for Handling Space Insertion Problem in Urdu Script. World Academy of Science, Engineering and Technology 60. 2009

[11] L. Rabiner, A Tutorial on Hidden Markov Model and Selected Application in Speech Recognition, in Proceeding on the IEEE, Vol. 77, Issue. 2, pages: 257-286 1989

[12] Papinni, Kishore, (2002), Bleu: a method for automatic evaluation of machine translation, ACL-2002: 40th Annual meeting of the Association for Computational Linguistics. pages 311–318

[13] Singh, UmrinderPal et.al (2012) "Named Entity Recognition System for Urdu". Processing of Colling 2012 pages: 2507-2518

[14] Sani, Tajinder Singh (2011) "Word Disambiguation in Shahmukhi to Gurmukhi Transliteration", Processing of the 9th Wordshop on Asian Language Resources, Chiang Mai, Thailand, November 12 and 13, 2011 pages: 79-87

APPENDIX 1 (TRANSLATION MODEL TRAINING FILE)

Training File	
1	پاکستانی کرکٹ ٹیم|||[ਪਾਕਿਸਤਾਨੀ ਕ੍ਰਿਕੇਟ ਟੀਮ]
2	نیوزی لینڈ کے خلاف|||[ਨਿਊਜੀਲੈਂਡ ਦੇ ਖਿਲਾਫ]
3	پہلا|||[ਪਹਿਲਾ]
4	میچ|||[ਮੈਚ]
5	31|||[31]
6	جنوری کو|||[ਜਨਵਰੀ ਨੂੰ]
7	ویلنگٹن|||[ਵੈਲਿੰਗਟਨ]
8	میں|||[ਵਿੱਚ]
9	جبکہ|||[ਜਦੋਂ ਕਿ]
10	دوسرا|||[ਦੂਜਾ]
11	تین|||[ਤਿੰਨ]
12	فروری کو|||[ਫਰਵਰੀ ਨੂੰ]
13	نیپیئر|||[ਨੇਪੀਅਰ]
14	کھیلے میں|||[ਵਿੱਚ ਖੇਡੇ]
15	گی|||[ਗਈ]
16	-|||[।]
17	پہلا|||[ਪਹਿਲਾ]
18	میچ|||[ਮੈਚ]
19	نو|||[ਨੌਂ]
20	فروری کو|||[ਫਰਵਰੀ ਨੂੰ]
21	بنگلہ دیش کے|||[ਬਾਂਗਲਾਦੇਸ਼ ਦੀ]
22	خلاف|||[ਖਿਲਾਫ]
23	اور دوسرا|||[ਅਤੇ ਦੂਜਾ]
24	میچ|||[ਮੈਚ]
25	11|||[11]
26	فروری کو|||[ਫਰਵਰੀ ਨੂੰ]
2	انگلینڈ|||[ਇੰਗਲੈਂਡ]
28	کے خلاف|||[ਦੇ ਖਿਲਾਫ]
29	کھیلا|||[ਖੇਡਿਆ]
30	جائے گا|||[ਜਾਵੇਗਾ]
31	-|||[।]
32	دلچسپ بات یہ|||[ਦਿਲਚਸਪ ਗੱਲ ਇਹ]
33	ہے کہ کپتان|||[ਹੈ ਕਿ ਕਪਤਾਨ]
34	مصباح الحق آسٹریلیا|||[ਮਿਸਬਾਹ ਉਲ ਹੱਕ ਆਸਟਰੇਲਿਆ]
35	میں دو ٹیسٹ|||[ਵਿਚ ਦੋ ਟੇਸਟ]
36	میچ|||[ਮੈਚ]
37	کھیلے|||[ਖੇਡਣੇ]
38	ہیں|||[ਹਨ]

Training file with probability distribution		
1	پاکستانی کرکٹ ٹیم|||[ਪਾਕਿਸਤਾਨੀ ਕ੍ਰਿਕੇਟ ਟੀਮ]	1.0
2	نیوزی لینڈ کے خلاف|||[ਨਿਊਜੀਲੈਂਡ ਦੇ ਖਿਲਾਫ]	0.89
3	پہلا|||[ਪਹਿਲਾ]	0.90
4	میچ|||[ਮੈਚ]	1.0
5	31|||[31]	1.0
6	جنوری کو|||[ਜਨਵਰੀ ਨੂੰ]	1.0
7	ویلنگٹن|||[ਵੈਲਿੰਗਟਨ]	1.0
8	میں|||[ਵਿੱਚ]	0.76
9	جبکہ|||[ਜਦੋਂ ਕਿ]	0.98
10	دوسرا|||[ਦੂਜਾ]	0.64
11	تین|||[ਤਿੰਨ]	0.90
12	فروری کو|||[ਫਰਵਰੀ ਨੂੰ]	1.0
13	نیپیئر|||[ਨੇਪੀਅਰ]	1.0
14	کھیلے میں|||[ਵਿੱਚ ਖੇਡੇ]	1.0
15	گی|||[ਗਈ]	0.34
16	-|||[।]	1.0
17	پہلا|||[ਪਹਿਲਾ]	0.90
18	میچ|||[ਮੈਚ]	1.0
19	نو|||[ਨੌਂ]	0.98
20	فروری کو|||[ਫਰਵਰੀ ਨੂੰ]	1.0
21	بنگلہ دیش کے|||[ਬਾਂਗਲਾਦੇਸ਼ ਦੀ]	0.89
22	خلاف|||[ਖਿਲਾਫ]	0.76
23	اور دوسرا|||[ਅਤੇ ਦੂਜਾ]	0.89
24	میچ|||[ਮੈਚ]	1.0
25	11|||[11]	1.0
26	فروری کو|||[ਫਰਵਰੀ ਨੂੰ]	1.0
27	انگلینڈ|||[ਇੰਗਲੈਂਡ]	1.0
28	کے خلاف|||[ਦੇ ਖਿਲਾਫ]	0.68
9	کھیلا|||[ਖੇਡਿਆ]	0.87
30	جائے گا|||[ਜਾਵੇਗਾ]	0.89
31	-|||[।]	1.0
32	دلچسپ بات یہ|||[ਦਿਲਚਸਪ ਗੱਲ ਇਹ]	0.78
33	ہے کہ کپتان|||[ਹੈ ਕਿ ਕਪਤਾਨ]	1.0
34	مصباح الحق آسٹریلیا|||[ਮਿਸਬਾਹ ਉਲ ਹੱਕ ਆਸਟਰੇਲਿਆ]	1.0
35	میں دو ٹیسٹ|||[ਵਿਚ ਦੋ ਟੇਸਟ]	0.98
36	میچ|||[ਮੈਚ]	1.0
37	کھیلے|||[ਖੇਡਣੇ]	0.97

APPENDIX 2 (PHRASE SEPARATED FILE TO GENERATE LANGUAGE MODEL)

[ਪਾਕਿਸਤਾਨੀ ਕ੍ਰਿਕੇਟ ਟੀਮ] [ਆਪਣੇ] [ਸਭ ਤੋਂ ਵੱਡੇ] [ਇਮਤਿਹਾਨ] [ਦੇ ਲਈ] [ਮੰਗਲਵਾਰ] [ਦੀ ਰਾਤ] [ਕ੍ਰਿਕੇਟ ਦੇ ਵਿਸ਼ਵ ਕੱਪ] [ਦੀ ਯਾਤਰਾ ਤੇ] [ਰਵਾਨਾ ਹੋ] [ਰਿਹਾ ਹੈ] [।] [ਇਸ] [ਸਫਰ] [ਵਿੱਚ ਓਹ] [ਪਹਿਲਾ] [ਪੜਾਵ] [ਨਿਊਜੀਲੈਂਡ ਵਿੱਚ] [ਪਾਏ] [ਗਈ] [ਜਿੱਥੇ] [ਉਸਨੂੰ] [ਨਿਊਜੀਲੈਂਡ ਦੇ ਖਿਲਾਫ] [ਦੋ] [ਇੱਕ ਰੋਜਾ] [ਅੰਤਰਰਾਸ਼ਟਰੀ] [ਮੈਚ] [ਖੇਡਣੇ] [ਹਨ ਜਿਸਦੇ] [ਬਾਅਦ ਉਹ] [ਆਸਟਰੇਲਿਆ] [ਪਹੁੰਚ] [ਜਾਵੇਗੀ] [।]

[ਪਾਕਿਸਤਾਨੀ ਕ੍ਰਿਕੇਟ ਟੀਮ] [ਨਿਊਜੀਲੈਂਡ ਦੇ ਖਿਲਾਫ] [ਪਹਿਲਾ] [ਮੈਚ] [11] [ਜਨਵਰੀ ਨੂੰ] [ਵੈਲਿੰਗਟਨ] [ਵਿੱਚ] [ਜਦੋਂ ਕਿ] [ਦੂਜਾ] [ਤਿੰਨ] [ਫਰਵਰੀ ਨੂੰ] [ਨੇਪੀਅਰ] [ਵਿੱਚ ਖੇਡੇ] [ਗਈ] [।]

[ਪਾਕਿਸਤਾਨੀ ਟੀਮ] [ਵਿਸ਼ਵ ਕੱਪ ਤੋਂ] [ਪਹਿਲਾਂ] [ਸਿਡਨੀ] [ਵਿੱਚ ਦੋ] [ਵਾਰਮ ਅੱਪ] [ਮੈਚ] [ਵੀ] [ਖੇਡੇ] [ਗਈ] [।]

Bigram Phrase based Language Model

Bigram Phrases: $count(phrase_{i-1}, phrase_i)$	Count
ਪਾਕਿਸਤਾਨੀ ਕ੍ਰਿਕੇਟ ਟੀਮ, ਆਪਣੇ	10
ਆਪਣੇ, ਸਭ ਤੋਂ ਵੱਡੇ	8
ਸਭ ਤੋਂ ਵੱਡੇ, ਇਮਤਿਹਾਨ	1
ਇਮਤਿਹਾਨ, ਦੇ ਲਈ	1
ਦੇ ਲਈ, ਮੰਗਲਵਾਰ	2
ਮੰਗਲਵਾਰ, ਦੀ ਰਾਤ	2
ਦੀ ਰਾਤ, ਕ੍ਰਿਕੇਟ ਦੇ ਵਿਸ਼ਵ ਕੱਪ	3

Bigram Word based Language Model

Bigram Words: $count(w_{i-1}, w_i)$	Count
ਪਾਕਿਸਤਾਨੀ, ਕ੍ਰਿਕੇਟ	123
ਕ੍ਰਿਕੇਟ ਟੀਮ	421
ਟੀਮ, ਆਪਣੇ	321
ਆਪਣੇ, ਸਭ	131
ਸਭ, ਤੋਂ	634
ਤੋਂ, ਵੱਡੇ	102
ਵੱਡੇ, ਇਮਤਿਹਾਨ	1

APPENDIX 3

Output comparison of all domain. Mistakes are underlined.

Political Input Text	پاکستان کی بّری فوج کے سربراہ جنرل راحیل شریف نے کہا ہے کہ سکیورٹی فورسز چین پاکستان اقتصادی راہداری منصوبے کے خلاف چلائی جانے والی تمام مہمات سے آگاہ ہیں اور اس شاندار منصوبے کو حقیقت کا رنگ دینے کے لیے کسی قربانی سے بھی دریغ نہیں کریں گی۔
Our Translator output	ਪਾਕਿਸਤਾਨ ਦੀ ਫੋਜ ਦੇ ਪ੍ਰਮੁੱਖ ਜਨਰਲ ਰਅਹੀਲ ਸ਼ੀਫ ਨੇ ਕਿਹਾ ਹੈ ਕਿ ਸੁਰੱਖਿਆ ਬਲਾਂ ਚੀਨ ਪਾਕਿਸਤਾਨ ਆਰਥਕ ਰਾਹਦਾਰੀ ਯੋਜਨਾ ਦੇ ਖਿਲਾਫ ਲੀਕੇ ਜਾਣ ਵਾਲੀਆਂ ਸਭ ਮੁਹਿੰਮਾ ਦੀ ਜਾਣੂ ਹਨ ਅਤੇ ਇਸ ਸ਼ਾਨਦਾਰ ਯੋਜਨਾ ਨੂੰ ਸਚਾਈ ਦਾ ਰੰਗ ਦੇਣ ਲਈ ਕਿਸੇ ਕੁਰਬਾਨੀ ਤੋਂ ਵੀ ਗੁਰੇਜ ਨਹੀਂ ਕਰਨ ਗਏ ।
Google Translator output	ਫੌਜ ਸਟਾਫ ਜਨਰਲ ਰਾਹੀਲ ਸ਼ਰੀਫ ਦੇ ਪਾਕਿਸਤਾਨ ਦੇ ਪ੍ਰਧਾਨ ਨੇ ਕਿਹਾ ਹੈ ਕਿ ਪਾਕਿਸਤਾਨ ਦੇ ਸੁਰੱਖਿਆ ਬਲ, ਚੀਨ ਲਈ ਸਭ ਮੁਹਿੰਮ ਆਰਥਿਕ ਕੋਰੀਡੋਰ ਪ੍ਰਾਜੈਕਟ ਨੂੰ ਦੇ ਖਿਲਾਫ ਸ਼ੁਰੂ ਕੀਤੀ ਹੈ ਅਤੇ ਇਸ ਤੱਥ ਦੇ ਜਾਣੂ ਹਨ ਬਲੀ ਨਾ ਕਰੇਗਾ, ਜੋ ਕਿ ਇਸ ਸ਼ਾਨਦਾਰ ਪ੍ਰਾਜੈਕਟ ਦੀ ਰੰਗ ਹੈ.
Health Input Text	منہ کے اندرونی حصے کی مختلف بیماریوں میں دانتوں کا کم زور ہو جانا، کیڑا لگنا، دانتوں کا ٹیڑھا ہو جانا اور مسوڑھوں کا انفیکشن عام ہے۔ دانتوں اور مسوڑھوں کی زیادہ تر بیماریوں کی علامات فوراً ظاہر نہیں ہوتیں، اور شدید تکلیف کا باعث بنتی ہیں۔
Our Translator output	ਮੁੰਹ ਦੇ ਅੰਦਰੂਨੀ ਹਿੱਸੇ ਦੀ ਵਿਭਿੰਨ ਬਿਮਾਰੀਆਂ ਵਿੱਚ ਦੰਦਾਂ ਦਾ ਕੰਮਜੋਰ ਹੋ ਜਾਣਾ , ਕੀੜਾ ਲੱਗਣਾ , ਦੰਦਾਂ ਦਾ ਟੇਢਾ ਹੋ ਜਾਣਾ ਅਤੇ ਮਸੂੜਿਆਂ ਦੀ ਇੰਫੈਕਸ਼ਨ ਆਮ ਹੈ । ਦੰਦਾਂ ਅਤੇ ਮਸੂੜਿਆਂ ਦੀਆਂ ਜ਼ਿਆਦਾਤਰ ਬਿਮਾਰੀਆਂ ਦੇ ਲਕਸ਼ਨ ਤੁਰੰਤ ਪ੍ਰਗਟ ਨਹੀਂ ਹੁੰਦੇ , ਅਤੇ ਗੰਭੀਰ ਤਕਲੀਫ ਦਾ ਕਾਰਨ ਬਣਦੀਆਂ ਹਨ ।
Google Translator output	ਮੂੰਹ ਤਣਾਅ ਨੂੰ ਘੱਟ ਕਰਨ ਲਈ, ਕੀੜਾ ਲੱਗਦਾ ਹੈ, ਟੇਢੇ ਦੰਦ ਅਤੇ ਗਮ ਲਾਗ ਜਾਣ ਦੇ ਅੰਦਰ ਦੰਦ ਦੇ ਵੱਖ-ਵੱਖ ਰੋਗ ਵਿੱਚ ਆਮ ਹਨ. ਜ਼ਿਆਦਾਤਰ ਦੰਦ ਅਤੇ ਗੰਮ ਦੀ ਬਿਮਾਰੀ ਦੇ ਲੱਛਣ ਤੁਰੰਤ ਵਿਖਾਈ ਨਾ ਕਰੇ, ਅਤੇ ਗੰਭੀਰ ਦਰਦ ਦਾ ਕਾਰਨ ਬਣ.
Entertainment Input Text	کبھی جڑواں بھائیوں میں فرق ظاہر کرنے کے لیے ایک مونچھ یا تل سے فرق دکھانے والے میک اپ آرٹسٹ اب پورے چہرے کا میک اپ ہی الگ طریقے سے ڈیزائن کرتے ہیں۔

Our Translator output	ਕਦੇ ਜੁੜਵਾਂ ਭਰਾਵਾਂ ਵਿੱਚ ਫਰਕ ਦਿਖਾਉਣ ਲਈ ਇੱਕ ਮੁੱਛਾ ਜਾਂ ਤਿਲ ਨਾਲ ਫਰਕ ਵਿਖਾਉਣ ਵਾਲੇ ਮੇਕਅੱਪ ਆਰਟਿਸਟ ਹੁਣ ਪੂਰੇ ਚਿਹਰੇ ਦਾ ਮੇਕਅੱਪ ਹੀ ਅਲੱਗ ਤਰਾਂ ਨਾਲ ਡੀਜਾਈਨ ਕਰਦੇ ਹਾਂ ।
Google Translator output	Twin ਬਣਤਰ ਕਲਾਕਾਰ, ਜੋ ਕਿ ਇੱਕ ਹੋ◌ੇਏ ਜ ਤਿਲ ਨੂੰ ਦਿਖਾਉਣ ਲਈ ਭਰਾ ਸਾਰੀ ਚਿਹਰੇ ਨੂੰ ਵੱਖ ਵੱਖ ਢੰਗ ਨੂੰ ਤਿਆਰ ਕਰ ਰਹੇ ਹਨ ਫ਼ਰਕ ਕਦੇ ਹੋਵੇਗਾ.
Tourism Input Text	لاطینی امریکہ کے ملک ارجنٹینا میں ساحل سمندر پر جانے والے ان افراد پر سخت نکتہ چینی کی جاری ہے جنھوں نے ناپید ہونے والی نسل کی ایک ڈولفن کے ساتھ سیلفی لینے کے لیے اسے سمندر سے باہر نکال لیا.
Our Translator output	ਲੈਟੀਨ ਅਮਰੀਕਾ ਦੇ ਦੇਸ਼ ਅਰਜਨਟੀਨਾ ਵਿੱਚ ਸਮੁੰਦਰੀ ਤੱਟ ਉੱਤੇ ਜਾਣ ਵਾਲੇ ਉਨ੍ਹਾਂ ਲੋਕਾਂ ਦਾ ਕਠੋਰ ਵਿਰੋਧ ਜਾਰੀ ਹੈ ਜਿਨ੍ਹਾਂ ਨੇ ਲੁਪਤ ਹੋਣ ਵਾਲੀ ਨਸਲ ਦੀ ਇੱਕ ਡਾੱਲਫਿਨ ਦੇ ਨਾਲ ਸੈਲਫ਼ੀ ਲੈਣ ਲਈ ਉਸਨੂੰ ਸਮੁੰਦਰ ਤੋਂ ਬਾਹਰ ਕੱਢ ਲਿਆ ।
Google Translator output	ਨਸਲ, ਜੋ ਜਿਹੜੇ ਬੀਚ 'ਤੇ ਜਾਣ ਦੀ ਆਲੋਚਨਾ ਕੀਤੀ ਹੈ ਦੇ ਸਵੈ-ਤਬਾਹ ਲੈ ਲਈ ਇੱਕ ਡਾਲਫਿਨ ਨਾਲ ਲਾਤੀਨੀ ਅਮਰੀਕੀ ਦੇਸ਼ ਵਿਚ ਅਰਜਨਟੀਨਾ ਸਮੁੰਦਰ ਨੂੰ ਉਸ ਨੂੰ ਬਾਹਰ ਲੈ ਗਿਆ.
Sports Input Text	بھارتی کرکٹ بورڈ نے کہا ہے کہ قومی کرکٹ ٹیم کے کپتان مہندر دھونی پیر کو معمول کی تربیت کے دوران کمر کے درد میں مبتلا ہونے کے بعد ایشیا کپ میں ٹیم کا حصہ نہیں ہوں گے۔
Our Translator output	ਭਾਰਤੀ ਕ੍ਰਿਕੇਟ ਬੋਰਡ ਨੇ ਕਿਹਾ ਹੈ ਕਿ ਰਾਸ਼ਟਰੀ ਕ੍ਰਿਕੇਟ ਟੀਮ ਦੇ ਕਪਤਾਨ ਮਹਿੰਦਰ ਧੋਨੀ ਸੋਮਵਾਰ ਨੂੰ ਰੁਟੀਨ ਸਿਖਲਾਈ ਦੇ ਦੌਰਾਨ ਕਮਰ ਦੇ ਦਰਦ ਤੋਂ ਪੀੜਿਤ ਹੋਣ ਦੇ ਬਾਅਦ ਏਸ਼ਿਆ ਕੱਪ ਵਿੱਚ ਟੀਮ ਦਾ ਹਿੱਸਾ ਨਹੀਂ ਹੋਣਗੇ ।
Google Translator output	ਭਾਰਤੀ ਕ੍ਰਿਕਟ ਟੀਮ ਦੇ ਕਪਤਾਨ ਮਹਿੰਦਰ ਸਿੰਘ ਧੋਨੀ ਨੇ ਕਿਹਾ ਹੈ ਕਿ ਸੋਮਵਾਰ ਨੂੰ ਰੁਟੀਨ ਦੀ ਸਿਖਲਾਈ ਦੌਰਾਨ ਪਿੱਠ ਦੇ ਦਰਦ ਨਾਲ ਪੀੜਤ ਦੇ ਬਾਅਦ, ਏਸ਼ੀਆਈ ਕੱਪ 'ਚ ਟੀਮ ਦਾ ਹਿੱਸਾ ਨਾ ਹੋਵੇਗਾ.

Holistic Evaluation Framework for Automated Bug Triage Systems: Integration of Developer Performance

Dr.V.Akila
Dept. of Computer Science and Engineering
Pondicherry Engineering College
Pondicherry, India

Dr.V.Govindasamy
Dept. of Information Technology
Pondicherry Engineering College
Pondicherry, India

***Abstract*—Bug Triage is an important aspect of Open Source Software Development. Automated Bug Triage system is essential to reduce the cost and effort incurred by manual Bug Triage. At present, the metrics that are available in the literature to evaluate the Automated Bug Triage System are only recommendation centric. These metrics address only the correctness and coverage of the Automated Bug Triage System. Thus, there is a need for user-centric evaluation of the Bug Triage System. The two types of metrics to evaluate the Automated Bug Triage System include Recommendation Metrics and User Metrics. There is a need to corroborate the results produced by the Recommendation Metrics with User Metrics. To this end, this paper furnishes a Holistic Evaluation Framework for Bug Triage System by integrating the developer performance into the evaluation framework. The Automated Bug Triage System is also to retrieve a set of developers for resolving a bug. Hence, this paper proposes Key Performance Indicators (KPI) for appraising a developer's effectiveness in contribution towards the resolution of the bug. By applying the KPIs on the retrieved set of developers, the Bug Triage System can be evaluated quantitatively.**

***Keywords*—Bug Management; Bug Triage; Recommendation Metrics; Key Performance Indicators; Developer Performance; Bug Resolution Time**

I. INTRODUCTION

Open Source Software (OSS) is a commercial software where full access to the code for viewing, modification, and redistribution is granted to all the users by agreeing to a free-of-cost license. Bug management is a central component of the software maintenance of the OSS. Bug Management in OSS is usually performed using Bug Management Systems like Bugzilla. The new bugs that arise after the deployment of a new version of software are first reported to the Bug Management system. The new bugs are manually verified and important attributes like Component and Severity are fixed. Following this, the bugs are assigned to a developer for resolution by a human triager. In summary, Bug management comprises the following three activities: (i) Bug Triaging, (ii) bug assignment to the software developer for solution and (iii) solving of the bug. Software maintenance expenditure is about 50% of the overall expenditure of the software project. In OSS development, the expenditure translates to time. Bug Triaging comprises checking for validity of the bug, assigning priority, severity and assigning the bug to a correct software developer. Manual Bug Triaging is time- consuming and fault prone [1],[2],[3].

The bugs are reported to the Bug Management System. The reported bug is verified for validity and is assigned a new tag and is assigned to a developer. If the developer is unable to resolve the bug he may reassign the bug to a new developer. This activity is captured in a bug tossing graph. The summary that is in the bug report and the Bug Tossing graph serves as the basis for any Automated Bug Triage system. The metrics that are used to evaluate the Automated Bug Triage system are: (i) Accuracy (ii) Precision(iii) Recall and (iv) Mean Steps To Resolve. Precision is a better parameter for software developer recommendation because the cost of false recommendation is much higher than in search engine. Further, the Mean Steps to Resolve parameter encodes only the number of steps in the predicted path. While the reduction in the number of steps to resolve is required, it is also vital to compare how far the predicted path is similar to the original path. The structure and the ordering of nodes in the predicted path needs to be compared with that of the original path. Metrics based on graph edit distance were used for this purpose[4][5].

The evaluation of the Automated Bug Triage System is based only on the Recommendation metrics[6]. This paper integrates the developer performance in the evaluation framework of the Automated Bug Triage System. The quality of the developer extracted by the Automated Bug Triage System is evaluated by the Key Performance Indicators.

II. RELATED WORK

The related work has been studied with a perspective of metrics used for evaluation of the Automated Bug Triage System. The summary of the survey is depicted in the Table 1. It is observed from the survey that the Automated Bug Triage System is evaluated only with Recommendation Metrics. It is essential to integrate the User metrics into the evaluation process in order to build confidence in the Automated Bug Triage System. The user metrics are built over the Developer Performance. The following section shows the developer performance assessment incorporated in Open Source Systems.

TABLE I. SUMMARY OF THE SURVEY

Paper	Machine Learning	Information Retrieval				Bug Tossing Graph
	Accuracy	Precision	Recall	F-Measure	Mean Reciprocal Rank	Mean Steps to Resolve
2]	✓	-		-	-	✓
1]	✓	-		-	-	✓
3]	✓	-		-	-	✓
7]	✓	-		-	-	✓
8]	✓	-		-	-	-
9]	✓	-		-	-	-
10]		-		✓	✓	-
11]		✓		✓	-	-
12]		✓		-	-	-
13]		-		-	-	-
14]		-		-	-	✓

A. Developer Performance Assessment

Developer Performance Assessment is a necessity in identifying the strength and weakness of a developer, for career advancement, and fine tuning a business organization[15]. The contribution of a developer towards the software maintenance is quite different from a developer's contribution in developing a software product. Measuring a developer's contribution towards the maintenance of an OSS System is even more complicated. This complication is due to the fact that there are no explicitly assigned roles for the developers. However, there are different roles a developer may assume in the course of bug resolution.

The different roles that the developer may play in the bug resolution process are reporter of a bug, triager, commenter, and assignee [16]. In OSS, usually, there are metrics for evaluating the bug characteristics. These metrics focus on the program slicing characteristics of the bug like a number of lines of code affected by the bug and Cyclomatic Complexity of the bug. [17]. However, these metrics are underutilized in evaluating the Bug Triage System. Further, the developer's performance may be assessed based on Buggy commits, code contributions, and priority bugs. Buggy commits are used to identify developers who performed less buggy commits. Code contribution is measured regarding code addition, code removal, method addition, and method modification. The developer may also be assessed in terms of the number of high priority bugs that he has resolved [18]. In most of the existing works, developer's performance assessment is treated as an independent module. In the following section, the developer's performance assessment is integrated into the evaluation of the Bug Triage System. There are several KPIs proposed to assess the developer. These indicators are further utilized in quantifying the performance of the Bug Triage System.

B. Key Observations from the Dataset

This section gives a brief preview of the various factors that affect the bug resolution which is observed in the dataset. The bug reports of Eclipse project from www.bugzilla.org from 2009 to 2013 were analysed. The developers contribution for the various fields in the bug report like CC, status, Keywords, Summary priority, Assignee, and resolution are given in Figure 1. It is evident that 62% of the developers change the status of the bug to 'resolved'.

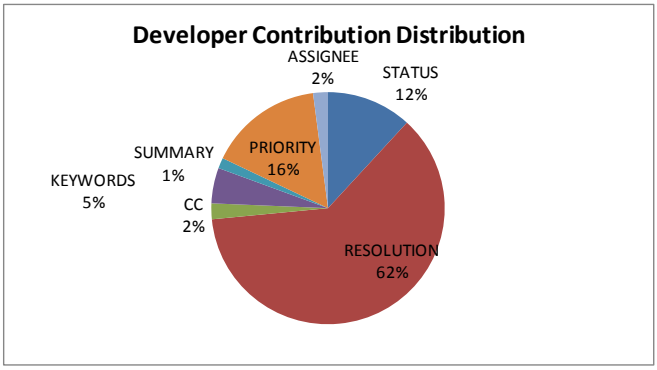


Fig. 1. Developer Contribution Distribution

The average time spent by a developer on a particular field of the bug report is given in Figure 2. As observed, the time spent to set the assignee field, status field and resolution field contributes mostly in the bug resolution time.

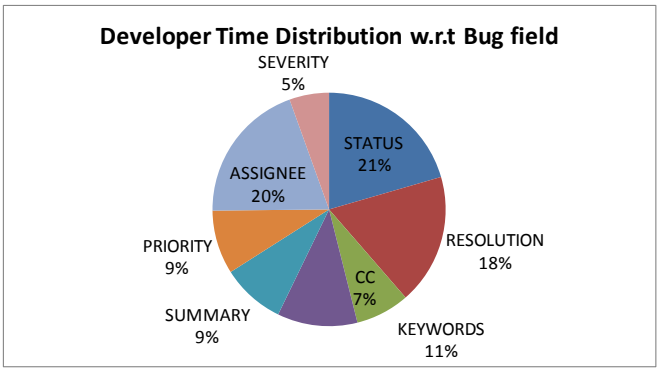


Fig. 2. Developer Time Distribution w.r.t Bug Field

The developer distribution with respect to the time spent by a developer on a particular bug is given in the Figure 3. It can be observed that 39% of the developers spend 121 to 700 days on a particular bug. Only 17% of the developers spend less than 6 months on a bug. Any Bug Triage System that extracts its set of developers mostly from this pool of 17% is a successful triage system.

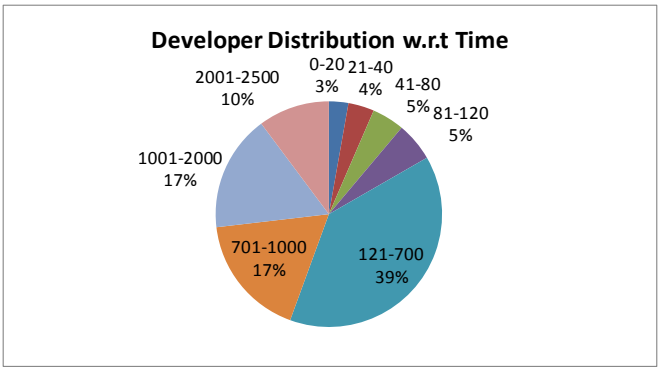


Fig. 3. Developer Distribution w.r.t Time

Figure 4 shows the Developer Distribution against the range of Bug Resolution Time. It can be observed that the most ineffective bug resolution is when the bug resolution time is more than 2 years. There are 44% of developers who spend time on bug whose resolution time is > 2 years. It can be observed from the chart that only 25% of developer has spent time in bugs that were resolved before six months. The motivation behind any Bug Triage System is to retrieve the developers from this pool of 25% of developers.

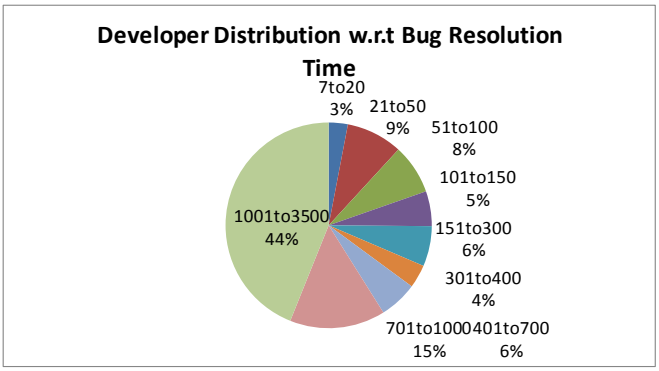


Fig. 4. Developer Distribution w.r.t Bug Resolution Time

Based on these observations, the Key Performance Indicators for assessing the Developer are introduced in the next section.

III. Key Performance Indicators for Assessing Developer Performance

The KPIs devised to evaluate the developer are Developer Time Index, Developer Effective index, and Developer Productivity. The Developer Time Index, Developer Effective index, and Developer Productivity are derived from Developer Contribution Count and Developer Contribution Time. The dependencies among the KPIs are depicted in Figure 5.

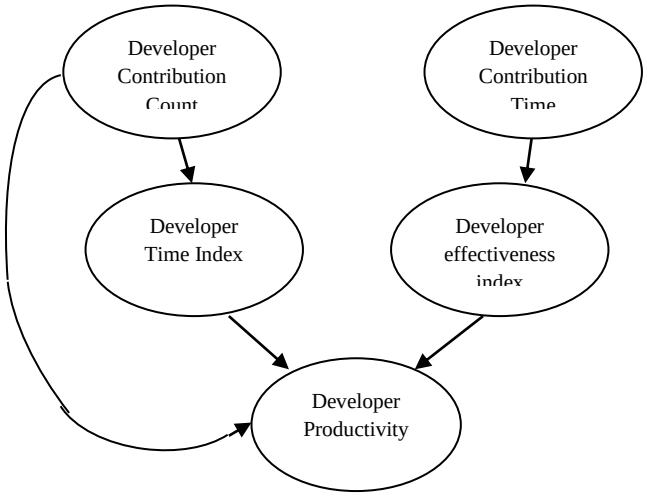


Fig. 5. Dependency in the Key Performance Indicators

The developer may have made the following types of contribution: CC, reassign the bug, change the Status field, or finally resolve the bug. For convenience sake, all the contributions are equally treated.

A. Developer Contribution Count

Developer Contribution Count (DCC) is defined as the number of contributions made by each developer in the process of resolving them.

$$DCC = \sum_{i}^{n} C_i$$

where,

C_i - Contribution by a developer to a single bug.

n - Total number of bugs assigned to a developer

B. Developer Contribution Time

Developer Contribution Time (DCT) is defined as the time taken by each developer to make a contribution on a single bug.

$$DCT = \forall \text{ Bug(Developer)} \sum(DBR - DBA)$$

where, DBR- Date of Bug Reassignment

DBA – Date a Bug Assignment

C. Developer Time Index

Developer Time Index (DTI) is defined as the ratio of DCT to DCC. This indicator captures the amount of time taken by a developer to make a single contribution.

$$DTI = \frac{DCT}{DCC}$$

D. Developer Effectiveness Index

The bug resolution time is considered to calculate the Developer Effectiveness Index (DEI). The intuition behind DEI is that, if a developer has contributed towards a bug that has been resolved with less time, then the developer's effectiveness is increased. Contrarily, if a developer has contributed towards a bug that has taken a long time to resolve, then the weight assigned to the developer is reduced.

TABLE II. WEIGHT ASSIGNMENT TABLE

Bug Resolution Time (in days)	Weights
7– 20	7
21-50	6
51-100	5
101-150	4
151-300	3
301-400	2
401-700	1
701-1000	-0.25
1001-3500	-0.50

The bugs for 10 years of Eclipse project were studied and the Resolution Time (RT) was extracted. RT varies from lower to higher values. RT was divided into nine ranges and their weights were assigned as given in Table 2. The highest weight is assigned to the range of Resolution Time that falls between 7 and 20 days. Negative weights are assigned to a range which took more than 700 days to resolve a bug.

The weights given here are inversely proportional to RT.

$$\text{Weight } (W_i) \alpha \frac{1}{RT}$$

DEI is defined as the ratio of the summation of Weights W_i of the bugs to the DCC.

$$\text{Developer Effectiveness Index DEI} = \frac{1}{DCC} \sum_{i}^{DCC} W_i$$

E. Developer Productivity

Developer Productivity (DP) is defined as the product of Developer Effectiveness Index, Developer Contribution Count, and the Developer Time Index.

$$DP = DEI * DTI * DCC$$

F. A Holistic Evaluation Framework with Developer Performance

The framework for evaluating the Bug Triage System is given in the Figure 6. The Bug Triage System extracts the optimal set of developers. KPIs of the retrieved developers are calculated and thereby, the Bug Triage System is assessed.

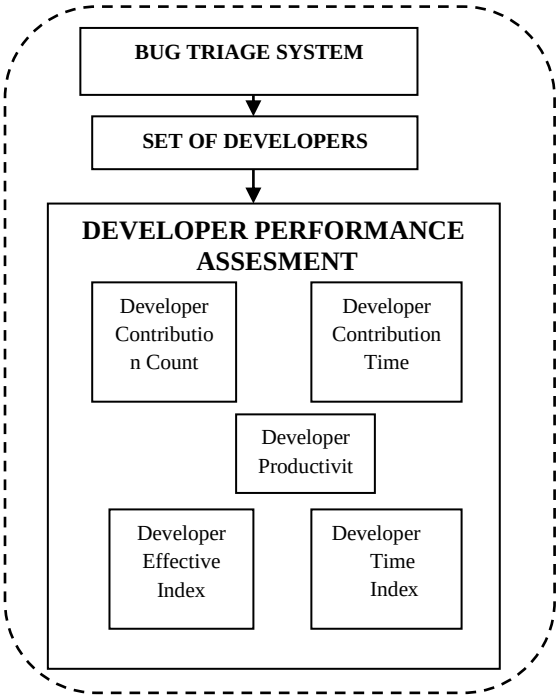


Fig. 6. A Holistic Evaluation Framework with Developer Performance

IV. PERFORMANCE EVALUATION OF BUG TRIAGE SYSTEM WITH KPIS

The performance of the existing system GP-WBFS [2], BT-ANT [19] and the Multiple Ant Colony System (MACS) [20] were analysed using Developer Productivity, Developer Effectiveness and Developer Time Index. The Goal-oriented Path model with Weighted Breadth First Search (GP-WBFS) algorithm was compared only with Bug Triaging based on Ant System (BT-ANT) and the MACS because only in these systems adaptive learning was adopted. The graph for Developer Time Index is given in the Figure .7. It is evident from the Figure .7 that the Developer Time Index for the MACS as well as the BT-ANT is skewed towards Developer Time Index of<300. Almost 85% of the retrieved developers by MACS have a Developer Time Index of <300 and 65% of the developers retrieved by BT-ANT has a Developer Time Index of <300. Whereas in the existing GP-WBFS, 77% of the developers have a Developer Time Index >300.

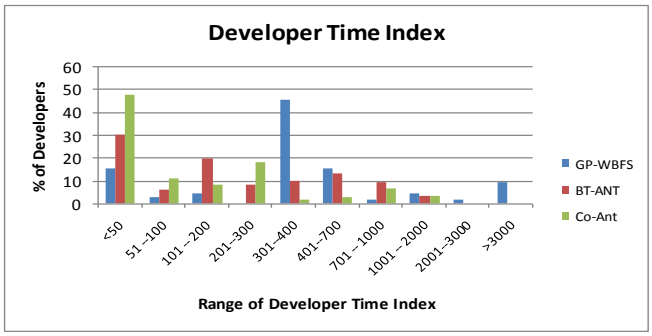


Fig. 7. Developer Time Index

The performance of the systems for Developer Effectiveness Index is given in Figure.8. Developer Effectiveness Index encodes the contribution of the developers for bugs that were resolved in a shorter period of time. From the graph, it is evident that 88% of the developers retrieved by MACS possess a Developer Effectiveness Index of >60 and 78% of the developers retrieved by the BT-ANT possess a Developer Effectiveness Index of >60. Whereas, in the developers retrieved by GP-WBFS, 78% of the developers have a Developer Effectiveness Index of <60.

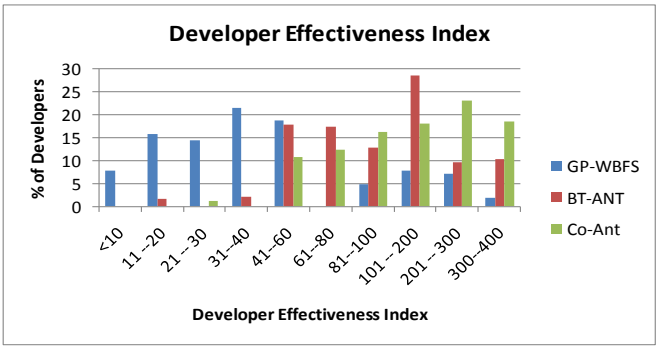


Fig. 8. Developer Effectiveness Index

The performance of the systems for Developer Productivity is given in Figure 9. Developer Productivity is a cumulative index that encodes the Developer Effectiveness, Developer Time Index and Developer Contribution Count. From Figure 9, it is evident that 91% of the developers retrieved by MACS possess a Developer Productivity of >50 and 75% of the developers retrieved by the BT-ANT possess a Developer Productivity of>50. Whereas in the developers retrieved by GP-WBFS, 73% of the developers have a Developer Productivity of <50.

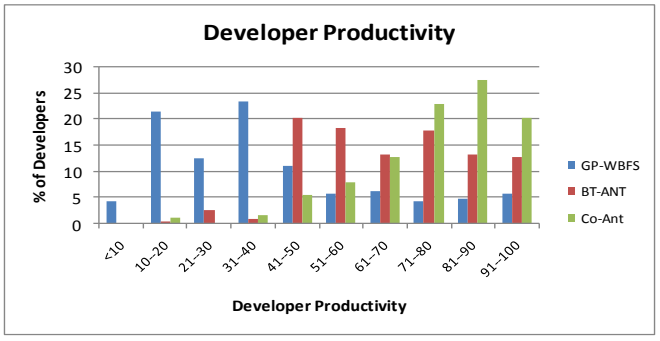


Fig. 9. Developer Productivity

V. Conclusion

This paper presents an Holistic Evaluation Framework for Bug Triage using developer performance. The metrics available in the literature that were used to evaluate the Bug Triage System were recommendation centric. The recommendation centric metrics evaluated the correctness and completeness of the recommendation mostly based on Precision and Recall measures. This paper adds a new dimension to the evaluation of the Bug Triage System. The evaluation framework factors in the quality of the developers extracted by the Bug Triage System in assessing the performance of the Bug Triage System. The evaluation metric based on the usefulness of the Bug Triage System is proposed. This is done by computing Key Performance Indicator values for the performance of the developers involved in the bug resolution. These calculated indices are then utilized to evaluate the developers extracted by the system. The proposed Key Performance Indicators are coarse grained in nature. A more fine grained analysis comprising the role analysis of the developers can be performed. The role analysis may be based on Social Network Analysis. Further , on the extracted roles of the developers more fine grained Key Performance Indicators are to be proposed.

Reference

[1] Pamela Bhattacharya and Iulian Neamtiu, "Fine-grained Incremental Learning and Multi-feature Tossing Graphs to Improve Bug Triaging," in IEEE International Conference on Software Maintenance, 2010, pp. 1-10.

[2] Pamela Bhattacharya, Iulian Neamtiu, and Christian R. Shelton, "Automated, Highly-accurate, Bug assignment using Machine Learning and Tossing Graphs," Journal of Systems and Software, vol. 85, no. 10, pp. 2275-2292, 2012.

[3] Gaeul Jeong, Sunghun Kim, and Thomas Zimmermann, "Improving Bug Triage with Bug Tossing Graphs," in 7th Joint meeting of the European Software Engineering Conference and the ACM SIGSOFT Symposium on The Foundations of Software Engineering (ESEC/FSE '09), New York, NY, USA, 2009, pp. 111-120.

[4] V.Akila, G.Zayaraz "Bug Triage in Open Source System- A Review", International Journal of Collaborative Enterprise, Inderscience Publishers, vol.4, no.4, pp.299–319,2014

[5] V.Akila, G.Zayaraz, "Novel Metrics for Bug Triage", Journal of Software, vol. 9, no.12, pp.3035-3040, Dec 2014

[6] Iman Avazpour, Teerat Pitakrat, Lars Grunske, and John Grundy, "Dimensions and Metrics for Evaluating Recommendation Systems," in Recommendation Systems in Software Engineering.: Springer Berlin Heidelberg, 2014, pp. 245-273.

[7] Jifeng Xuan, He Jiang, Zhilei Ren, and Weiqin Zou, "Developer Prioritization in Bug Repositories," in 34th International Conference on Software Engineering (ICSE), Zurich, 2012, pp. 25 - 35.

[8] Wen Zhang, Song Wang, Ye Yang, and Qing Wang, "Heterogeneous Network Analysis of Developer Contribution in Bug Repositories," in International Conference on Cloud and Service Computing (CSC), Beijing, 2013 , pp. 98 - 105.

[9] Song Wang, Wen Zhang, Ye Yang, and Qing Wang, "DevNet: Exploring Developer Collaboration in Heterogeneous Networks of Bug Repositories," in ACM / IEEE International Symposium on Empirical Software Engineering and Measurement, Baltimore, MD, 2013, pp. 193 - 202.

[10] Tao Zhang and Byungjeong Lee, "A Hybrid Bug Triage Algorithm for Developer Recommendation," in 28th Annual ACM Symposium on Applied Computing, Coimbra, Portugal, 2013, pp. 1088-1094.

[11] Geunseok Yang, Tao Zhang, and Byungjeong Lee, "Utilizing a Multi-Developer Network-based Developer Recommendation Algorithm to Fix Bugs Effectively," in 29th Annual ACM Symposium on Applied Computing (SAC '14)., Gyeongju, Republic of Korea, 2014, pp. 1134-1139.

[12] Tao Zhang and Byungjeong Lee, "An Automated Bug Triage Approach: A Concept Profile and Social Network Based Developer Recommendation," Intelligent Computing Technology,Lecture Notes in Computer Science, vol. 7389 , pp. 505-512, 2012.

[13] Shadi Banitaan and Mamdouh Alenezi, "DECOBA: Utilizing Developers Communities in Bug Assignment," in 12th International Conference on Machine Learning and Applications (ICMLA), Miami, 2013, pp. 66 - 71.

[14] Liguo Chen, Xiaobo Wang, and Chao Liu, "Improving Bug Assignment with Bug Tossing Graphs and Bug Similarities," in International

Conference on Biomedical Engineering and Computer Science (ICBECS), 2010 , Wuhan, 2010, pp. 1 - 5.

[15] Ayushi Rastogi, Arpit Gupta, and Ashish Sureka, "Samiksha: Mining Issue Tracking System for Contribution and Performance Assessment," in 6th India Software Engineering Conference, New Delhi,India, 2013, pp. 13-22.

[16] Tao Zhang, Geunseok Yang, Byungjeong Lee, and Ilhoon Shin, "Role Analysis-based Automatic Bug Triage Algorithm," Technical Report 2012.

[17] Raula Gaikovina Kula, Kyohei Fushida, Shinji Kawaguchi, and ajimu Iida, "Analysis of Bug Fixing Processes Using Program Slicing Metrics," in Product-Focused Software Process Improvement, Lecture Notes in Computer Science.: Springer Berlin Heidelberg, 2010, vol. 6156, pp. 32-46.

[18] Daniel Alencar da Costa, Uirá Kulesza, Eduardo Aranha, and Roberta Coelho, "Unveiling Developers Contributions Behind Code Commits: An Exploratory Study," in 29th Annual ACM Symposium on Applied Computing, Gyeongju, Republic of Korea, 2014, pp. 1152-1157.

[19] V.Akila, G.Zayaraz, V.Govindasamy, "Bug Triage based on Ant Systems", International Journal of Bi- Inspired Computation , vol. 7 no. 4, pp. 263-268 , August 2015

[20] Govindasamy V, Akila V, Banu Priya, "Bug Triaging Using Multi-Attribute Bug Tossing Graph" Discovery Journal, 46(213), pp.101-106, 2015,

An Analysis on Host Vulnerability Evaluation of Modern Operating Systems

Afifa Sajid
Department of Computer Science
COMSATS Institute of Information Technology,
Islamabad, Pakistan

Munam Ali Shah
Department of Computer Science,
COMSATS Institute of Information Technology,
Islamabad, Pakistan

Muhammad Kamran
Department of Distance Continuing & Computer Education
University of Sindh, Hyderabad,
Pakistan

Qaisar Javaid
Department of Computer Science & Software Engineering,
International Islamic University, Islamabad, Pakistan

Sijing Zhang
Department of Computer Science & Technology
University of Bedfordshire, Luton, UK

***Abstract*—Security is a major concern in all computing environments. One way to achieve security is to deploy a secure operating system (OS). A trusted OS can actually secure all the resources and can resist the vulnerabilities and attacks effectively. In this paper, our contribution is twofold. Firstly, we critically analyze the host vulnerabilities in modern desktop OSs. We group existing approaches and provide an easy and concise view of different security models adapted by most widely used OSs. The comparison of several OSs regarding structure, architecture, mode of working, and security models also form part of the paper. Secondly, we use the current usage statistics for Windows, Linux, and MAC OSs and predict their future. Our forecast will help the designers, developers and users of the different OSs to prepare for the upcoming years accordingly.**

Keywords*—Security; Operating system; Virtualization; kernel; Reliability; Vulnerability evaluation

I. Introduction

In today's world, information security is very important. It can relate to our national policy, civilization, economy, and military affairs, etc. Hence, for the information security system, operating system (OS) security is very important [1]. The OS is the core intermediary that resides on the hardware and controls the security of a computing environment. Any security vulnerability that exists in the OS causes the vulnerability at a host, therefore, the OS security is considered as the most critical part of a computing environment. Any OS that gives the reliability and supports and addresses the security issues efficiently to meet a certain set of requirements is called a secure or trusted operating system [1]. An OS manages the working and operation of all the complex applications on a computer system. It must also be capable of coping up with a largely increasing number of dangerous malicious attacks, software bugs, and hardware failures.

Through the Internet, all the web browsers accept download and run executable files. Desktop OSs and many other applications are compatible with the plug-in technology making the host environment more vulnerable to attacks. Furthermore, there is a growing trend that more users pay attention and respond to different executable codes that are present on the Internet. If these executable codes are directed from unauthorized sources or infected by viruses, then execution of such files can bring the system security at high risks [2]. Tragically, most of the desktop personal computer (PC) operating systems only give basic protection. Before running the executable files they do not authorize properly [3]; it is one of the many reasons for the accelerating spread of harmful software. Nowadays, everyone uses smartphones, desktop computers or laptops in their routine tasks. In this case, these systems are handling different environments like web browsing. At the same time, these systems are also handling data which is private or sensitive and is used for the confidential corporate matters. There exists the risk of malicious attacks on the Internet. These attacks can modify the confidential corporate data. Hence, the security of end user and/or security of the whole corporation is at risk. If this problem is neglected, it may cause severe security flaws. When a virus attacks any host or file, it propagates from one host to the other and replicates itself. To stop the propagation of virus from one host machine to the other host isolation is needed.

The existing research does not provide critical analysis and comparison of different desktop OSs and the tools that can make an OS more secure. This lag in existing literature motivated us to survey the security aspects of modern OSs. In this paper, a comparison of different OSs regarding their kernel security, network security, and system security is provided. The rest of the paper is organized as follow. Section II presents the security mechanism for different OSs. We also analyze different OSs for different parameters such as kernel-based containment and virtualization. The modules and techniques to secure an OS have been discussed in Section III. In Section IV, the security performance comparison of different OSs has been provided. The open issues are discussed in Section V where we also predict the future for different operating systems. The paper is concluded in Section VI.

II. Operating System Security Mechanism

Security is a challenging task in computing systems. Operating system is a component that manages all the other applications, so it needs to be secured [3]. For analyzing the

characteristics of an operating system, there is no general set of metrics available that can access the possible risks present in any operating system.

A. Security Methods

The methods used for operating system security comprises of software security and hardware security [4]. Hardware security method consists of I/O protection, running protection and storage protection while software security method relates mainly to the following features.

a) Identification and authentication: User identification and authentication is mandatory. Identification is that system needs the user's identity. And the process of connecting user identifier with the user is known as authentication.

b) Access Control: Computer system security's primary mechanism is access control. It consists of three steps. First is authorization second is access permission and the third is impose access permission.

c) Least Privilege: Give all users, only the kind of privileges that they need to complete the task.

d) Trusted Channel: Usually in computers an interaction between the Operating system and a user is through middle application layer which is not trustworthy. So the operating system needs to make sure that during communication Trojan horse cannot be able to capture the information.

e) Virus protection: In real world protecting your computer system from viruses is a very difficult task. In general, certain functions can be protected using security MAC (Mandatory Access Control) mechanism of operating system security.

B. Quantifying Risks

Unfortunately it is not easy to quantify the risks related to operating system [5]. This fact is reflected by the lack of research in this area. The problem lies in recognizing such data. If we can easily identify such data only then we can take the right actions accordingly. For both personal purposes and professional purposes Linux, Windows, Mac, and Solaris are considered the best operating system and billions of users use them. Before selecting any of them, there is a need to answer the question that which of them performs the required functions in the best way. In this section, we will analyze all these operating systems and certain security modules.

C. Windows Operating System

Windows are the most popular operating system. The desktop environment is dominated by Windows operating system. It is portable and extensible operating system, hence, it can take benefit of new hardware and new techniques. Windows allow multiprocessing (symmetric) and several operating environments [6]. It supports Non-uniform Memory Acce8uss (NUMA) computers, 32-bit and 64-bit processors. To offer basic services, Windows use the kernel objects with client server computing to give support to lots of application environments. It provided preemptive scheduling, virtual memory and integrated caching. The statistical data obtained from [7] has been used to plot the OS usage for different operating systems in 2015. It can clearly be observed that Windows is the most widely used OS.

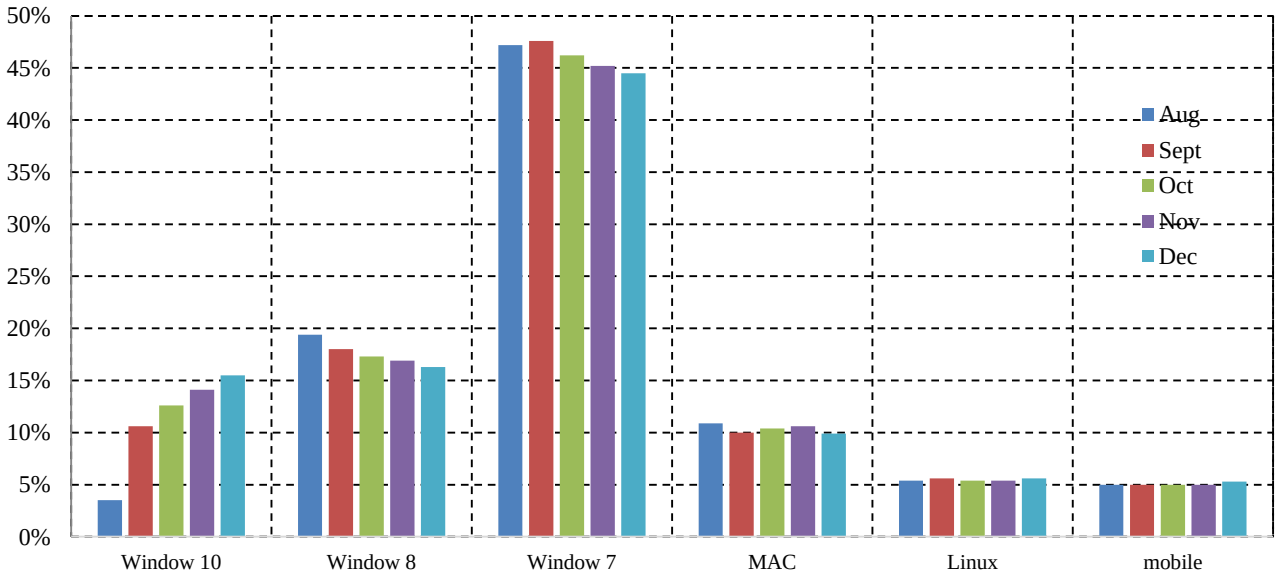


Fig. 1. Graph representing OS usage in 2015

By design, Window operating system is monolithic. In Windows operating system design, much focus is given to GUI (Graphical User Interface). In Windows operating system end points are unprotected and susceptible to malware. It is not considered efficient in managing obscure changes and software updates. In addition to updates, it gives poor end user experience. If we consider administration, actions performed by an administrator can be easily isolated from the user actions.

a) Windows OS Kernel: Kernel system for both Windows and Linux is very secure. Although the design decisions of Windows are different than Linux, but to exploit either of operating system by kernel is very rare. Windows kernel provides virtualization [8]. The virtualization process seperates one virtual machine from the other that's why it is considered good for security [9]. Certain security models and techniques are presented in literature for Windows operating system kernel to make it more secure.

- *Kyongi Information Security System:* This mechanism shows fast response against any attack. An effective response mechanism is built by analyzing kernel attack types [10]. It gathers information and analyzes the Kernel information and then use this information against windows kernel by using response mechanism. Kernel level programming skills are required to implement this system.
- *Symbolic Execution:* This system was proposed for testing paths to achieve high-security coverage [11]. The program is executed with symbolic inputs and do not use the concrete inputs. Whenever a branch instruction is executed it preserves the path condition. This path condition is updated with every execution of the program. Test generation is performed by solving the collected constraints with the help of constraint solver. It checks for the assertion violation or run-time errors, and it creates test inputs that trigger those errors. It checks all the supported functions. There still exists scalability issues because of the analysis of a number of large paths and constraints complexity that is produced.

b) Windows OS System Security: Security subsystem in Windows is made up of certain components that cannot grant access to a user without proper authentication and identification. Only security subsystem function offers the access control [12]. This can be implemented by giving different privileges and rights to the user. Another similar feature known as capabilities, it is present in Linux system [6]. There is a need to find out the possible risks associated with privileges, rights and capabilities. By doing this we can find the risks and bugs more efficiently.

c) Windows OS Network Security: The network security is achieved by data monitoring using socket. In this technique data is monitored using socket [13]. At any stage of the process when it realizes the need, it will monitor the net data, then apply certain development on net base data to allow secure transmission.

d) Linux Operating System

Linux is built on UNIX principles and is free, open source operating system. The user interface and programming interface is compatible with the regular UNIX system. A kernel of Linux operating system is fully original, but it allows many UNIX-based applications to run. Because of the performance reasons its kernel is implemented as a monolithic kernel, but at run time, drivers can be loaded and unloaded dramatically [14]. Linux design is modular enough, and Linux servers are considered best for non-local administration. It supports a multi-user environment. Using time-sharing scheduler it gives protection among several running processes. It allows multithread programming. In Linux to reduce the repetition of data shared by many processes, memory management uses copy-on-writing and page sharing. Because of the open source theory of Linux it allows more peer evaluation of the code to find bugs and to fix the code [15]. But it does not mean that Linux is secure. In past Windows had been the most attacked operating system. Because of this reason most Windows-based servers were shifted to Linux. Ultimately, the attackers started targeting Linux accordingly.

e) Linux Kernel Security: Kernel is the most sophisticated and complex piece of software on the system. It is the core of the system [16]. If the kernel or any of its parts is affected by virus or any malicious code then the system will be corrupted and all secret information can be stolen, and files can be deleted. The Linux kernel provides virtualization.

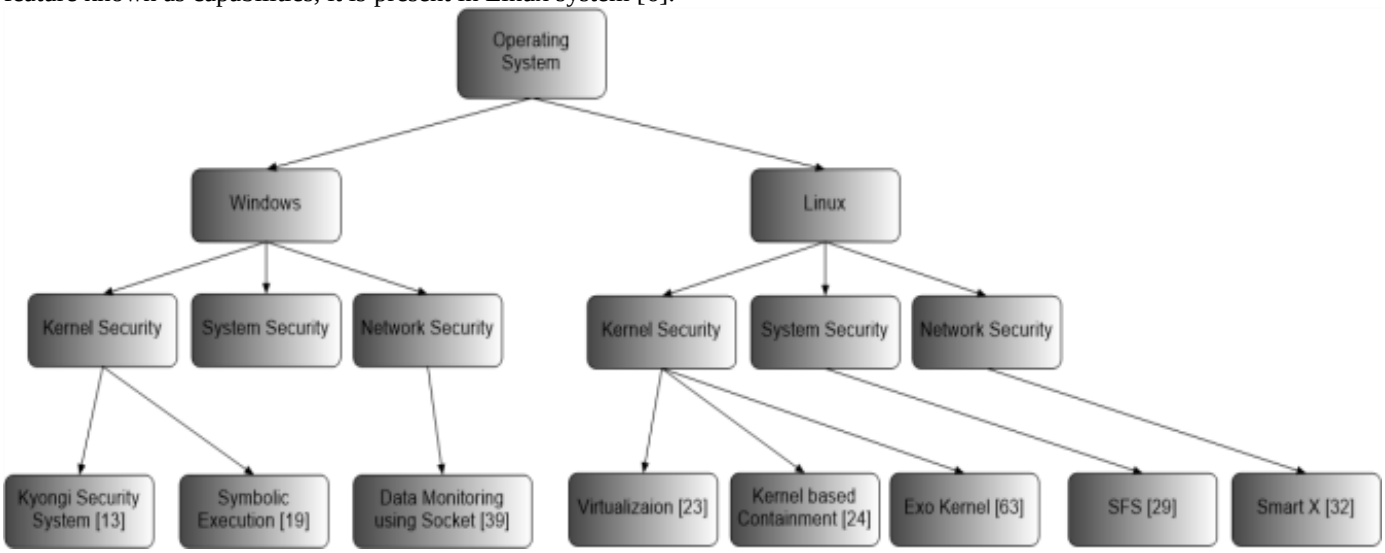


Fig. 2. Taxonomy Diagram of Linux, Windows, and their security modules

- *Virtualization:* For many server-oriented systems an approach has been taken to diminish the propagation of viruses or any other malicious attacks by separating the various services on different machines [17]. By using the technique of virtualization, we can achieve the isolation by running several virtual machines (VMs) on one physical machine, an application gets corrupted and run under a given VMs cannot be able to interrelate with different other applications in separate VMs.
- *Kernel Based Containment:* The isolation of different activities from one another. can be done by using the access rights like HRU [18]. Each user then can access only the subset of files, but in the case of corruption of any application, the attacker can get a hand on every file of the user and can also change the permission of these files [18]. Mandatory Access Control (MAC) systems such as TOMOYO Linux [19] and SELinux [20] allows limiting the damage caused by the compromised application. Due to this security system the administration can set access control policy for every user so the attacker gets fewer privileges that could be gained by the application diversion phenomenon [16]. The disadvantage in MAC solution is that it has very high cost for the administrative purposes because the policies used are very complex and large.
- *Virtualization Based Containment:* There is a virtualization technology called Xen through which different domains can be placed in various virtual machines [21]. As compared to modern kernel code, virtualization code is simpler [10], thus, it needs more security. The project proposes simplified mandatory access control kernel.
- *Exokernel:* It is an extension of the microkernel theory [22]. Abstraction of the physical devices is not provided by the base kernel, putting down that to the applications that required devices. It separates the security from abstraction, hence making parts of the operating system which are non-override-able to do nothing but to multiplex the hardware. The objective is to avoid forcing any abstraction on applications and let the abstraction use whatever abstraction best suited for their task. Exokernel technology is still much not researched comprehensively.

f) File Data Security in Linux: In Linux and Unix files are thought to be one of the important unit for storing information. The idea of a file is of much significance that all the devices that are input/output are believed to be files [23]. So in Linux problem of protecting data is considered to be protecting file data. In current computer environments file data security is mandatory. Utilities like "aescrypt" and "crypt" are used for securing file data. In these utilities as input password and file name is given to produce a file which will then encrypted [12]. Linux file system is a hierarchical tree that follows the Unix. To manage different file systems, Linux makes use of an abstraction layer [24]. Virtual file system, device oriented and networked these all are supported. Different modules are used to enhance the file security system some of these are

- *Secure File System (SFS):* To secure the file data, SFS uses cryptographic techniques. Encryption of SFS services is done in kernel space of Linux [23]. After integration, it provides security services as it is inherited [20]. For storing file data, SFS ask the user to make a directory named "ecrypt" like ecryptdir. Without user intervention, all of the files placed in this directory will be encrypted transparently.

g) Linux Network Security In Linux, certain access permissions are defined for network security. First is password security, to access the Linux system users have to go through an authentication process in this process all users needs to input the password and account then system validates the user and grant access. In Linux, each user is given different level of permissions to access any data. Some important users are given more access than the others who are given less access to the data [25]. To increase the security of the system Linux removes unnecessary services. In a network Linux limits the number of IP (Internet Protocol) addresses. A super user password is used which is generally known to the system administrator and needs to be revised periodically [26]. Also, user of Linux are given a lower level account, not the root account so if the system is compromised only the certain programs, and local files will be affected by that vulnerability.

- *Smart-X:* This software framework can be used in Windows and Linux for efficient and secure communication connecting two nodes. The working of this module involves creating a tunnel between two end points using a single context switching and a single copy of data. An algorithm is used to secure the two end points. For encryption/decryption purposes advanced standard encryption 128-bit algorithm is used [27]. This software framework remains on Network Driver Interface Specification (NDIS). Through this framework, some modifications are performed on a packet before transmission. The framework encrypts the whole packet from the start of the header to the end of the header and then generates one or more UDP packets from it. If the encrypted packet size exceeds then, it is divided into two UDP packets. These UDP packets hit the wire and send out to the destination, where they are reassembled if necessary and then decrypted.

III. Modules and Techniques for OS Security

There are some more modules and techniques available which can be used to enhance the operating system security. Some of these are:

a) Chinese Wall Security Policy Model: Based on the realistic commercial business model this security policy was devised by Nash and Brewer [28]. In this security policy, the company information is divided into three storage levels. Single data elements make the base level; company data elements form the next level and the level on top of all this consists of the information with regards to conflict of

interest(COI) [29]. One company can belong to only one COI. A user may visit any CD (Control Domain) within a COI without any limiting or enforcement factor [28]. But after the user makes any decision, then user cannot get access to other elements of the same COI, it is like a wall has been created around the CD.

TABLE I. MODULES AND TECHNIQUES FOR MAKING OS SECURE

Ref.	*Name*	*Operating System*	*Security level*	*Features*
[21]	Symbolic Execution	Windows	Kernel Security	It checks all the supported functions to provide high-security coverage.
[20]	Kyongi Information Security System	Windows	Kernel Security	An effective response mechanism is built by analyzing kernel attack types.
[26]	Smart-X	Windows/ Linux/ UNIX	Network Security	Working of the module involves creating a tunnel between two endpoints using single context switching and single copy of the data.
[10]	Socket Data Monitoring	Windows	Network Security	Data is monitored using a socket and this technology is known as cutting age data monitoring.
[28]	Secure file system	Linux/ UNIX	System security/file data security	To secure the file data SFS uses cryptographic techniques.
[30]	Signature based code scanning	Windows/ Linux/ UNIX	N/A	scanning is performed to find any malicious code in it if identification of any malicious code is confirmed then execution of program file is denied
[13]	Chinese wall security policy	Windows /Linux/ UNIX	N/A	Information is divided into different levels for making it secure.

b) Signature Based Code Scanning: A mostly used and well-known technique that is signature based malicious code scanning is used for the purpose of authorization and inspection of executable codes [30]. In operating system, any program file scanning is performed to find any malicious code in it. If the identification of any malicious code is confirmed then execution of program file is denied. The problem with this code scanning is, it can only defend against the attacks which are known. Solution to this is an approach known as combined integrity measurement and access control. Only authentic programs can run in this model [17]. If there is an existing code which is from an authentic source and not considered malicious. It shows malevolent manners and at run time can be trapped and manipulated.

c) Security Enhanced Linux: There was a project under NSA (National Security Agency) called security enhanced Linux. This project is to implement MAC on Linux for military purposes. For the architecture La Padula model [31] is followed and supports RBAC (Rule Based Access Control). In file system EA (Extended Attribute) is used to label the files. The security context is also supported for any known application. During execution of an application, security context can be switched by the application. It is a Linux modification and was initially released by NSA in January 2000 [32]. It extends the features of Linux with certain security capabilities to make it more secure. SE (Security Enhanced) Linux gives a language for all the security policies of Linux. Deletion: Delete the author and affiliation lines for the second affiliation. This specified security policy covers all the aspects of the system like file management, network communication and process control [29]. Policy enforcement uses the method in Flask Architecture. Here a server gives the decision, for policies whether user request should be granted to the operating system or not. For making a decision this security server refers to the other internal policies.

IV. OTHER OPERATING SYSTEM

In this section, we provide the performance comparison of some other OSs. Our focus is on security of Solaris and Macintosh OS.

A. Solaris Operating System

Solaris is a Unix-based operating system [33]. In Solaris multiple software isolated applications can run on a single system, that's how linking between servers become easy. Some abilities of Solaris 10 like Process Rights Management, Predictive Self-Healing and Dynamic Tracing helps in attaining good utilization without causing any harm to privacy or security levels [33]. Solaris operating system offers adequate CPU and memory to applications and also preserve the ability to utilize the idle resources [34]. It has an ability to recover automatically from disastrous problems that occur in the system by using both Solaris containers and self-healing functionalities.

Solaris 10 offers advanced security features [32]. Solaris containers work with process rights and user management to give secure hosting of hundreds of applications and several customers on the system. To apply secure foundation, security administrators can harden and minimize Solaris. It offers the following functionalities to the users.

TABLE II. COMPARATIVE ANALYSIS OF DIFFERENT OSS

Features	*Linux/Unix*	*Windows*	*Solaris*	*MAC*
What is it?	Open source development and free operating system [51].	Operating system from Microsoft, not free [52].	Unix-based operating system introduced in 1992 [32].	Series of graphical user interface based operating system [40].
Manufacturers	Developed by a society Linus Torvalds manage things [20].	The developer of Windows is Microsoft.	Developed by Sun Microsystems. Now be in possession of Oracle Corporation [32].	Developed by Apple Inc in 1984 [40].
Cost	Can be downloaded free of charge. Some priced editions are also offered [53].	For desktop users, it can be costly depending on the version. From $50 to $450 [13].	Costly ranging from $500 to $720 [54].	Mac OS is too expensive [54].
Users	Anyone from home users to developers.	Anyone from home users to developers.	Not for home users, for developers.	Anyone from home users to developers.
GUI	Two default GUI, Gnome and KDE others are like twm, Unity, Mate, LXDE, Xfce are also offered [19].	GUI is an important part of operating system and is not replaceable [55].	Provides Gnome and CDE [34].	N/A
File system support	NTFS, FAT, FAT32, Xfs, Btrfs, Ext2, Ext3, Ext4, Jfs, ReiserFS [15]	NTFS, FAT, exFAT, FAT32 [56].	ZFS, UFS, HSFS, NFS, TFS, PCFS [57]	HFS or HFS+ , Macintosh file system [45].
Text Mode Interface	BASH (Bourne against SHell) is default shell, can maintain several command interpreters [58].	A command shell and each version of Windows uses single command interpreter with a command like DOS have, there is an addition of non-compulsory PowerShell [59].	Solaris console and kernel terminal emulator [46].	N/A
Security	Till date 100-200 viruses programmed not keenly spreading [60].	There exist lots and lots of viruses; antivirus costs $20 to $450 [38].	Fully virus protected. Assured and tested low-risk platform [60].	Much more secure than Windows, viruses designed for Windows processors won't run on Mac [61].
Threat detection and solution	In Linux case, threat detection and solution is very quick, as Linux is mostly society driven and at any time when any Linux user posts any type of threat, some developers start working on it from different parts of the world [16].	After detecting threat, Microsoft releases a patch which fixes the problem and takes more than 2/3 month sometimes sooner. Updates and patch are weekly based [13].	N/A	Apple provides software updates, works with the incident response community such as CERT, FIRST, and FreeBSD Security Team, to proactively identify and quickly correct operating system vulnerabilities [45].
Processors	Dozens of different kinds.	Limited.	Limited.	Different kind.
Gaming	Very little number of games are offered natively. It can be used to play some, but frequently not all features are offered [62].	Almost all games are compatible with Windows. Some CPU intensive and graphics intensive games are exclusive to Windows PC's [63].	It supports few games.	Games are available for the Mac, gives environment much like Windows but not better than it [64].
Version	Fedora, Red Hat, Debian, Android, Arch Linux, Ubuntu, etc.	Vista, XP, Window 7, 8, 8.1 etc.	Solaris 4.1x, 5.1, 5.2-5.11	Mac OS x 10.1, 10.8.1, 10.8 .3 etc.
User experience	Although there are many GUI applications, most of the work is done through Terminal (a console window), and if a problem occurs GUI is not often usable to fix them [65].	Everything can be controlled through GUI and incompatibility problems are exceptional [66].	Not user-friendly not recommended for home users.	Average not more customizable and user-friendly.
Graphics Performance	Because hardware manufacturers, such as NVidia, regularly does not provide documentation for Linux developers, drivers can' not use full card performance [38].	Integrated with newest DirectX versions and full graphics card support the performance is approximately as good as it can get [31].	Graphics performance is poor, lack of good GUI [40].	N/A
Supported Platforms	All [38]	PowerPC: versions 1.0 - NT 4.0, DEC Alpha: versions 1.0 - NT 4.0, MIPS R4000: versions 1.0 - NT 4.0, IA-32: versions 1.0 – 8, IA-64: version XP, x86-64: versions XP – 8, ARM: version RT [38]	SPARC, IA-32, PowerPC and i86PC (which includes both x86 and x86-64) platforms [67].	Compatible with only power pc processors version 10 and version 10.3, 68k processors.
Preceded by	Basic terminal (CLI) [38]	MS-DOS[31].	N/A	N/A

Features	*Linux/Unix*	*Windows*	*Solaris*	*MAC*
Terminal	Multiterminal windows	With Solaris/X86 2.4 not configured with Solaris 2.4-7 configured [36].	N/A	N/A

a) With the use of file verification and Solaris secure execution, it validates the system integrity.

b) Grant access only to the privileges needed processes and users, hence reduce the risk level.

c) For File encryption Solaris uses open standard-based cryptographic framework and hence makes the administration easier.

d) For network traffic security Solaris uses IP-Filter firewall.

Solaris is not for the home users; it is considered good for developers. Solaris lacks GUI (Graphical User Interface).

B. Macintosh Operating System

For home users Apple Macintosh was a debatable first GUI operating system. At first, it was designed for Apple computers, then slowly it was by Microsoft Windows. Mac (Macintosh) operating system has several benefits, apart from all these benefits two main serviceable benefits are protected memory and pre-emptive multitasking [35]. The scheduler is a layer which is a point of an interaction between microprocessor and application. This concept of the scheduler was introduced by Mac. When user opens an application the scheduler takes its control and allocates certain amount of CPU (Central Processing Unit) memory to the process. When the allocated time of the process is up, then scheduler takes control back and gives the CPU memory to another application [28]. Memory activity needs to be securely controlled, so for this when an application is executed, checks are given to confirm that memory activity of application is within bound [36]. If an application tries to interfere with critical resources or tries to write in another application space, it would be infeasible. This concept is called protected memory. Because of these two features, any rogue process or application can't make the whole system hostage [37]. Scheduler supports dynamic feedback and application with time constraints. Table 1 given below summarizes different security modules/techniques and their features and the Table 2 represents the comparative analysis of different operating systems.

V. Discussion and Open Issues

To support developers and the end users in their daily tasks, the modern OSs continue to evolve. Users do not need to have underlying architectural details of the resources. With the growing scope of resource type and growing demand for functionality, the OS has thus grown to become a flexible and large platform for which change by now is not convenient. For example, in Windows 7, to remove a dispatcher lock, one needs to write 6000 lines of code. For such complex system, execution cost is compensated by stronger and faster processor generation. With the advent of multi-core systems, the above assumption no longer holds. To provide powerful services, even though the underlying hardware will not increase exponentially anymore, the system needs to exploit the infrastructure and its specifications.

It will take effort and time into adapting OSs and programs for particular high-performance computing usage. Desktop OSs must be able to cope with any infrastructure and must demonstrate high portability. The actual objective which is behind the Microsoft Windows OS development is to hide the peculiarities of the underlying infrastructure which is complex and instead offer ease of portability of applications. The next generation operating systems must value these characteristics for usability and growth. It is obvious that Microsoft Windows has been the most popular and widely used OS until 2015. We reuse the statistics provided and Figure 1 and derive a relation to predicting the future growth of Windows OS up to the year 2020. We estimate that the Windows OS will continue to grow in its adaptation and usage by the end users. However, taking into account the worst scenario, the Microsoft Windows OS usage may decrease. This means that ideally, there will be 70% of the entire users who will be using Windows OS by the year 2020. Even, if the end users stop using the Windows OS, the total usage will still be 50% or more. The growth trends for Microsoft Windows OS are provided in Figure 3. We also predict that the Microsoft will be naming its OS as *Windows'20* in the year 2020. This would be because of the trends and the rapid growth in the concept of *Internet of Things* (IoT).

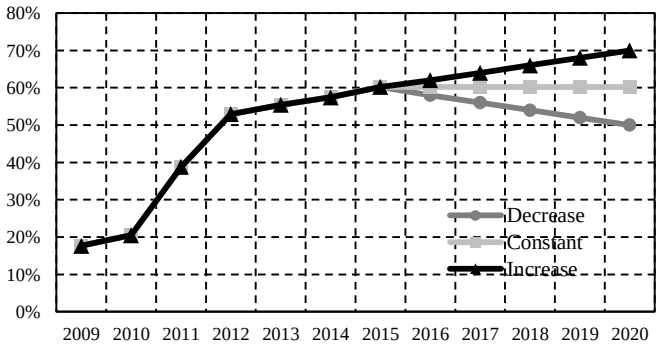


Fig. 3. Statistics representing the growth trends of Windows operating system in the future

Based on the current statistics for Linux OS, we predict that the growth in the usage of Linux OS will rise to approximately more than 7%. The future prediction of Linux usage is provided in Figure 4. One obvious reason for this inference is that Windows security model is not resilient to threats and vulnerabilities, and the end user would require some strong, secure OS, which can provide sufficient security to the confidential data. However, it is also possible that the current Linux based users migrate to Windows OS and the Linux usage may drop to 4% or lesser.

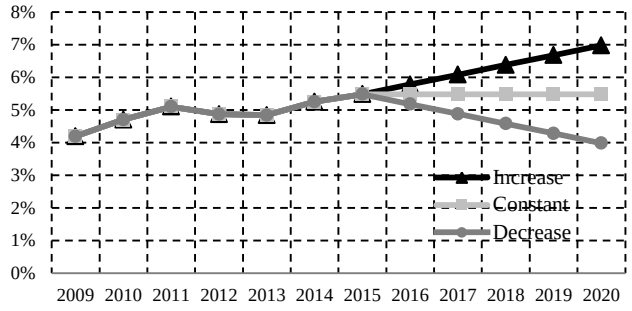


Fig. 4. Statistics representing the growth trends of Linux operating system in the future

We then predict the future usage of the Macintosh OS in the year 2020. We strongly believe that the MAC OS will continue to evolve, and the end user will adapt MAC OS more frequently in near future. However, it should be noted that comparatively, the MAC based PCs are still costly. The vendor must focus on reducing the cost of the hardware to make MAC based PCs affordable to the general public. The growth trend of MAC OS is given in Figure 5.

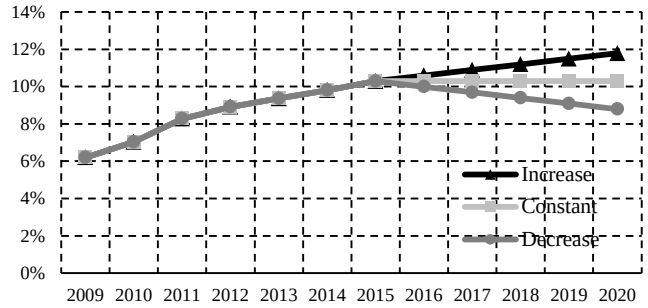


Fig. 5. Statistics representing the growth trends of Macintosh operating system in the future

VI. Conclusion

Achieving security is a challenging task. The operating system is the one which coordinates amongst all the other systems. This is the reason; the security, is of primary and major concern. In this survey paper, we analyzed several different features such as network security, system security and kernel security etc. of multiple OSs including Windows, Linux, and MAC. Some of the salient features of different OSs have been summarized in Table 3. We believe that the choice of an OS solely depends on the user requirement. A home user with suitable performance in multi-tasking selects Windows OS. When it comes to handling a large amount of data and to secure the resources, Solaris is a better choice. For multiuser environment, Linux is best. If the graphics and GUIs are concerned, MAC OS is a leading. There are certain modules and techniques available which can be adapted to make an OS more secure. Furthermore, it is believed that absolute security is impossible to achieve, and no OS is 100% secure against all types if threats and vulnerabilities. However, the designers and developers of the OS can strive for maximizing the security in all possible aspects and must satisfy the end users' security needs.

TABLE III. Advantages And Disadvantages Of Different Operating Systems

Operating System	Advantages	Disadvantages
Linux/ UNIX	• Provides multiuser environment. • Free of cost. • Modular design.	• Some Window programs may not run. • Learning curve for new people. • Not user friendly like Windows.
Windows	• Good GUI. • Universal plug and play feature available. • User friendly as compared to other OS.	• Easily compromised by virus, loopholes are there. • Heavy system old hardware is not able to run it. • Single user license.
Solaris	• No reboot needed, can run 24/7. • Fully virus protected. • Good backup tools.	• Not user friendly. • Not good GUI. • Not for home users.
MAC	• Much more secure than Windows. • Can get bootcamp. • Not popular as Windows, so not a target.	• Too expensive. • Some programs that run on Window won't run on Mac. • Few games available.

References

[1] M. Bishop, Computer security: art and science, vol. 200. Addison-Wesley, 2012.

[2] N. L. darek and O. Jae, "A system dynamics model for information security management," Information & mnagement , vol. 52, no.1, pp. 123-134, 2015.

[3] M. Maekawa, K. Shimizu, X. Jia, P. Sinha, K. S. Park, H. Ashihara, and N. Utsunomiya, Operating System. Springer Science & Business Media, 2012.

[4] J. Song, G. Hu, and Q. Xu, "Operating System Security and Host Vulnerability evaluation," in Management and Service Science, 2009. MASS'09. International Conference on, 2009, pp. 1–4.

[5] J. K. Guo, S. Johnson, and I.-P. Park, "An operating system security method for integrity and privacy protection in consumer electronics," in Consumer Communications and Networking Conference, 2006. CCNC 2006. 3rd IEEE, 2006, vol. 1, pp. 610–614.

[6] A. Silberschatz, P. B. Galvin, G. Gagne, and A. Silberschatz, Operating system concepts, vol. 9. Addison-Wesley Reading, 2008.

[7] w3school, "OS Statistics," http://www.w3schools.com/browsers/browsers_os.asp. .

[8] D. H. Lee, J. M. Kim, K.-H. Choi, and K. J. Kim, "The study of response model & mechanism against windows kernel compromises," in Convergence and Hybrid Information Technology, 2008. ICHIT'08. International Conference on, 2008, pp. 600–608.

[9] T. Ni, Z. Yin, Q. Wei, and Q. Wang, "High-Coverage Security Testing for Windows Kernel Drivers," in Multimedia Information Networking and Security (MINES), 2012 Fourth International Conference on, 2012, pp. 905–908.

[10] J. Park, J. Park, J. Lee, B. Kim, G. Lee, and B. Cho, "Windows Security Patch Auto-Management System Based on XML," in Advanced Communication Technology, The 9th International Conference on, 2007, vol. 1, pp. 407–411.

[11] C. Cadar, P. Godefroid, S. Khurshid, C. S. P\uas\uareanu, K. Sen, N. Tillmann, and W. Visser, "Symbolic execution for software testing in practice: preliminary assessment," in Proceedings of the 33rd International Conference on Software Engineering, 2011, pp. 1066–1071.

[12] E. B. Fernandez and T. Sorgente, "A pattern language for secure operating system architectures," in Proceedings of the 5th Latin American Conference on Pattern Languages of Programs, 2005, pp. 16–19.

[13] M. Howard and S. Lipner, "Inside the windows security push," IEEE Secur. Priv., vol. 1, no. 1, pp. 57–61, 2003.

[14] E. William, H. Nadkarni, A. Sadeghi, A. Reza and H. Stephan, "Programmable interface for extending security of application-based Operating System," US Patent 20,160,042,191. 2016.

[15] R. K. Pal and I. Sengupta, "Enhancing file data security in linux operating system by integrating secure file system," in Computational Intelligence in Cyber Security, 2009. CICS'09. IEEE Symposium on, 2009, pp. 45–52.

[16] G. Zhai and Y. Li, "Analysis and Study of Security Mechanisms inside Linux Kernel," in Security Technology, 2008. SECTECH'08. International Conference on, 2008, pp. 58–61.

[17] C. Border, "The development and deployment of a multi-user, remote access virtualization system for networking, security, and system administration classes," in ACM SIGCSE Bulletin, 2007, vol. 39, no. 1, pp. 576–580.

[18] T. Harada, T. Horie, and K. Tanaka, "Task oriented management obviates your onus on Linux," in Linux Conference, 2004, vol. 3.

[19] N. S. A. Peter Loscocco, "Security Vulnerabilities: An Impediment Against Further Development of Smart Grid," in Smart Grids from a Global Perspective, pp 77-93, 2016.

[20] N. Petreley, "Security report: Windows vs Linux. The Register, October 2004."

[21] A. Lackorzynski and A. Warg, "Less is More--A Secure Microkernel-Based Operating System," in SysSec Workshop (SysSec), 2011 First, 2011, pp. 103–106.

[22] R. K. Pal and I. Sengupta, "Enhancing File Data Security in Linux Operating System by Integrating Secure File System," in Computational Intelligence in Cyber Security, 2009. CICS '09. IEEE Symposium on, 2009, pp. 45–52.

[23] D. D. Clark and D. R. Wilson, "A comparison of commercial and military computer security policies," in Security and Privacy, 1987 IEEE Symposium on, 1987, p. 184.

[24] N. Provos, "Improving Host Security with System Call Policies.," in Usenix Security, 2003, vol. 3, p. 19.

[25] S. Schaefer, "Operating system abstraction and protection layer," in Proceedings of the 16thSymposium on Operating Systems Principles (SOPS), 'Online!, 2006, pp. 1–14.

[26] R. Luniya, A. Agarwal, M. Bhatnagar, V. Rathod, and D. Unwalla, "SmartX--Advanced Network Security for Windows Opearating System," in Intelligent Systems, Modelling and Simulation (ISMS), 2012 Third International Conference on, 2012, pp. 680–683.

[27] D. F. C. Brewer and M. J. Nash, "The chinese wall security policy," in Security and Privacy, 1989. Proceedings., 1989 IEEE Symposium on, 1989, pp. 206–214.

[28] B. Armstrong, P. England, S. A. Field, J. Garms, M. Kramer, and K. D. Ray, "Computer security management, such as in a virtual machine or hardened operating system," in Proceedings of the Computer Society Symposium on Research in Security and Privacy, IEEE Pub., 2008, pp. 2–19.

[29] G. Ateniese, R. Burns, R. Curtmola, J. Herring, O. Khan, L. Kissner, Z. Peterson, and D. Song, "Remote data checking using provable data possession," ACM Trans. Inf. Syst. Secur., vol. 14, no. 1, p. 12, 2011.

[30] D. E. Bell, "Looking back at the Bell-La Padula model," in null, 2005, pp. 337–351.

[31] B. McCarty, Selinux: Nsa's open source security enhanced linux. O'Reilly Media, Inc., 2004.

[32] M. Richard, Solaris" Performance And Tools: Dtrace And Mdb Techniques For Solaris 10 And Opensolaris. Pearson Education India, 2007.

[33] K. K. Yue and D. J. Lilja, "Dynamic processor allocation with the Solaris operating system," in Parallel Processing Symposium, 1998. IPPS/SPDP 1998. Proceedings of the First Merged International... and Symposium on Parallel and Distributed Processing 1998, 1998, pp. 392–397.

[34] D. Price and A. Tucker, "Solaris Zones: Operating System Support for Consolidating Commercial Workloads.," in LISA, 2004, vol. 4, pp. 241–254.

[35] C.-Y. Yang, Y.-K. Huang, N.-C. Perng, J.-J. Chen, Y.-H. Lee, C.-M. Hung, H.-R. Hsu, S.-W. Huang, H.-W. Tseng, A.-C. Pang, and others, "Another real-time operating system and unified MAC protocol for home controlling and monitoring," in Software Technologies for Future Embedded and Ubiquitous Systems., 2006,

[36] E. H. B. M. Gronenschild, P. Habets, H. I. L. Jacobs, R. Mengelers, N. Rozendaal, J. Van Os, and M. Marcelis, "The effects of FreeSurfer version, workstation type, and Macintosh operating system version on anatomical volume and cortical thickness measurements," PLoS One, vol. 7, no. 6, p. e38234, 2012.

[37] C. Yi, Z. Zhi-rong, and S. Chang-xiang, "Design and implementation MAC in security operating system," in TENCON'02. Proceedings. 2002 IEEE Region 10 Conference on Computers, Communications, Control and Power Engineering, 2002, vol. 1, pp. 216–219.

[38] "Linux vs Windows - Difference and Comparison | Diffen." [Online]. Available: http://www.diffen.com/difference/Linux_vs_Windows. [Accessed: 08-Dec-2014].

[39] "Solaris Operating System." [Online]. Available: http://www.slideshare.net/wangluxh09/solaris-operating-system-36793409?related=1. [Accessed: 08-Dec-2014].

[40] "Mac, Windows And Solaris." [Online]. Available: http://www.slideshare.net/shivendra007/mac-windows-and-solaris. [Accessed: 08-Dec-2014].

[41] "Oracle Solaris 10 | Operating System | Oracle." [Online]. Available: http://www.oracle.com/us/products/servers-storage/solaris/solaris10/overview/index.html. [Accessed: 08-Dec-2014].

[42] Global Pricing and Licensing." [Online]. Available: http://www.oracle.com/us/corporate/pricing/index.html. [Accessed: 08-Dec-2014].

[43] "Solaris 11 GNOME 2.30." [Online]. Available: http://toastytech.com/guis/sol11.html. [Accessed: 08-Dec-2014].

[44] "Overview of File Systems - Oracle Solaris Administration: Devices and File Systems." [Online]. Available: https://docs.oracle.com/cd/E23824_01/html/821-1459/fsoverview-51.html. [Accessed: 08-Dec-2014].

[45] OS X: Mac OS Extended format (HFS Plus) volume and file limits. Support.apple.com, 2008.

[46] "Solaris Consoles and the Kernel Terminal Emulator." [Online]. Available: https://docs.oracle.com/cd/E19253-01/816-4854/6mb1o3bg7/index.html. [Accessed: 08-Dec-2014].

[47] "UNIX System Administration: Solaris, AIX, HP-UX, Tru64, BSD.: Gaming on Solaris." [Online]. Available: http://blog.boreas.ro/2007/08/gaming-on-solaris.html. [Accessed: 08-Dec-2014].

[48] Lineage online RPG now available as Mac OS X beta. .

[49] Oracle Completes Acquisition of Sun. Yahoo, 2010.

[50] "Mac OS X versions (builds) for computers - Apple Support." [Online]. Available: http://support.apple.com/en-us/ht1159. [Accessed: 08-Dec-2014].

[51] C. Zemao, W. Xiaoping, and T. Weimin, "An Executable Code Authorization Model for Secure Operating System," in Electronic Commerce and Security, 2008 International Symposium on, 2008, pp. 292–295.

[52] J. Viega and J. Voas, "The pros and cons of Unix and Windows security policies," IT Prof., vol. 2, no. 5, pp. 40–47, 2000.

[53] P. Levis, S. Madden, J. Polastre, R. Szewczyk, K. Whitehouse, A. Woo, D. Gay, J. Hill, M. Welsh, E. Brewer, and others, "TinyOS: An operating system for sensor networks," in Ambient intelligence, Springer, 2005, pp. 115–148.

[54] O. Services, "Oracle Solaris 10," Available: http://www.oracle.com/us/products/servers-storage/solaris/solaris10/overview/index.html. [Accessed: 08-Dec-2014]. 2014.

[55] X. Song, M. Stinson, R. Lee, and P. Albee, "An approach to analyzing the Windows and Linux security models," in Computer and Information Science, 2006 and 2006 1st IEEE/ACIS International Workshop on Component-Based Software Engineering, Software Architecture and Reuse. ICIS-COMSAR 2006. 5th IEEE/ACIS International Conference on, 2006, pp. 56–62.

[56] M. E. Russinovich, D. A. Solomon, and J. Allchin, Microsoft Windows Internals: Microsoft Windows Server 2003, Windows XP, and Windows 2000, vol. 4. Microsoft Press Redmond, 2005.

[57] E. Zadok, R. Iyer, N. Joukov, G. Sivathanu, and C. P. Wright, "On incremental file system development," ACM Trans. Storage, vol. 2, no. 2, pp. 161–196, 2006.

[58] J. Dike, User mode linux, vol. 2. Prentice Hall Englewood Cliffs, 2006.

[59] J. Fox, "Getting started with the R commander: a basic-statistics graphical user interface to R," J. Stat. Softw., vol. 14, no. 9, pp. 1–42, 2005.

[60] B. Leiba, "Aspects of Internet security," IEEE Internet Comput., no. 4, pp. 72–75, 2012.

[61] R. Sailer, T. Jaeger, E. Valdez, R. Caceres, R. Perez, S. Berger, J. L. Griffin, and L. Van Doorn, "Building a MAC-based security architecture for the Xen open-source hypervisor," in Computer security applications conference, 21st Annual, 2005, p. 10–pp.

[62] M. G. Martinek, M. D. Jackson, D. R. Kingham, and T. S. Wasinger, "Video gaming apparatus for wagering with universal computerized controller and I/O interface for unique architecture." Google Patents, 2005.

[63] M. P. Casey, A. G. Englman, D. P. Fiden, J. M. Hornik, and J. R. Jaffe, "Gaming machine with interactive pop-up windows providing enhanced game play schemes." Google Patents, 2009.

[64] R. Godwin-Jones, "Emerging technologies: Messaging, gaming, peer-to-peer sharing: Language learning strategies & tools for the millennial generation," Lang. Learn. Technol., vol. 9, no. 1, pp. 17–22, 2005.

[65] Z. H. S. H. W. Fengke, "Application study of development kit gtk+ technique on linux gui [j]," Comput. Appl. Softw., vol. 1, p. 50, 2009.

[66] M. Gouy, S. Guindon, and O. Gascuel, "SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building," Mol. Biol. Evol., vol. 27, no. 2, pp. 221–224, 2010.

The Problem of Universal Grammar with Multiple Languages: Arabic, English, Russian as Case Study

Nabeel Imhammed Zanoon
Department of Applied Science, Al- Balqa' Applied University (BAU)
Aqaba-Jordan

***Abstract*—Every language has its characteristics and rules, though all languages share the same components like words, sentences, subject, verb, object and so on. Nevertheless, Chomsky suggested the theory of language acquisition in children instinctively through a universal grammar that represents a universal grammar for all human languages. Since it has its declaration, this theory has encountered criticism from linguists. In this paper, the criticism will be presented, and the conclusion that the general rule is not compatible with all human languages will be suggested by studying some human languages as a case namely Arabic, English, and Russian.**

Keywords—Chomsky; linguistic; Universal Grammar; Arabic; English; Russian

I. INTRODUCTION

All languages contain in their structure of the categories of the word, a noun and a verb. However, there is a difference through the category of open-class words such as nouns, verbs, and adjectives, which are regarded as the foundation stone in the formation of sentences in a language, while the categories of the closed-class words consist of the articles, conjunctions as well as prepositions. These words are empty of meaning [1]. Their abstract form which demonstrates the meaning of the phrase represents the deep structure while the surface structure is what we write and say. However, there is a relationship of transformations between them such as combination, addition, and deletion. Chomsky has presented transformational rules which he built on the duality of linguistic structure [2].

Philosophers and psychologists started since the twentieth-century research in the phenomena of language learning and mastery. It became obvious that knowledge of the language does not depend only on the connection between the words; it is made up by knowing how to put words together because language is made up of sentences that express our thoughts. If the knowledge of a language is acquired by knowing all the mysterious rules, a question is raised on how children can learn complex rules in language [3]. Linguists believe that the sentence is the basic structure and characteristic of the human language as all human languages are made up by of syntactic patterns. A Syntactic pattern is a model which identifies human language. Chomsky, in 1972, showed that human beings had a language acquisition device and put forward the Universal Grammar (UG) theory [4]. This is a comprehensive grammar theory which assumes that there are general rules common to all languages. It explains the principles of language acquisition , and it is not concerned with describing specific languages [5], due to the fact that universal rules show that children use them to understand and acquire their mother tongue because they stipulate that the universal rules are rules for all languages [6]. The language acquisition device is called Universal Grammar which provides children with the principles of a universal language and grammatical structures with an instinctive hypothesis; it suggests that our ability to learn language rules is already found in the brain. This theory states that language ability appears by itself without being taught and that there are characteristics common to all human languages [7]. The problem of the universal grammar with other languages will be discussed by studying the syntactic structure in several languages and comparing them to each other, and by designing a finished cases device that represents the structure and arrangement of words in the languages in order to prove that each language is specific in the linguistic structure and word order, and any change in that leads to a difference in meaning. Moreover, some of the criticism by linguists will be discussed and cited to support the idea put forward in this paper.

II. RELATED WORK

In [8] showed in his study that the universal grammar is a suspected, and that the evidences put forward are weak, since there are some arguments in favor of the general rules without any evidence to support them, but there is no general model for general rules. They are a set of proposals, with a presentation of the views of some researchers and scientists, including, that child language learning and development varies from child to child in terms of syntax. In his paper, he stated that the general rules, in fact, do not exist, and presented a series of criticisms from different sources.

In [9] showed in his study that Chomsky's hypothesis was not widely accepted, and that it is just a theory, citing Piaget that the hypothesis is contrary to the truth, because knowledge acquisition is through experience and work. He said that if the language learning is a simple acquisition process from childhood, the child will not be in need of learning anything related to language.

He denied that the universal grammar is innate, and illustrated this by several examples. He concluded that learning the grammatical rules is endless, since languages have infinite probabilities in the formation of sentences and learning languages, that is to say, children may take a long time to learn the language rules in order to stop committing grammatical errors.

III. General Rules and Human Languages

The human languages consist of letters, words and sentences. Each language has its special script and terminology, for all languages consist of nominal and verbal sentences, and these sentences include a noun which functions as the subject, object or case. What regulates and adjusts the sentences and word order in language is the language rules, and each language has its rules and word order. Language is generated through an input of words, groups of words and sentences, and in order to produce sentences that represent language the input must be processed, and through the application of the rules, the output will be compatible with the input as in Fig (1).

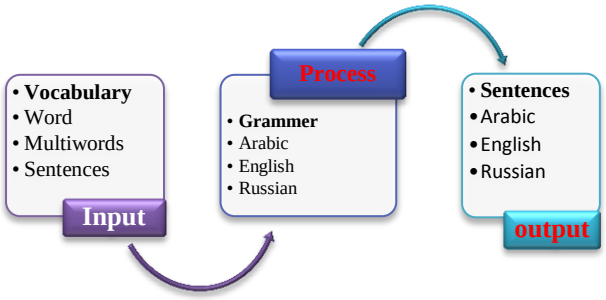


Fig. 1. The system to understand language

For example, if the input is in Arabic, can you process the input by the rules of the English language or any other language, and would the output be in Arabic or English? If we took an Arab child from birth and made him live with two boys who speak English and live in a country that speaks English, will he speak the Arabic or the English Language? Of course, he will speak English because the input was in English and through the linguistic application which accompanies a child's growth the rules of the English language will gradually be applied, and the output will consequently be in English.

We go back to the previous example of the Arab child if we apply the universal grammar theory what the output of the child will be? Based on the text of the theory, the child will speak Arabic because he acquires language and rules by instinct, that is, through the device of comprehensiveness of language found in the brain, regardless of input from the environmental.

IV. Acquisition Process

With the increasing vocabulary of the child, it is believed that he is somehow taught the language which means that children are using what is said to build possible ways to use the language. The linguistic production of children shows that it is often a kind of an experience or test for some structures as to whether they were correct or not. One of the factors that appear to be important in the process of a child acquisition of language is the actual use of sounds and words, whether in communicating with others or in dealing with words alone.

The human system, just like a computer system, consists of input, processing and output; language acquisition by children takes place by mingling and communicating, for conversation and dialogue are kinds of input. Processing takes place in the brain and outputs through language, that is, conversation and dialogue. Eyesight is also regarded a sort of input, and the processing takes place in the brain while the output is done through behavior and other matters such as touch. The brain does not create something out of nothing but it has some processes depending on the input the proof is that if we put a child in isolation from humans beginning from the first month to the age of 3 or 4 years, will he speak any language, and from where will he acquire the language? So the child's brain does not contain a linguistic device that automatically generates language.

Language is made up of sentences composed of words, and the child stores a large number of words and through dialogue sentences are structured out of these words. Sentences are made by dictating orders; for example, if a child stored several words such as door, open, come, here, box, book ... etc., he will receive the sentences in the form of order 'come here', and upon hearing these sentences, he will obey the order. In this way, these sentences become compound ones. In case he wants to speak to someone else he will say 'come here' and a few sentences are made that way, and so the child learns the language. When he enters school, the language is learned in order develop the language skills he learned. In other words, he learns the rules of grammar which represent the controls for the syntax of sentences, and so he starts making sentences by himself without the help of dictation method.

V. Sentence Structure In Language

All human languages consist of sentences, but they vary in the sentence structure, as it shows the physical nature of the sentence and explains the elements from which the sentence is made up [10]. The word order has to do with the arrangement of the grammatical structure of language, for human languages differ in the order of words, that is to say, the way sentences are structured of the language fundamental components. This is a feature which distinguishes a language from another as seen by linguists. One of the divisions of these scholars of languages is based on the way sentences are structured in the discourse of a particular human group. They divide languages into various types according to the succession of a sentence (Subject), (Verb) and (Object) as well as the (complements), which is regarded as a distinctive feature of a particular language. A sentence, any sentence, consists basically of a verb, a subject, and an object, with other additions [11]. There are six patterns that represent the word order in a language: they are add (SVO) subject, verb, object, (SOV) subject, object, verb, (VSO) verb, subject, object, (VOS) verb, object, subject, (OSV) object, subject, verb, and (OVS) object, verb, and subject. The overwhelming majority of the world's languages follow either SVO or SOV patterns [12]. Some languages have a fixed word order, and others have a free unfixed word order [13].

The word order in the human language is arranged on several structures that consist of the subject (S), the object (O), and the verb (V), and there are six structures for word order. Languages have been classified into categories according to the word order structure that can be found in human languages [14] as in table (1). There is a study on the classification of languages and distribution of word order in the map of the

world as in the world Atlas of Linguistic Structures Online [15].

TABLE I. WORD ORDER AND DISTRIBUTION IN HUMAN LANGUAGES

Word order	English equivalent	Proportion of languages	Example languages
SOV	"She him loves."	45%	Pashto, Latin, Japanese, Afrikaans
SVO	"She loves him."	42%	English, Hausa, Mandarin, Russian
VSO	"Loves she him."	9%	Biblical Hebrew, Irish, Filipino, Tuareg
VOS	"Loves him she."	3%	Malagasy, Baure
OVS	"Him loves she."	1%	Apalaí?, Hixkaryana?
OSV	"Him she loves."	0%	Warao

A. Sentence Structure in the Arabic Language

The Arabic language is different from other languages, where the sentence has various word orders like SVO and VSO. The Arabic language is rich in grammatical structures and is different from English in that the word order is not fixed. When the word order in the sentence changes it will not affect the sentence meaning unlike the English language, as can be shown in the examples [16].

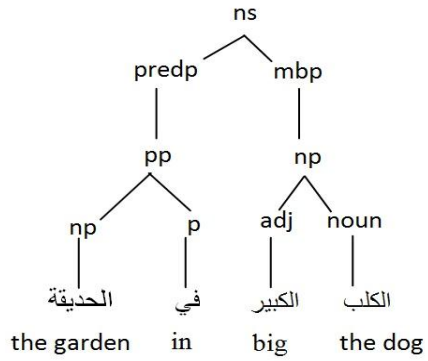


Fig. 2. Sentence Structure in Arabic ,An Example of nominal sentence

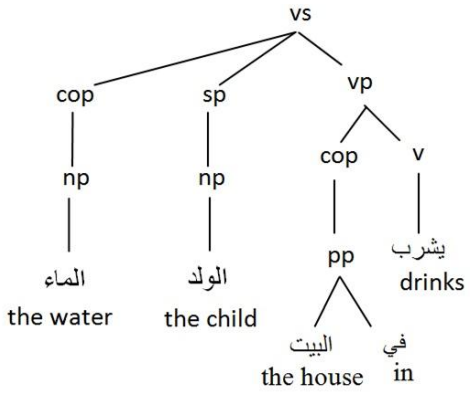


Fig. 3. Sentence Structure in Arabic ,An Example of verbal sentence

The Arabic language is characterized by the relatively free order of words; the Arabic sentence is diverse by the several word order forms such as VSO, SVO, and VOS, as in the examples in (Table 1) [17].

B. Sentence Structure in the Russian Language

The word order in Russian is not fixed; the sentence in the Russian language may be composed of a combination of the word order, that is to say, the place of the subject, the verb and object of the sentences can be changed without any change in the meaning of the sentences [18]. Flexibility in the Russian language means that the sentence admits the six-word order structures (SVO, SOV, VSO, VOS, OSV, and OVS) without any change in the meaning of the sentence. For example, in the English sentence "the boy read the paper" if we make all the possible six orders of the sentence in the Russian language, there will be no change in meaning, as in Table (2) below [19]:

TABLE II. COMPARISON BETWEEN SENTENCES IN RUSSIAN AND ENGLISH

Russian language	Order word	English language
Коля купил машину	S V O	**Kolya** Bought the car (neutral)
Коля машину купил	S O V	**Kolya** BOUGHT the car
Купил **Коля** машину	V S O	**Kolya** did bought the car
Купил машину **Коля**	V O S	**KOLYA** bought the car
Машину **Коля** купил	O S V	the car, **Kolya** BOUGHT it
Машину купил **Коля**	O V S	The car, it was **Kolya** who bought it

C. Sentence Structure in the English language

The sentence structure of the English language consists of a subject, a verb and an object (SVO) known as (canonical word order). Word order in the English language is fixed, since the subject comes first in the English sentence [20], like other languages which consist of the nominal sentence and the verbal sentence. The structure of grammatical sentences is shown in (Figure 4) [21]:

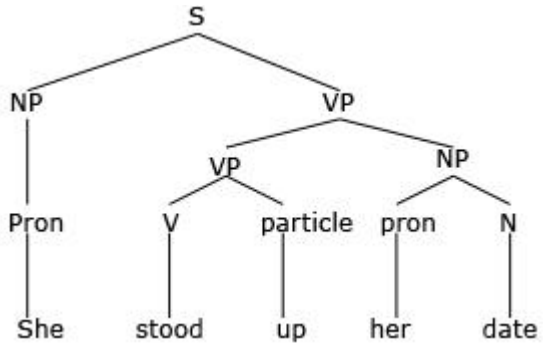


Fig. 4. Sentence Structure in the English Language

VI. Criticism

First, when Chomsky proposed the universal grammar and child language acquisition theory he did not study languages, that is, he did not take samples from several languages, but he thought of this language only at the level of the nature of the English language, and that is why the rule was void of analysis, evidence, and application as forms of proof of their existence. Languages are similar it in general, but in particular, they are different, as already discussed in previous sections of this paper. Second, this grammar and theory were proposed on the principle of thought and theory analysis which are far from the application. Third, language is based on knowledge and skills just like other sciences and skills. So, is there a part of the brain that is devoted to each science and skill, for music, art and sports? All science and skills are learned and acquired through practice, education, and application; if all people acquire language innately and learn everything by instinct, it will be a sign that the level of knowledge among all human beings is equal.

The Universal grammar and child language acquisition were not received warmly since they were initiated by Chomsky. Rather, they have received criticism by linguists. This criticism will be identified and discussed below.

Many linguists opposed the universal grammar theory, including Jeffrey Sampson, who talked about the theory as an incorrect or false one and described it as unrealistic observations and views for language. There are many opinions which suggested that there is no basis for the universal grammar theory [22] and that it does not have any evidence or proof, including the underlying items. This was Ray Daniel Everett's view, but some others denied the existence of universal grammar altogether and advocated that it was unrealistic, and there was no evidence to prove its existence. There are several factors that play a role in the organization of communication and dialogue, and this is what was issued before [23].

There is a lot of criticism and among the most prominent critics is Jean Piaget with whom Chomsky has a debate that the theory lacks a concrete reality, where one can acquire knowledge of a certain thing through practice, experience and comprehension. Chomsky suggests acquiring knowledge of language by providing general rules for all languages. However, Jean Piaget expressed his opinion that Chomsky's hypothesis could not be accepted by the premise of the "fixed innate nucleus" because he neither interpreted nor proved it.

Ray Skinner had well-known views on language acquisition. He stated that knowledge is acquired through the environmental and reinforcement, where children learn language through input which is "the environmental conditions as a result of training for by caregivers [24]. People around the child have an influence on the acquisition of language. The scholar Tomasello, who is one of Chomsky's critics, also shows his opinion. He states that children acquire language by understanding how to use the language of others around them [25].

VII. Finite State Automaton and Universal Grammatical Rules

All human languages consist of words, but these words are subject to an arranged and tidy grammatical order since each language adopts a particular word order as has been discussed in the previous sections of this paper. Finite state automaton leads to the correct meaning of the sentence. If the arrangement is not consistent with the approved order of the language, the meaning of the sentence would be incorrect. Accordingly, the general rules are not consistent with all languages. In this paper, a finite state automaton was set up; cases represent the subject (S) the verb (V) and the object (O). A sentences in several languages (Arabic and English and Russian) will be tested. The finite state automaton represents here the role of the general rules in order to see the compatibility of the general rules with the languages.

The word order in the English language plays a key role because the grammatical meaning depends on the order of words. In the Russian sentence, if we change the position of the words within the sentence, the general meaning of the sentence will not change. The Russian language is compatible with all arrangements, that is to say, it is free in the arrangement, and so is the Arabic language which has a free feature of free word order. Arabic is compatible with the range arrangements SVO, VSO, VOS and OVS, for if the position of the subject in the sentence is changed the meaning remains the same, unlike the English language where the meaning changes, as shown in the examples [26].

A finite state automaton acts as the universal grammar in this paper. In figure 5 a finite state automaton accepts languages that are compatible with all word orders such as Arabic and Russian while the finite state automaton in Figure 6 accepts all languages that comply with the word order that begins with the subject such as English. Table (3) shows the application of sentences of different languages to the finite state automaton and cases of acceptance and rejection in terms of the word order of a language and the correct meaning of the sentence.

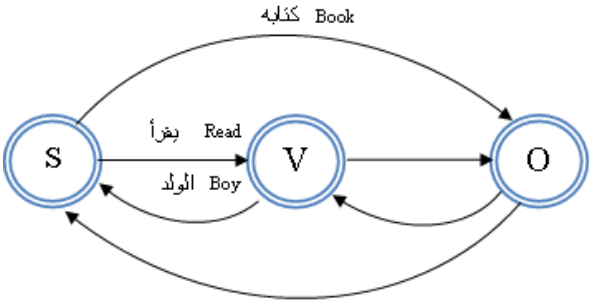


Fig. 5. finite-state automaton, multi-lingual

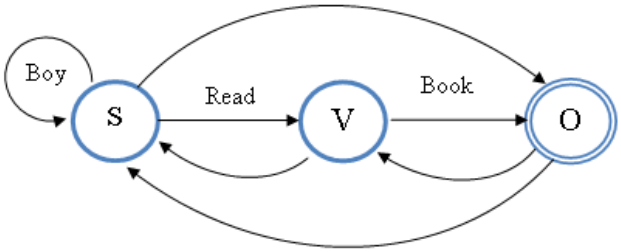


Fig. 6. finite-state automaton, English

We note that there is a difference between languages in the structure of sentences which adjusts the structure of sentences in the language, that is the language rules, and in light of this we conclude that languages share the components of words and differ in their structure and rules. Nevertheless, it is difficult to find general rules that represent all languages, that is, each language is unique by its rules and characteristics. As shown in the table (4), the derivation of sentences from the same rules is different and some sentences are no longer correct regarding of structure and meaning.

TABLE III. RESULTS OF THE APPLICATION OF SENTENCES ON THE FINITE STATE AUTOMATON

The final status (meaning of the sentence)	Order Word	Language	Sentence
تقبل	SVO	العربية	الولد يقرأ الكتاب
تقبل	SVO	الانجليزية	The boy reads the book
تقبل	SVO	الروسية	Мальчик читает книгу
تقبل	VSO	العربية	يقرأ الولد كتابه
لاتقبل	VSO	الانجليزية	Read the boy the book
تقبل	VSO	الروسية	Читайте мальчику книгу
تقبل	OVS	العربية	الكتاب يقراه الولد
لاتقبل	OVS	الانجليزية	The book reads the boy
تقبل	OVS	الروسية	Книга читать мальчика

TABLE IV. SENTENCE DERIVATION FROM RULES BY WORD ORDER

Order Word	Grammar Rules	English language	Arabic Language	Russian Language
SVO	**sentence** –> **<subject> <verb-phrase> <object>** **subject** –> boy | I **verb-phrase** –> **<adverb> <verb>** | **<verb>** **adverb** –> always **verb** –> is | read | am **object** –> the **<noun>** | a **<noun>** | **<noun>** **noun** –> book | Newspaper	The boy reads the book	الولد يقرأ الكتاب	Мальчик читает книгу
VSO	**sentence** –> **<verb-phrase> <subject> <object>** **subject** –> boy | I **verb-phrase** –> **<adverb> <verb>** | **<verb>** **adverb** –> always **verb** –> is | read | am **object** –> the **<noun>** | a **<noun>** | **<noun>** **noun** –> book | Newspaper	Read the boy the book	يقرأ الولد كتابه	Читайте мальчику книгу
VOS	**sentence** –><**verb-phrase**> **<object><subject>** **subject** –> boy | I **verb-phrase** –> **<adverb> <verb>** | **<verb>** **adverb** –> always **verb** –> is | read | am **object** –> the **<noun>** | a **<noun>** | **<noun>** **noun** –> book | Newspaper	Read the book boy	يقرأ الكتابه الولد	Читайте мальчику книгу
OVS	**sentence** –><**object**> **<verb-phrase> <subject>** **subject** –> boy | I **verb-phrase** –> **<adverb> <verb>** | **<verb>** **adverb** –> always **verb** –> is | read | am **object** –> the **<noun>** | a **<noun>** | **<noun>** **noun** –> book | Newspaper	The book reads the boy	الكتاب يقراه الولد	Книга читать мальчика

VIII. CONCLUSION AND FUTURE WORK

Several languages have been studied as case studies regarding of structure, word order, and the rules in order to know the compatibility of the universal grammar rules with those languages. We conclude that there is a problem in compatibility between the rules of universal grammar and human languages, and this shows that the child learns the mother language in the surrounding environment by acquiring skills and knowledge. This has between shown by linguists where the theory was criticized, and a finite state automaton has been set up which acts as a language device that has a universal grammar to examine languages compatibility. By

citing criticism, analysis and discussion, the universal grammar rules are having a trouble of incompatibility with human languages; each language has its special rules, but which share other languages only the elements of verb, subject, and object.

In the future, the theory of transformational generative grammar and probability theory will be applied, by choosing some words and forming sentences in several languages. These sentences will be checked as to whether they comply with the rules of languages in terms of meaning and structure. The percentage of the sentence accuracy for each language will be calculated, which gives an indication that there is no universal grammar for all languages.

REFERENCES

[1] C. Brown , P. Hagoort, and M. Ter Keurs , "Electrophysiological signatures of Visual Lexical Processing: Open- and Closed-Class Words," Journal of Cognitive Neuroscience, vol. 11, pp. 261 – 281, May 1999.

[2] B. Burke, " Rituals and beliefs ingrained in world language pedagogy: Defining deep structure and conventional wisdom," Journal of Language Teaching and Research, vol. 2, pp. 1-12, January 2011.

[3] Fiona Cowie ,Innateness and Language, http://plato.stanford.edu/entries/innateness-language/,*Jan* 2008.

[4] C. Derek, J. Keith, and J. Daniel, "Darwin's mistake: Explaining the discontinuity between human and nonhuman minds." Behavioral and Brain Sciences ,vol 31, pp. 109-130, April 2008.

[5] E. Kandel, E-Study Guide for Principles of Neural Science, Cram101 Textbook Reviews , 5th Edition,2013.

[6] V. Cook., Multilingual Universal Grammar as the norm. In I. Leung (ed.) Third Language acquisition and Universal Grammar, Bristol: Multilingual Matters, pp. 55-70,2009.

[7] B. Rowe, D. Levine, "A Concise Introduction to Linguistics",ch8, pp. 233,Routledge 2015.

[8] Dąbrowska, E. (2015). What exactly is Universal Grammar, and has anyone seen it?. Frontiers in psychology, 6.

[9] Samúelsdóttir, B. (2015). The Innateness Hypothesis: Can Knowledge of Language be Inborn?.

[10] A. Akmajian, A. Demer, and K. Farme , M. Harnish "Linguistics: An Introduction to Language and Communication," pp.153-604, MIT Press 2001.

[11] D. Carroll,"e-Study Guides for: Psychology of Language," Cram101 Textbook Reviews, 2013.

[12] D. Kemmerer,"The Cross-Linguistic Prevalence of SOV and SVO Word Orders Reflects the Sequential and Hierarchical Representation of Action in Broca's Area," Language and Linguistics Compass, vol 6, pp.50–66, January 2012.

[13] I. Laka , K. Erdocia, "Linearization Preferences Given "Free Word Order" Subject Preferences Given Ergativity: A Look At Basque," Of grammar, words, and verses, In honor of Carlos Piera, Ch 6, pp.115-140,2012.

[14] F. Meyer, "Introducing English Linguistics International Student Edition," Cambridge University Press 2010.

[15] Dryer, Matthew S. & Haspelmath, Martin (eds.) 2013. The World Atlas of Language Structures Online.

[16] Leipzig: Max Planck Institute for Evolutionary Anthropology.(Available online at http://wals.info, Accessed on 2016-01-15.)

[17] M. Al Aqad, "Syntactic analysis of Arabic adverb's between Arabic and English: X bar theory," International Journal of Language and Linguistics,vol 1(3), pp.70-74, August 2013.

[18] H. Bassam, M. Asma, and O. Nadim, T. Abeer," Formal description of Arabic syntactic structure in the framework of the government and binding theory," Computación y Sistemas, vol 18(3), pp. 611-625, sep 2014.

[19] L. Isurin, " Cross Linguistic Transfer in Word Order: Evidence from L1 Forgetting and L2 Acquisition," Proceedings of the 4th International Symposium on Bilingualism, pp.1115–1130. Somerville, MA: Cascadilla Press 2005.

[20] A.Kharitonova, " Lexical transfer and avoidance in the acquisition of English phrasal verbs," MA Thesis, University of Oslo,May 2013.

[21] K. Christianson, F. Ferreira, "Conceptual accessibility and sentence production in a free word order language (Odawa) ," International Journal of Cognitive Science, Elsevier Cognition vol 98, pp. 105–135,2005.

[22] N. Oostdijk , "Corpus Linguistics and the Automatic Analysis of English," pp-89-267, Rodopi 1991.

[23] J. Hurford, "Nativist and Functional Explanations in Language Acquisition," Logical Issues in Language Acquisition, pp.88-136. 1995.

[24] Universal Grammar, (n.d.). In *Wikipedia*. Retrieved 10 February 2016from https://en.wikipedia.org/w/index.php?title=Universal_Grammar&oldid=704299963.

[25] B. Samúelsdóttir," The Innateness Hypothesis Can Knowledge of Language be Inborn?," http://hdl.handle.net/1946/20344,Thesis, University of Iceland, 2015.

[26] all about linguistics to discover and understand, https://sites.google.com/a/sheffield.ac.uk/all-about-linguistics/branches/language-acquisition/who-studies-language-acquisition, University of Sheffield 2015.

The Application of Fuzzy Control in Water Tank Level Using Arduino

Fayçal CHABNI, Rachid TALEB, Abderrahmen BENBOUALI, Mohammed Amin BOUTHIBA
Electrical Engineering Department, Hassiba Benbouali University, Chlef, Algeria
Laboratoire Génie Electrique et Energies Renouvelables (LGEER)

***Abstract*—Fuzzy logic control has been successfully utilized in various industrial applications; it is generally used in complex control systems, such as chemical process control. Today, most of the fuzzy logic controls are still implemented on expensive high-performance processors. This paper analyzes the effectiveness of a fuzzy logic control using a low-cost controller applied to a water level control system. The paper also gives a low-cost hardware solution and practical procedure for system identification and control. First, the mathematical model of the process was obtained with the help of Matlab. Then two methods were used to control the system, PI (Proportional, Integral) and fuzzy control. Simulation and experimental results are presented.**

Keywords*—Fuzzy control; PI; PID; Arduino; System identification

I. Introduction

The extraordinary development of digital processors (Microprocessors, Microcontrollers) and their wide use in control systems in all fields have led to significant changes in the design of control systems. Their performance and low cost make them suitable for use in control systems of all kinds that require a lot more capabilities and performance than those provided by the analog controllers.

In certain industry branches, the liquid level control problem is often encountered. The nature of the liquid and friction of control mechanism and other factors make the system nonlinear [1, 2]. In nowadays, the best-known industrial process controller is the PID controller because of its simplicity, robustness, high reliability and it can be easily implemented on any processor, but using a PID controller is not fully convenient when it comes to dealing nonlinear systems [3, 4]. But these systems can be successfully controlled using fuzzy logic controllers because of their independency from the mathematical model of the system.

In this paper, PI (Proportional, Integral) and fuzzy logic controllers are applied to water level control system; the proposed fuzzy controller has better performance than conventional control methods with a simpler algorithm that can be easily implemented on a microcontroller. The PI and fuzzy controllers are implemented on Arduino, which is an open source development board.

This paper is organized as follows: in the next section, a description of the system is presented. The system identification phase and the mathematical model are presented in the 3rd section; the 4th section describes the PI controller implementation, the simulation, and experimental results are presented. In the 5th section a general description of a fuzzy logic controller implementation, simulation, and experimental results are presented. The conclusion is given in the last section

II. System Description

Adjusting a liquid level in a tank is the main objective of this work, the structure of the entire system is as shown in Fig. 1. The system consists of a water tank, a liquid level sensor, a pump based on a 12V direct current motor, an electronic circuit (Arduino and a DC/DC step-down converter).

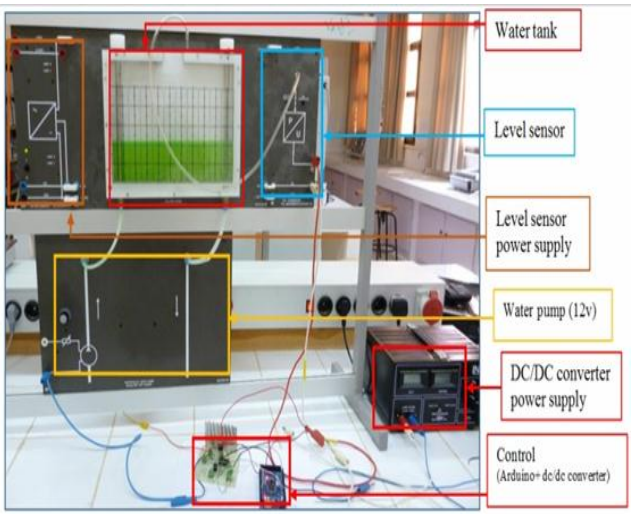


Fig. 1. Structure of water level control

The structure chart of the water tank level system is shown in Fig. 2 which the liquid flows into the top of the tank with the help of a dc motor pump and leaves from the bottom, through a pip equipped with an adjustable valve to adjust manually the flow rate of the liquid leaving the tank and to simulate leaks (disturbances).

The Arduino is used as an acquisition board in identification phase, once the mathematical model of the system is obtained, the Arduino will play the role of an independent controller. A computer is needed to display signals and to impose set points for the controller; the computer will communicate with the Arduino through the RS232 communication protocol.

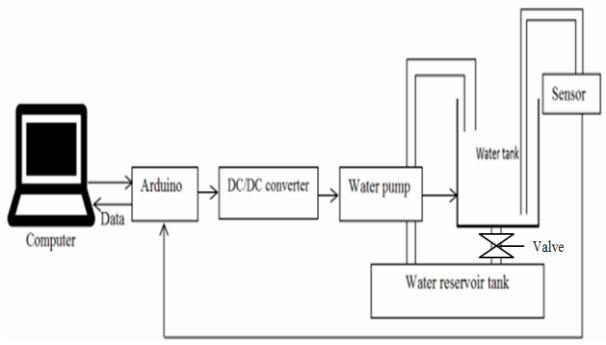


Fig. 2. Structure chart of water tank control system

III. System Identification

To obtain the mathematical model of the process, the Arduino board was used as an interface between the computer and the system. The computer is equipped with software that can store incoming samples from the board. "MATLAB identification toolbox" shown in Fig.3 has been used to process the samples and to obtain the mathematical model of the system.

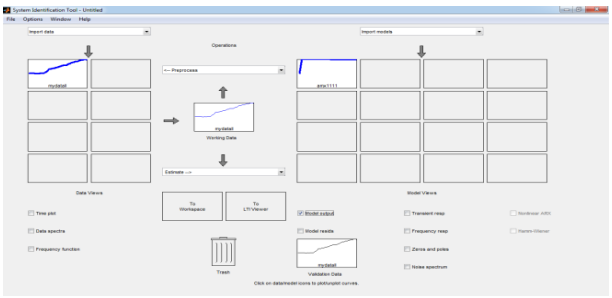


Fig. 3. Graphical user interface (GUI) of the identification tool box

Fig. 4 shows the open loop response of the system to a constant input $u(t)$, 15.8cm is the final value of the output $y(t)$ to a 7.6cm input. This difference between $u(t)$ and $y(t)$ corresponds to the steady-state error of 8.2 cm. That is why a controller is needed to minimize the steady-state error. The transfer function of the system was found with the help of MATLAB identification toolbox, and it is as follows:

$$G(z) = \frac{0.004483}{z + 0.8852}, \text{sampling time } Ts = 0.2s \quad (1)$$

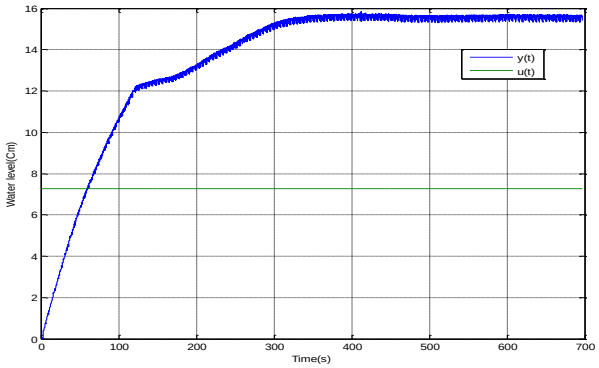


Fig. 4. Response of the system

Fig. 5 represents a comparison between system response and transfer function response to the same input. And it can be seen that the transfer function response, almost matches the real system response.

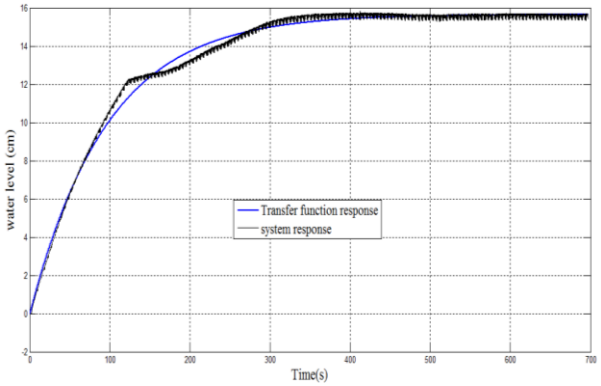


Fig. 5. Comparison between system response and transfer function response to the same input

IV. PI Controller

A Proportional-Integrate-Derivative controller (PID) is a control mechanism, the role of this controller is to minimize the error between a set point and the system response, the control algorithm contains three terms proportional, integrate and derivative term [5, 6]. The most popular controller industrial field is the PI (Proportional-Integrate) controller, and it is a special case of a PID controller, it has only two constant parameters K_p and K_i, where K_p is the proportional gain and K_i is the integral gain [7, 8]. The control algorithm $u(t)$ and the controller transfer function $C(p)$ are given by the following equations:

$$u(t) = K_p(\varepsilon(t) + \frac{1}{\tau_i}\int_0^t \varepsilon(t)dt \quad (2)$$

$$C(p) = K_p \frac{1+\tau_i p}{\tau_i p} = K_p(1 + K_i \frac{1}{p}) \quad (3)$$

The design of the PI controller was done using Matlab/Simulink, and it was based on the mathematical model obtained from the identification phase. The simulation is shown in Fig.6; it was used to test the performance of the controller, the gains (K_p and K_i) were calculated using pole placement method, (K_p = 1.145 and K_i = 0.015). Fig. 7 shows the results obtained by the simulation.

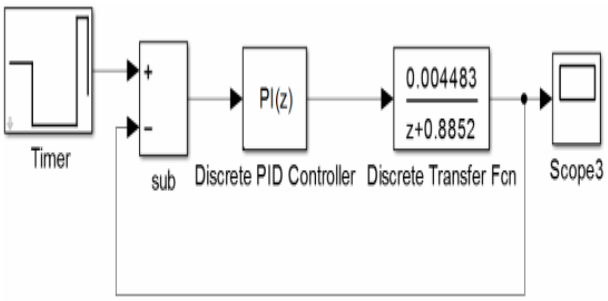


Fig. 6. Simulation of PI controller in Simulink

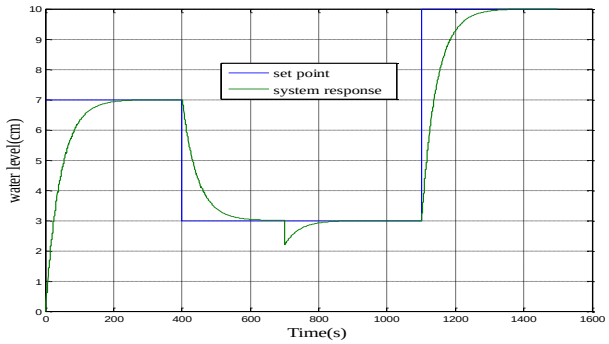


Fig. 7. Behavior of the process with a PI controller (simulation)

After the controller had been designed and tested in Matlab/Simulink, the function of the controller mentioned earlier was implemented in Arduino to control the system. Fig 8 presents the behavior of the system with PI controller.

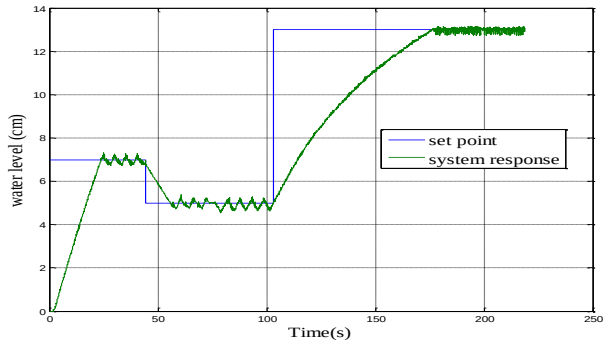


Fig. 8. Behavior of the process with a PI controller (experimental results)

V. FUZZY LOGIC CONTROLLER

The Fuzzy Logic controller consists basically of four parts: fuzzification interface, knowledge base, inference engine, and a defuzzification interface. Fig. 9 shows the basic configuration of a fuzzy logic controller. Each of these parts plays a different role in the control process and affects the performance of the controller and the behavior of the whole system. The fuzzification is the transformation of numerical data from the input to linguistic terms. The knowledge base provides necessary information for all the components of the fuzzy controller [9-11]. The fuzzy inference engine or the logical decision-making is the core (brain) of the controller. It is capable of simulating the decision-making of human beings. At the end of the inference step, the obtained result is a fuzzy value that cannot be directly used to control the process, so the value should be defuzzified to obtain a crisp value, and that is the role of the defuzzification interface.

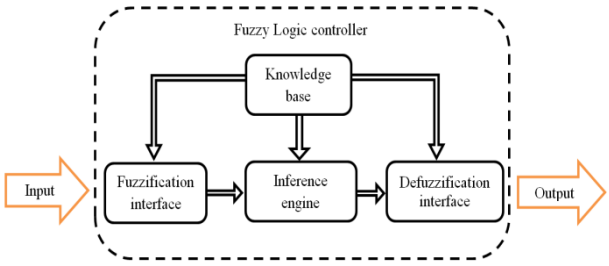


Fig. 9. Basic configuration of a fuzzy logic controller

The fuzzy logic controller usually works with more than two input signals, the system error $e(k)$ and the change rate in the error $\Delta e(k)$. The error of the system is defined as the difference between the set point $y_r(k)$ and the plant output $y(k)$ at a moment k:

$$e(k) = y_r(k) - y(k) \quad (4)$$

The variation of the error signal at the moment k is given by the following equation:

$$\Delta e(k) = e(k) - e(k-1) \quad (5)$$

The configuration of the proposed fuzzy controller is shown in Fig.10. In1 is the system error, and In2 is the variation of the error signal.

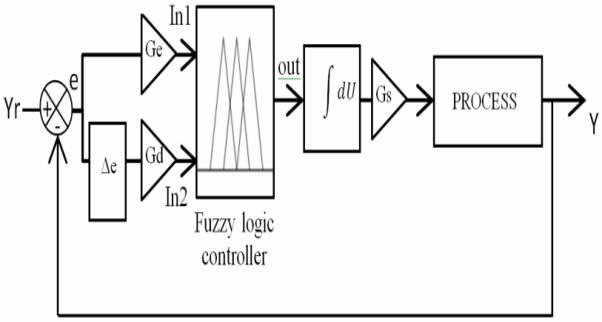


Fig. 10. Fuzzy controller in a closed loop system

The simulation shown in Fig. 11 was used to test the performance of the fuzzy controller and to determine the controller gains.

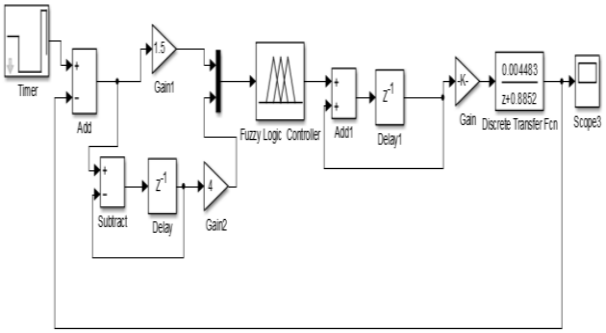


Fig. 11. Simulink model

Using Matlab toolbox "fuzzy logic toolbox", shown in Fig. 12, a fuzzy logic controller was designed containing two inputs (error and error derivative) and one output. The proprieties of our controller are given in the Table. 1.

TABLE I. PROPRIETIES OF THE FUZZY LOGIC CONTROLLER

Controller type	Mamdani
And method	Min
Or method	Max
Implication	Min
Defuzzification	Centroid

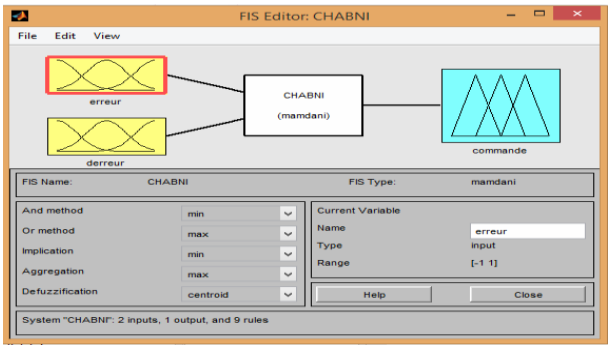


Fig. 12. Graphical user interface of the fuzzy logic toolbox

The chosen membership functions of output and input signals are all similar; they are shown in Fig. 13.

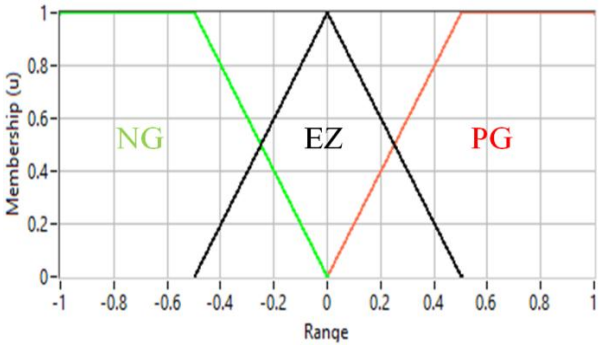


Fig. 13. Membership functions of In1 and In2 and out

The design of the table below (Table .2) was based on the principles of a basic control system which are: If the error is big, and the error rate changes quickly, then the controller should eliminate the error quickly and if the error is small, and the error rate change is not fast, then the controller should eliminate the error slowly and if the error is zero, and the error rate doesn't change, then the control command should be zero. The labels inside the table are the linguistic variables.

TABLE II. FUZZY RULES

In2 \ In1	NG	EZ	PG
NG	NG	NG	EZ
EZ	NG	EZ	NG
PG	EZ	PG	PG

The labels in Table 2 are as follows: NG = very low, EZ = zero and PG = very high.

The values of the controller constants were found after preforming simulations in Matlab. Table 3 shows the constants values. The result of the last simulation is presented in Fig. 14.

TABLE III. CONTROLLER GAINES

Error gain (Ge)	1.5
Error changing rate gain (Gd)	4
Output gain (Gs)	150

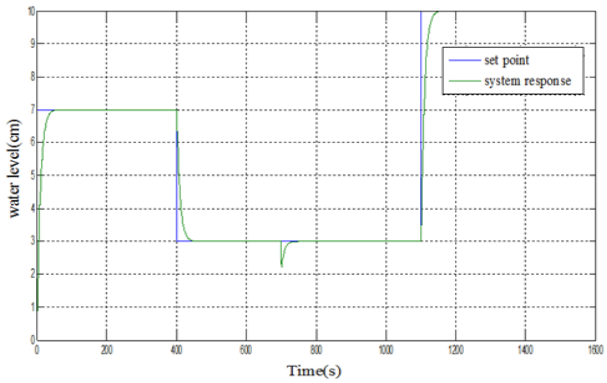


Fig. 14. Behavior of the process with a fuzzy controller (simulation)

After the controller was designed and tested in Matlab/Simulink, it was implemented on Arduino to control the system. Fig. 15 presents the behavior of the system with a fuzzy logic controller.

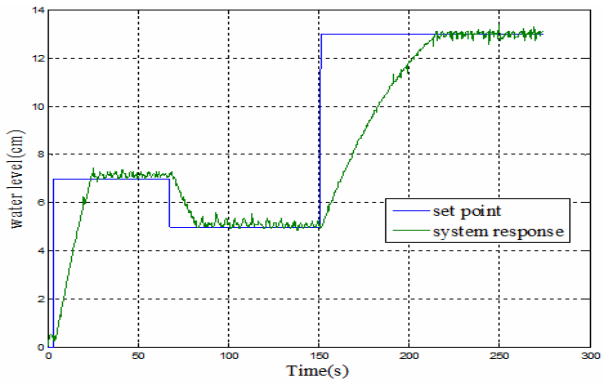


Fig. 15. Behavior of the process with a fuzzy controller (experimental results)

The system was subjected to disturbances (in simulation and experiment). It is seen from Figs. 7, 8, 14 and 15 that the fuzzy controller had better performance and stability in every given set point and fast error compensation.

VI. CONCLUSIONS

This paper proposed a low-cost solution to apply fuzzy logic control for a water tank level control system using an Arduino, and using it also as a low-cost acquisition board for system identification. The main objective of this work has been reached, which is to test the effectiveness of fuzzy logic control using Arduino, by comparing it to a PI controller. The general structure of both controllers (PI and fuzzy) were presented in this work. The simulations and experimental results showed the superiority of fuzzy control over the conventional control systems.

REFERENCES

[1] S. Krivic, M. Hujdur, A. Mrzic, S. Konjicija, "Design and implementation of fuzzy controller on embedded computer for water level control", Proceedings of the 35th International Convention, MIPRO, Opatija, pp. 1747-1751, 21-25 May 2012.

[2] P. Liu, L. Li, S. Guo, L. Xiong, W. Zhang, J. Zhang, C.Y. Xu, "Optimal design of seasonal flood limited water levels and its application for the Three Gorges Reservoir", Journal of Hydrology, Elsevier Ltd, vol. 527, pp. 1045-1053, August 2015.

[3] X. Fang, T. Shen, X. Wang, Z. Zhou, "Application and Research of Fuzzy PID in Tank Systems", 4th International Conference on Natural Computation (ICNC'08), Jinan, vol. 4, pp. 326-330, 18-20 October 2008.

[4] K. Ou, Y.X. Wang, Z.Z. Li, Y.D. Shen, D.J. Xuan, "Feedforward fuzzy-PID control for air flow regulation of PEM fuel cell system", International Journal of Hydrogen Energy, Elsevier Ltd, vol. 40, no. 35, pp. 11686-11695, 21 September 2015.

[5] V. Vindhya, V. Reddy, "PID-Fuzzy logic hybrid controller for a digitally controlled DC-DC converter", International Conference on Green Computing, Communication and Conservation of Energy (ICGCE), Chennai, pp. 362-366, 12-14 December 2013.

[6] Reshmi P. Pillai, Sharad P. Jadhav, Mukesh D. Patil, "Tuning of PID Controllers using Advanced Genetic Algorithm", IJACSA Special Issue on Selected Papers from International Conference & Workshop On Advance Computing, pp. 1-6, 2013.

[7] C.S. Tan, B. Ismail, M.F. Mohammed, M.F.N. Tajuddin, S. Rafidah, A. Rahim, Z.M. Isa, "Study of Fuzzy and PI controller for Permanent-Magnet Brushless DC motor drive", 4th International on Power Engineering and Optimization Conference (PEOCO), Shah Alam, pp. 517-521, 23-24 June 2010.

[8] A. Terki, A. Moussi, A. Betka, N. Terki, "An improved efficiency of fuzzy logic control of PMBLDC for PV pumping system", Applied Mathematical Modelling, Elsevier Ltd, vol. 36, no. 3, pp. 934-944, March 2012.

[9] Arindam Sarkar, J. K. Mandal, "Secured Wireless Communication using Fuzzy Logic based High Speed Public-Key Cryptography (FLHSPKC)", International Journal of Advanced Computer Science and Applications (IJACSA), pp. 137-145, vol. 3, no. 10, 2012.

[10] G. Bosque, I. Del Campo, J. Echanobe, "Fuzzy systems, neural networks and neuro-fuzzy systems: A vision on their hardware implementation and platforms over two decades", Engineering Applications of Artificial Intelligence, Elsevier Ltd, vol. 32, pp. 283-331, June 2014.

[11] Kawser Wazed Nafi, Tonny Shekha Kar, Md. Amjad Hossain, M.M.A. Hashem, "An Advanced Certain Trust Model Using Fuzzy Logic and Probabilistic Logic theory", International Journal of Advanced Computer Science and Applications (IJACSA), pp. 164-173, vol. 3, no. 12, 2012.

A Comparative Study of Databases with Different Methods of Internal Data Management

Cornelia Győrödi
Department of Computer Science and Information Technology, University of Oradea
Oradea, Romania

Alexandra Ștefan
Department of Computer Science and Information Technology, University of Oradea
Oradea, Romania

Robert Győrödi
Department of Computer Science and Information Technology, University of Oradea
Oradea, Romania

Livia Bandici
Faculty of Electrical Engineering and Information Technology, University of Oradea
Oradea, Romania

***Abstract*—The purpose of this paper is to present a comparative study between a non-relational MongoDB database and a relational Microsoft SQL Server database in the case of an unstructured representation of data, in XML or JSON format. We mainly focus our presentation on exploring all the possibilities that each type of database offers us, in the case that the data, which has to be stored, cannot or is not wanted to be normalized. This is a scenario most often found in production when, for the application that is being developed we are extracting unstructured data from social networks or all kinds of different channels that the user might have. The comparative study is based on the creation of a benchmark application developed in C# using Visual Studio 2013, which accesses databases created beforehand with proper optimizations that will be described.**

Keywords*—MongoDB; Microsoft SQL Server; NoSQL; non-relational database

I. Introduction

Nowadays, applications must support millions of users simultaneously and be able to handle a huge volume of data. A relational database model has serious limitations when handling huge volume of data. These limitations have led to the development of non-relational databases, also commonly known as NoSQL (Not Only SQL) [10].

The relational database model has a rigid schema which means that a schema must be designed in advance before data had been loaded and all attributes of the schema are uniform for all elements, in the case of missing values, null values are used instead [11]. Relational databases are known for their usefulness in terms of data that can be normalized and data that requires transactional integrity.

Non-relational databases do not store data in tables, the schema is not fixed and have very simple data model, and they can handle unstructured data such as documents, e-mail, multimedia, and social media efficiently as shown in [12].

We often encounter unstructured data in XML or JSON format, which cannot be normalized or normalization is not desired. It is important to know this type of data, when and why we should use a relational data model such as SQL Server database instead of a document-oriented database, such as MongoDB and what are the advantages and disadvantages.

It is necessary to do a careful analysis and consider main factors as the amount of data, the flexibility of schema, the budget, the amount of transactions that would be made, when choosing the data model for the application [13].

Generally, for smaller and medium applications, a relational database would be advisable and for big applications, that use and manipulate large quantities of data, a non-relational database is more appropriate [13].

In the first part of this paper, we will be presenting some information about SQL Server and the XML data type in SQL Server, and then we will continue with MongoDB and BSON data type. These will constitute a general knowledge that one needs to have in order to understand the logic behind each database type. In the second part, we will focus on experiments conducted with the help of the benchmark application in order to determine which of these two types of databases is more efficient and in what case. We will also present the experimental results and comparative study with the scenarios in which these results have an impact.

II. Unstructured Representation of Data in Microsoft SQL Server

Microsoft SQL Server is a relational database management system and it is one of the most popular systems used. We can securely say that, at the moment, the database market is dominated by systems that support the relational data model [1].

E. F. Codd proposed the relational model in 1970; D. D. Chamberlin and others from the IBM research lab from San Jose have developed the language that we now call SQL (Structured Query Language) [1]. There are many database management systems that have incorporated SQL and one of them is obviously the Microsoft SQL Server.

This work was performed through the Partnerships Program in priority areas, PN-II-PT-PCCA-2013-4-2225 - No. 170/2014 developed with the support of MEN - UEFISCDI, "Electromagnetic methods to improve processes wine".

Database management systems, such as Microsoft SQL Server, enjoy a high popularity precisely because they are easy to use and databases are easy to create. Microsoft SQL Server offers reliable transaction processing which is why so many choose this database management system.

Keeping the integrity of our data is often times the most important thing and Microsoft SQL Server has great support for it.

A. The XML Data Type

If the data is structured then our best choice for storing this data is the relational model. If the data is unstructured or semi-structured, we have several options. One would be using a NoSQL database and we will describe this possibility in the next chapter. Another option is using XML data type and this is particularly a good choice if unstructured data are tied in some way to structured data that is already stored in relational database. In this way, we will get a model that is independent from the platform and can be ported easily as shown in [2].

There are many reasons for choosing XML, some of the best, according to Microsoft, are shown in [2]:

- we don't have big quantities of data or the structure of our data is not known at the moment, we maybe have to take into consideration that our data structure might change in the future;
- we have recursive data or the entities don't have references among themselves;
- we have to follow a specific order in our data;
- we hardly ever need to update the whole entity at once, we want to update specific parts of it, change the structure or just simply query.

We have two options of storing XML data: either store it in SQL Server database and use its native XML features or choose to manage it in the file system. We choose considering some of the best reasons as shown in [2]:

- we need transactional integrity, so the most important reason would be that we need to share, query and modify the XML data in an efficient and transacted way;
- we want our relational data to work with or use parts of our XML data;
- we need support for querying and updating data, especially for a cross-domain application;
- we want indexing for an efficient way of querying the data that is stored in XML format.

For choosing to store our XML data in an SQL Server database, one has the option to store it in varchar(MAX), but as we want to take full advantage of what Microsoft SQL Server can offer us, we are going to talk about storing XML in the *xml data type*. We will also keep in mind that storing XML in the *xml data type* is slower due to the validation that happens in the background, but this can give the advantage of having all kinds of information about the specific order in the document, about attribute and element values.

In order to obtain the results from the experiments we used the hybrid model. The hybrid model is a combination of relational and *xml data type* columns [2]. The choice was made in order for the performance to be considerably better.

III. Unstructured Representation of Data in MongoDB

MongoDB is a document-oriented, NoSQL database. NoSQL, or Not Only SQL, is an approach of managing data and designing databases, which is most useful in the case that we have big quantities of data [3]. NoSQL databases provide you with ways of storing and retrieving the data that is not modelled as the relational databases are modelled. Mainly, NoSQL databases are designed to allow us insertion of data for which we do not have a predefined schema as the structure of our data is not set.

A database like MongoDB does not a have the concept of a "row"; instead, we have a more flexible model called a "document" [3]. The format in which the documents are stored is called BSON which comes from binary JSON and which offers us a binary representation of the JSON documents.

We have an easy way of modifying the structure of our data as MongoDB does not restrict to certain types or sizes, without having a predefined schema, we can experiment with modelling our data and choose the best option according to the needs of the application [3].

Often times the most challenging thing that developers are confronted with is the ever-growing amount of data that our applications deal with. As we need to store this data, the problem of scaling arises.

There are two choices when it comes to scaling: either we can scale up or we can scale out. Scaling up implies upgrading the machine we already have, basically adding more resources, while scaling out is getting our data spread across multiple machines [3]. Scaling up is generally more expensive and the physical limitation will inevitably be reached at some point [3]. Scaling out will come with a requirement of a bigger effort in order to administer the multiple machines, but it is generally less expensive and easier to scale [3].

When wanting to scale out a relational database we have to understand that it is generally not an easy problem to solve. However, MongoDB was made precisely with this process in mind. Being document oriented makes it easy to split the data and MongoDB figures out how to spread the data across the newly added machines [3].

The important thing to note about MongoDB is that while it has many features that facilitate CRUD operations, some features that we most often use in relational databases like joins, are not possible in MongoDB. We have a way of simulating this type of operation, which will be presented later on in this paper.

IV. Experimental Setup

Our working scenario is when we have data that cannot be normalized, but is still connected in some way to existing data in the relational database. We have a choice between using the hybrid model for SQL Server against storing the data in

MongoDB and just retrieving it from there. For the SQL Server, we will store an ID and an *xml data type* field. For MongoDB, the data will be stored as documents, which together will form a collection.

In order to have the fairest comparison we created indexes designed to ensure the optimal performance. As a result, for the *xml data type* we created the following indexes:

- primary XML index – this is the most important index that we created as this one indexes all the XML tags, values and paths [4]. According to Microsoft, for the creation of this index we need a clustered index on the primary key of the table that contains *the xml data type* column as SQL Server will use the primary key to correlate rows in the primary XML index with rows in our table [4];
- secondary XML index – in order to be able to create two types of secondary indexes we needed a primary XML index. These are the types of secondary indexes created:
    - path – used for queries that specify path expressions because it makes searching faster [5];
    - value – used for value based queries, an example would be searching for a string [5].
- full-text index on a XML column – according to Microsoft, it indexes the content of the XML values, but ignores the XML mark-up [6].

MongoDB has a default index on the *_id* field, so if now of creation we do not set it, this *_id* will be automatically set. Like the concept of primary key in SQL, this *_id* prevents the introduction of two *_id* values that are the same and is unique [7]. These are the following indexes created for the MongoDB database:

- single field – it is either an ascending or descending index specified by the user on single field of the document [7];
- compound index – it is an index on multiple fields from the document and the order in which you specify the fields is very important as MongoDB will sort after the first field and then it will sort within each value of the first field by the second field specified [7];
- text index – it is an index that supports running text search queries in a string content. One can specify any field that has a string as a value or an array of strings, according to MongoDB [8].

In order to run the experiments, we created a benchmark application using C# and Visual Studio 2013 as an IDE (Integrated Development Environment). Using the repository pattern, we created two repositories for each database. For the execution of the SQL commands, we used *SqlCommand* from *SqlClient* that is the .NET Framework Data Provider for SQL Server and for MongoDB we used the *.NET MongoDB Driver*. Both provide asynchronous workflows.

The architecture of the computer used to run the experiments:

Operating System	Windows 10 Pro
Processor	Intel(R) Core(TM) i5-4200M CPU @ 2.50 GHz
Installed memory (RAM)	4.00 GB
Disk	SSD Crucial MX100 256GB

V. Experimental Results

A. Experiment 1

The first experiment consists of populating the two databases with 100.000 entries. The chart shown in Figure 1 presents the results of the experiment:

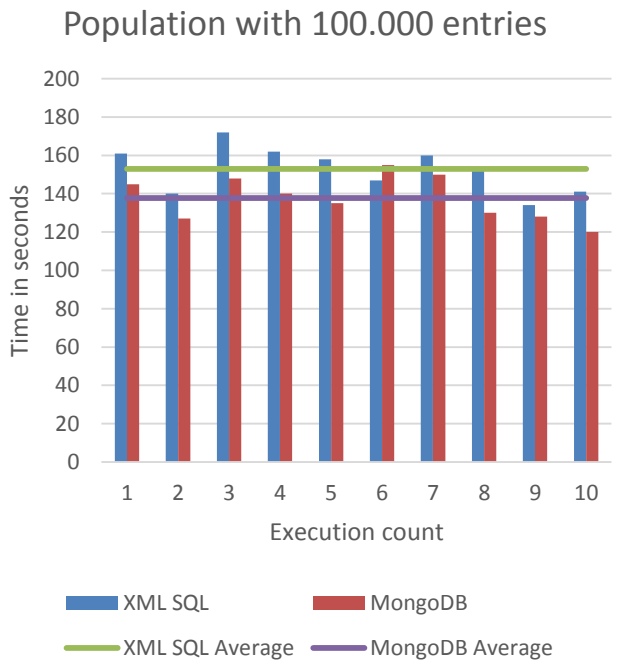


Fig. 1. The results of populating with 100.000 entries

We can easily see the implications of inserting an already considerable amount of data, both in SQL Server and in MongoDB. MongoDB is faster, usually being tens of seconds faster than SQL Server. The difference occurs also because of the XML validation done by SQL Server. The *xml data type* ensures us that each XML instance is correctly formed and this process slows down the insertion.

As it can be seen in Figure 1, MongoDB is faster than SQL Server in 9 out of 10 cases. The method used for insertion is similar in order to not give an advantage through implementation.

B. Experiment 2

The purpose of this experiment was to search by a randomly generated *ID* 1.000 times on each execution. As previously described the *ID* on the SQL Server database is a primary key and on the MongoDB database *_id* field has a default index on it.

The results of this experiment are pointing to the conclusion that searching by the *ID* field on which have a primary key or default index on it is yielding better results in SQL Server than in MongoDB. SQL Server is efficient and fast in these types of operations as it shown in Figure 2.

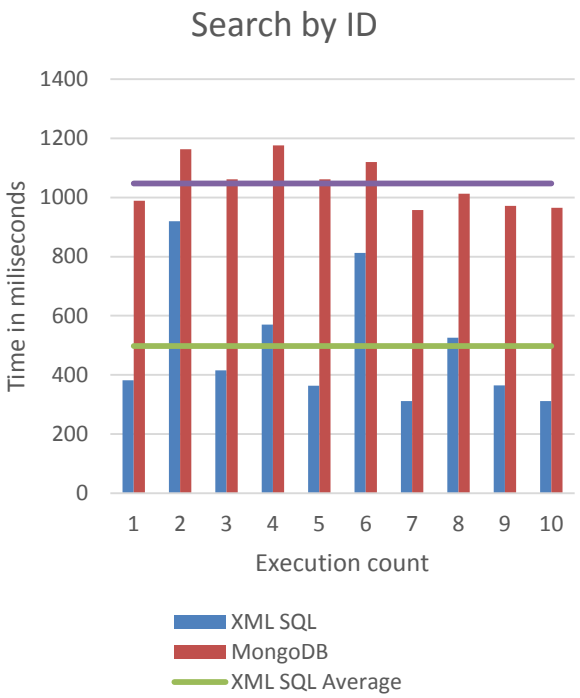


Fig. 2. The results of searching by ID

C. Experiment 3

The third experiment consisted in searching for a random string 1.000 times at every execution. In this experiment, we aim to test the full-text index and the text index that we set for each database type. The results, as shown in Figure 3, are rather dramatic as the difference between the two database types are quite big. MongoDB finds it simply easier and more efficient to search for particular occurrences of a string.

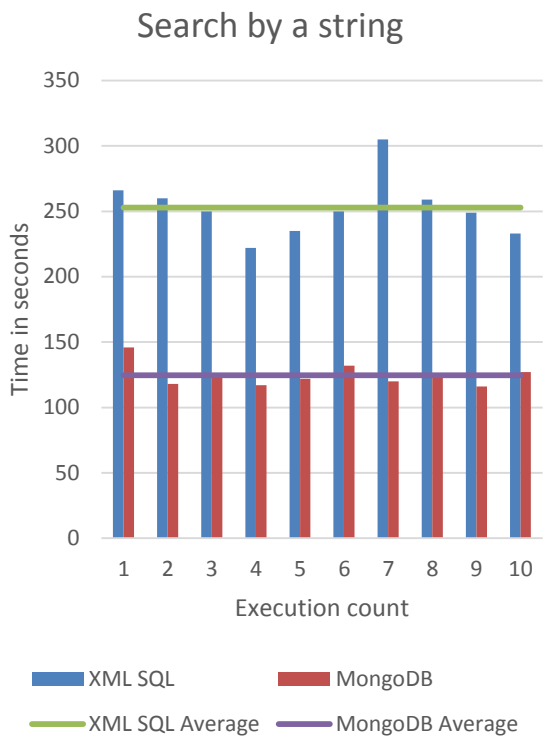


Fig. 3. The results of searching by string

D. Experiment 4

The purpose of this experiment is to update a field that has a randomly generated *ID* 1.000 times.

MongoDB clearly dominates in these types of operations, as we can see in Fig. 4, the difference is yet again major and in favour of our NoSQL database. MongoDB enables superior performance as querying in the XML using the *xml data type* methods, but is not nearly as fast as MongoDB's easy way of looking up the document by its ID and updating its field.

This is the method chosen to update a field in the SQL Server database:

```
UPDATE Post SET Xml.modify('replace value of
(/post/category/text())[1] with
(\"UpdatedCategory\")')
WHERE Id = @Id
```

Fig. 4. The results of updating a field

E. Experiment 5

The aim of this experiment is to compare the execution speed of join operations. MongoDB does not support join operations as it goes against the concept of data getting denormalized [9]. The addition of redundant data reduces the need for join operations. However, in certain scenarios we do not want to keep redundant data in our documents, and for this particular need MongoDB offers two solutions:

- manual references –meaning that we ought to have a field that will store the primary key of the document where the related data resides [9];
- *DBRefs* – this is a reference between two documents using the *_id* field, the name of the collection and optionally the name of the database [9].

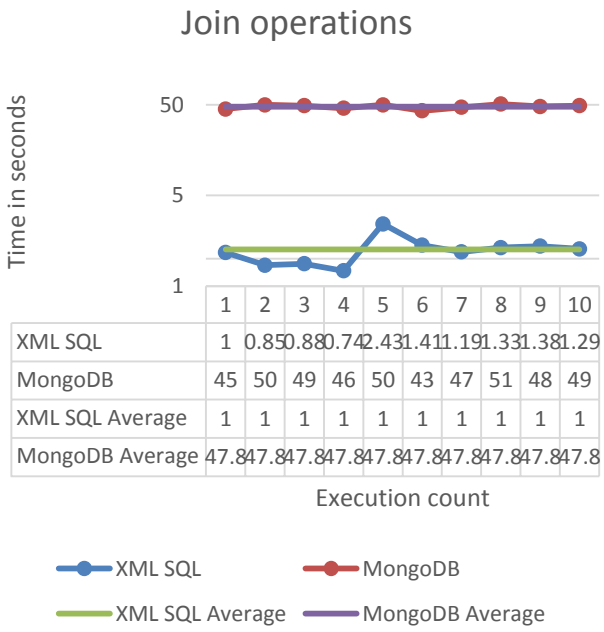
In this particular scenario, we know in which database the collection resides and we have all the information we need in our application meaning that *DBRef* does not give us any advantage over the manual reference. With that in mind, we chose to use the manual reference and to create an index on the reference field.

In the setup phase of our SQL Server database, we have created two tables – one that holds the posts, stores an *ID*, and has an *xml data type* field, which stores the XML and one table that holds the comments, which stores the *post_id* and some random content. We mirrored this in MongoDB by adding a comments document. It is necessary for us to create an index on the foreign key field *post_id* as this will speed up the operation drastically and will give fairness as we add an index on the *post_id* reference field from the MongoDB database as well.

We added, for each database, between 10 and 100 comments for 25,000 posts.

In our experiment, we made 1,000 join operations using the *ID* field which is, as previously mentioned, primary key in the SQL Server database and default *_id* index in MongoDB. In MongoDB's case, in order to simulate the joint operation we looked first for the post and then for all the comments made for that particular post.

As we can see in Figure 5, the results yielded by the SQL Server database are much better, which is as expected since with proper optimization there is no way that MongoDB can, beat SQL's JOIN. What is notable is that without having an index on the *post_id* foreign key field, SQL Server yielded much worse results than what we can observe here for MongoDB.



	1	2	3	4	5	6	7	8	9	10
XML SQL	1	0.85	0.88	0.74	2.43	1.41	1.19	1.33	1.38	1.29
MongoDB	45	50	49	46	50	43	47	51	48	49
XML SQL Average	1	1	1	1	1	1	1	1	1	1
MongoDB Average	47.8	47.8	47.8	47.8	47.8	47.8	47.8	47.8	47.8	47.8

Fig. 5. The results of join operations

VI. Conclusions

The purpose of all these experiments was to give an answer to the question: when do we use a relational database and when do we use a NoSQL database, like MongoDB?

The answer is not nearly as complex as one might think.

First, we need a proper analysis of the operations that we will do on our database and after that an analysis on the data that we work with. Microsoft SQL Server offers us transactional integrity and speed in JOIN operations, however MongoDB has the superior read and update speed. We must ask ourselves, "Do we have a rigid schema for our data?" Will the structure of our data suffer modifications? How flexible do we need to be when that happens? If our data cannot be normalized, we have to ask ourselves the question, "Does any

of the existing data in our relational database relate to our data that cannot be normalized?"

If yes, then is the hybrid model enough? Normalization often requires of us to store the data in many tables and in order for us not lose on performance we need many indexes. The same kind of structure can be modelled in a MongoDB database and in such a way that we completely get rid of the need to use JOIN type operations, which will drastically improve performance and will be considerably faster than any relational database. Identifying the needs of each application is key.

In this particular scenario, having data that cannot be normalized, it is very easy for us to conclude that given huge amount of data, MongoDB will always be the best solution. We can go as far as model our data in a single document, which will always be faster than storing XML in an *xml data type* column in an SQL Server database.

To sum up, while we cannot conclude that smaller amounts of data mean that the hybrid model becomes the best option, we can say that it entirely depends on the needs of the application that is being developed.

REFERENCES

[1] Matt Levene, George Loizou, "A Guided Tour of Relational Databases and Beyond" Published by Springer-Verlag, London, 1999, pp 1-2.

[2] XML Data Type and Columns (SQL Server) – Available: https://msdn.microsoft.com/en-us/library/hh403385.aspx., accessed January 2016.

[3] Kristina Chodorow, "MongoDB: The Definitive Guide, Second Edition" Published by O'Reilly, May 2013, pp 3-4.

[4] TechNet Library, "Primary XML Index" – Available: https://technet.microsoft.com/en-us/library/bb500237(v=sql.105).aspx, accessed January 2016.

[5] TechNet Library, "Secondary XML Index" – Available: https://technet.microsoft.com/en-us/library/bb522562(v=sql.105).aspx, accessed January 2016.

[6] TechNet Library, "Full-Text Index on an XML Column" – Available: https://technet.microsoft.com/en-us/library/bb522491(v=sql.105).aspx, accessed January 2016.

[7] MongoDB for Giant Ideas, "Index Introduction" https://docs.mongodb.org/manual/core/indexes-introduction/, accessed February 2016.

[8] MongoDB for Giant Ideas, "Text Indexes" – Available: https://docs.mongodb.org/manual/core/index-text/, accessed February 2016.

[9] MongoDB for Giant Ideas, "Database References" – Available: https://docs.mongodb.org/manual/reference/database-references, accessed February 2016.

[10] N. Jatana, S. Puri, M. Ahuja, I. Kathuria, D. Gosain, "A survey and comparison of relational and non-relational databases", International Journal of Engineering Research &Technology (IJERT) ISSN: 2278-0181, Vol 1, Issue 6, August 2012, pp. 1-5.

[11] R. D. Bulos, J. Bonsol, R. Diaz, A. Lazaro, V. Serra,"Comparative analysis of relational and non-relational database models for simple queries in a web-based application", Research Congress 2013, de la Salle University Manila, march 7-9, 2013.

[12] K. Sanobar, M. Vanita, "SQL Support over MongoDB using Metadata", *International Journal of Scientific and Research Publications*, Volume 3, Issue 10, October 2013.

[13] C. Győrödi, R. Győrödi, R. Sotoc, "A Comparative Study of Relational and Non-Relational Database Models in a Web- Based Application", International Journal of Advanced Computer Science and Applications, ISSN : 2158-107X(Print), ISSN : 2156-5570 (Online),Volume 6, Issue 11, 2015, pag. 78-83.

Improve Query Performance On Hierarchical Data. Adjacency List Model Vs. Nested Set Model

Cornelia Győrödi
Department of Computer Science and Information Technology, University of Oradea
Oradea, Romania

Robert Győrödi
Department of Computer Science and Information Technology, University of Oradea
Oradea, Romania

Romulus-Radu Moldovan-Duşe
Department of Computer Science and Information Technology, University of Oradea
Oradea, Romania

George Pecherle
Department of Computer Science and Information Technology, University of Oradea
Oradea, Romania

***Abstract*—Hierarchical data are found in a variety of database applications, including content management categories, forums, business organization charts, and product categories. In this paper, we will examine two models deal with hierarchical data in relational databases namely, adjacency list model and nested set model. We analysed these models by executing various operations and queries in a web-application for the management of categories, thus highlighting the results obtained during performance comparison tests. The purpose of this paper is to present the advantages and disadvantages of using an adjacency list model compared to nested set model in a relational database integrated into an application for the management of categories, which needs to manipulate a big amount of hierarchical data.**

Keywords—adjacency list model; nested set model; relational database; MSSQL 2014; hierarchical data

I. INTRODUCTION

Most of the database developers have dealt with hierarchical data in a relational database, and without a doubt, they reached the conclusion that relational database is not designed to manage data in a hierarchical way. Hierarchical data can be found in a great variety of database applications like the threads from forums or from emails, flowchart, content management categories or products categories [5].

Relational databases are widely used in most of the applications, and they have good performance when they handle a limited amount of data. To handle a huge volume of data like the internet, multimedia and social media the use of traditional relational databases is inefficient. Thus, more and more applications are beginning to use a non-relational database because they provide a more flexible structure that can shape after each user' needs; they are designed to store large amounts of data, and they have denormalized databases, which increases performance [6].

The tables from a relational database are not hierarchical. Hierarchical data have a parent-child relationship, which is not normally represented in a relational database table. A relational database does not store records in a hierarchical way. Hierarchical data is a collection of data where each item has a single parent and zero or more children, with the exception of the root item, which has no parent, as presented in [5].

These hierarchical data must also be stored in relational databases to obtain easier and more intuitive navigation through it [1] [8]. Because of these needs, there has always been an attempt to find solutions as close as possible to the application needs.

In this paper, we will examine two models dealing with hierarchical data in relational database namely adjacency list model and nested set model.

We will start with adjacency list model, and we will consider an example of the hierarchy of categories from an ads web-application as shown in Fig. 1.

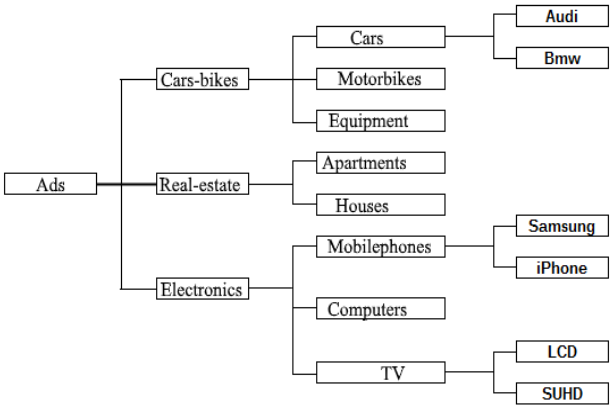


Fig. 1. Hierarchical data structure

The adjacency list model is very easy to maintain but is less efficient for queries. A hierarchical query is a method of reporting the branches of a tree in a specific order.

If we need a better performance from queries, and the hierarchical data does not have frequent changes in time, nested set model was proved by tests conducted by us to be a more efficient model.

In the next sections of this paper, we will describe the adjacency list model compared to the nested set model, query algorithm in adjacency list model compared to nested set model, study and analyse performance and the final conclusions.

II. Description of the Nested Set Model Compared to the Adjacency List Model in Relational Databases

To be able to understand why one is better than the other models in certain situations, we must first describe the two models.

The adjacency list model is probably the most common type of hierarchical structure found in databases and is characterized by nodes and lines. The reason for the popularity of this model is that it is easy to understand and maintain. The connection between nodes is made via an attribute called parent, that is the head of the hierarchy and has the parent attribute set to NULL and the rest of the elements have values corresponding to their parent's ids. Based on the previous section example, in Fig. 2 shows the hierarchical data structure as an adjacency list model and each colour represents a level in the hierarchy:

id	name	parentid
1	Ads	NULL
2	Car-bikes	1
3	Real-estate	1
4	Electronics	1
5	Cars	2
6	Motorbikes	2
7	Equipment	2
8	Apartments	3
9	Houses	3
10	Mobilephones	4
11	Computers	4
12	TV	4
13	Audi	5
14	BMW	5
15	Samsung	10
16	iPhone	10
17	LCD	12
18	SUHD	12

Fig. 2. Adjacency List Model exemplified on a four level hierarchical structure

Unlike the adjacency list models, in nested set models we will look at hierarchy in a different way. Each node in hierarchy will contain all the children from the nodes that are subordinated directly or indirectly to it. The nodes in the hierarchy will have two attributes, which we call right and left in which numbers will be stored to help us to identify all children of a node. The numbering technique was proposed by Joe Celko in [2], by starting from the element from the top of the hierarchy, all elements will be numbered two times, saving each value in the left, and right attributes, as you can see in the example from Fig. 3.

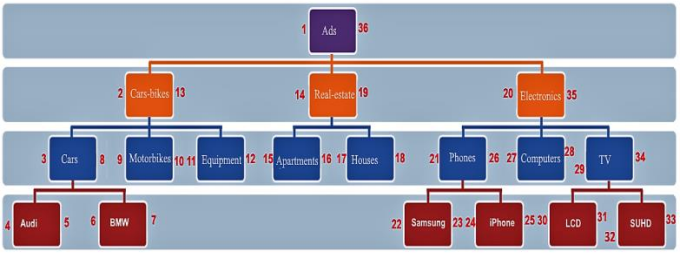


Fig. 3. Nested set model in a four level hierarchical structure

A. Adding and deleting of nodes in a nested set model

The values of the left and right attributes of the nested set model must be recalculated when a new node is added or deleted in the hierarchy. When adding a new node in the hierarchy the left and right attributes from the parent node that will contain the new element must always be taken into consideration [3]. If we want to add a new category called *Opel* in the hierarchy in Fig. 3, which belongs to the category *Cars*, according to the numbering technique, in the new node the value 8 will be stored in the left attribute and the value 9 in the right attribute. Therefore, all left, and right attributes with values greater than or equal to the value of the right attribute of the new category *Cars* will be incremented by 2, as shown in Figure 4.

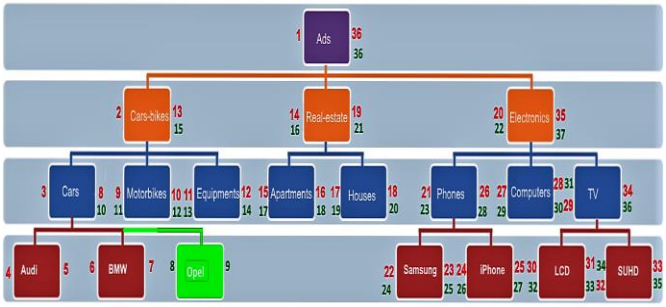


Fig. 4. Adding a subcategory in the nested set model

When we delete a node in the hierarchy, we must take into consideration the left and right attribute of the deleted node. Suppose we want to remove the *Real-estate* category in the hierarchy in Fig. 3. When deleting a category we must also take in consideration its subcategories, because they will also be deleted from the hierarchy when the category that they belong to is deleted [3]. An exact identification of all the subcategories that belong to the *Real-estate* category will be made based on the left and right attribute. We can see that all the subcategories from *Real-estate* have the left and the right attributes between 14 and 19. After we deleted the whole *Real-estate* branch, all the categories that have the value of the left and right attributes greater than 19, will be decremented with the difference between the right attribute and the left of the deleted category plus the value 1, in our case this will be 19-14+1 as you can see in the Fig. 5.

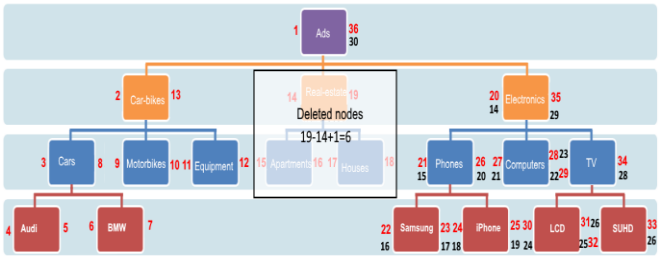


Fig. 5. Deleting a subcategory in the nested set model

III. Queries in the Adjacency List Model Vs. the Nested Set Model

A. Getting a subtree in the hierarchical data structure

We consider hierarchical structure presented above to exemplify queries. One of the most common queries, in this case, is to get all the subcategories that belong to a specific category.

If we want to get the subcategories that belong to the *Electronics* category, and we are considering the adjacency list model we have two options. First we could use self-join, in this case we must know the number of levels in the structure of categories and make one self-join for each level to get the categories from the lower level. For the second option we could use common table expression to build a recursive query which will return all the sublevels of category without needing to know the number of the levels existing.

The self-join query allows us to see the full path through our hierarchical data structure as shown below and in [5]:

```
SELECT t1.name AS nivel1, t2.name as nivel2,
t3.name as nivel3, t4.name as nivel4
FROM category AS t1
LEFT JOIN category AS t2 ON t2.ParentId = t1.Id
LEFT JOIN category AS t3 ON t3.ParentId = t2.Id
LEFT JOIN category AS t4 ON t4.ParentId = t3.Id
WHERE t1.name = 'Electronics';
```

```
SQL Server Execution Times:
   CPU time = 16 ms, elapsed time = 28 ms.
```

```
Table 'Workfile'. Scan count 0, logical reads 0,
physical reads 0, read-ahead reads 0, lob logical reads 0,
lob physical reads 0, lob read-ahead reads 0.
Table 'Worktable'. Scan count 0, logical reads 0,
physical reads 0, read-ahead reads 0, lob logical reads 0,
lob physical reads 0, lob read-ahead reads 0.
Table 'Category'. Scan count 3, logical reads 59,
physical reads 0, read-ahead reads 0, lob logical reads 0,
lob physical reads 0, lob read-ahead reads 0.
```

The main limitation of such an approach is that you need one self-join for every level in the hierarchy, and performance will naturally degrade with each level added as the joining grows in complexity, as presented in [5].

We use a recursive query to obtain all the sublevels of the category without needing to know the number of the existing levels.

Recursive query:

```
with CategoryBranch as (
     select Id, Name,ParentId
     from Category
     where name = 'Electronics'
union all
     select c.Id, c.Name, c.ParentId
     from Category c
     join CategoryBranch p on c.ParentId = p.Id
)select Name from CategoryBranch order by
ParentId;
```

```
SQL Server Execution Times:
  CPU time = 94 ms,  elapsed time = 89 ms.
```

```
Table 'Category'. Scan count 1, logical reads
2151, physical reads 0, read-ahead reads 0, lob
logical reads 0, lob physical reads 0, lob read-
ahead reads 0.
Table 'Worktable'. Scan count 2, logical reads
679, physical reads 0, read-ahead reads 0, lob
logical reads 0, lob physical reads 0, lob read-
ahead reads 0.
```

In the nested set model, the hierarchical data is maintained, as parent categories contain all the children from the nodes that are subordinated directly or indirectly to it. We can represent this form of hierarchical data in a table through the use of *TreeLeft* and *TreeRight* values.

```
CREATE TABLE NestedCategory (
     category_id INT AUTO_INCREMENT PRIMARY KEY,
     name VARCHAR(20) NOT NULL,
     TreeLeft INT NOT NULL,
     TreeReft rgt INT NOT NULL
);
```

We can see the full path of our hierarchical data using of a self-join that connects parents with nodes based on the fact that a node's *TreeLeft* value will always appear between its parent's left and right values as also shown in [5]:

```
SELECT node.name FROM NestedCategory AS node,
NestedCategory AS parent
WHERE node.TreeLeft BETWEEN parent.TreeLeft AND
parent.TreeRight
AND parent.name = 'Electronics'
ORDER BY node.TreeLeft;
```

```
SQL Server Execution Times:
  CPU time = 0 ms,  elapsed time = 2 ms.
```

```
Table 'NestedCategory'. Scan count 1, logical reads
233, physical reads 0, read-ahead reads 0, lob logical
reads 0, lob physical reads 0, lob read-ahead reads 0.
```

We can see from the performed tests and displayed parameters that query execution time for this type of query is the best when using the nested set model.

B. Finding all the leaf nodes

This type of query refers to obtaining all the categories from the tree that do not contain other subcategories. Thus we need all the parents who do not have children.

In the case of the adjacency list models, we need a self-join to get parent categories that have no children [5].

```
SELECT parent.name FROM category AS parent
LEFT JOIN category as child ON parent.Id =
child.ParentId
WHERE child.Id IS NULL;
```

```
SQL Server Execution Times:
CPU time = 16 ms,  elapsed time = 387 ms.
```

```
Table 'Category'. Scan count 2, logical reads 38,
physical reads 0, read-ahead reads 0, lob logical
```

```
reads 0, lob physical reads 0, lob read-ahead reads 0.
```

For nested set model, this type of query is simpler, because it is based on a rule from this model which says that leaf node has the left and right attributes with consecutive values. Thus we have to look at the categories where left + 1 = right [5].

```
SELECT name
FROM NestedCategory
WHERE TreeRight = TreeLeft + 1;

SQL Server Execution Times:
   CPU time = 16 ms, elapsed time = 371 ms.
```

```
Table 'NestedCategory'. Scan count 1, logical reads 22, physical reads 0, read-ahead reads 0, lob logical reads 0, lob physical reads 0, lob read-ahead reads 0.
```

The execution time is insignificant distinct between those two models but the execution plan is clearly better for the nested set model, as shown in Fig. 6.

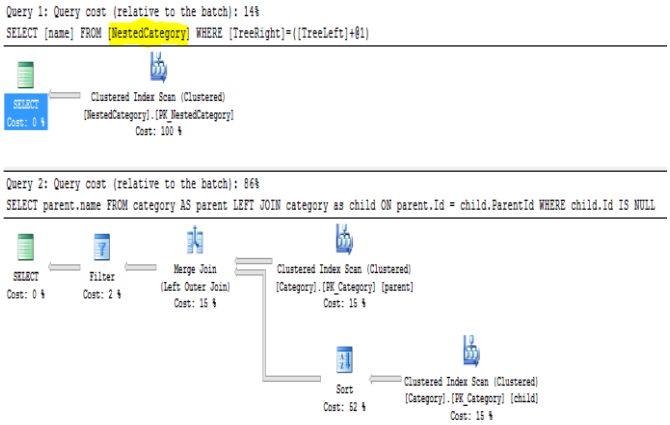


Fig. 6. Execution plan for the nested set model vs. adjacency list model

There are situations where working with adjacency list model directly in SQL can be difficult, such as the cases where we necessarily need self-joins and the exact number of existing levels in the tree or the level of the node to which we refer. In such a situation, it would be necessary to calculate the level on that is each node in the tree.

In the case of nested set model, to calculate the level of each category from the tree, we will be doing COUNT function on the parent nodes of each category based on the same rule that says any subcategory will have the left attribute value between the left and right values of the parent category.

```
SELECT node.name, COUNT(parent.name) AS CategoryLevel
FROM NestedCategory AS node,
NestedCategory AS parent
WHERE node.TreeLeft BETWEEN parent.TreeLeft AND parent.TreeRight
GROUP BY node.name
ORDER BY CategoryLevel;
```

A special case occurs when on the hierarchical data structure, we have assigned data from another table, and we want to know how they are distributed on each node in the structure. Supposing that we have a relational table with ads, and each ad belongs to a "leaf" category, if we want to display a list of all the categories and the number of ads each category, for the parent node categories we will consider the number of ads from each child subcategory that belongs to it.

In nested set model, as usual, we will use all of the attributes of the left and right to get the parent-child relationship between the categories and a join operation between the child table and table with articles to refer to articles assigned to a category and add a COUNT function in each category.

```
SELECT parent.name, COUNT(Articles.Id)
FROM NestedCategory AS child ,
        NestedCategory AS parent,
        Articles
WHERE child.TreeLeft BETWEEN parent.TreeLeft AND parent.TreeRight
        AND child.Id = Articles.NestedCategoryId
GROUP BY parent.name
ORDER BY parent.name;
```

IV. Implementation and Performance Analysis

The application we developed using Microsoft SQL Server 2014 Management Studio [7] has 4 sections. The first section we exemplified the navigation through the categories and the listing of ads from a selected category. In the second and the third sections we implemented the management of categories using the two models described above and in a fourth section we displayed response times to queries on ads comparing the performance of the two hierarchical data models.

Because the test results depend on the computer on which these tests are carried out, it is important to note that all the results presented below were obtained from studies conducted on a computer with the following characteristics: Windows 10 Home Edition 64-bit, processor Intel Core i5 (2.4 GHz), 4 GB RAM memory. The database contains 2902 categories and 310 818 ads distributed by category.

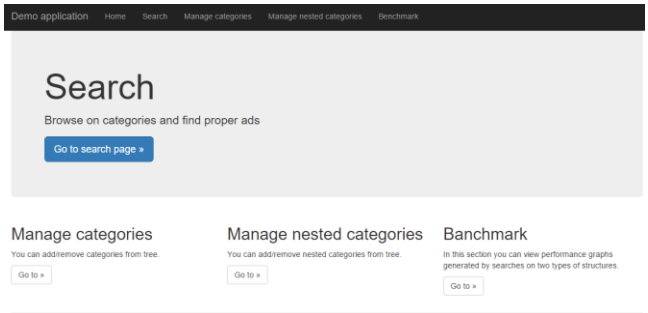


Fig. 7. Main page of the application

A. Navigation section

For this section of the application, we used two queries that highlight the usefulness of nested set models. For both queries, we used stored procedures, *sp_GetNestedCategoryIncludingCountArticles* and *sp_GetArticlesByNestedCategoryId*.

The first procedure, named *sp_GetNestedCategoryIncludingCountArticles* is used in the navigation menu on the left, where besides displaying the hierarchical data structure of categories is also shows the number of ads in each category. Thus,

sp_*GetNestedCategoryIncludingCountArticles* procedure returns a structure of categories with the number of ads in each category, as shown in Fig 8.

```
   CREATE PROCEDURE
[dbo].[sp_GetNestedCategoryIncludingCountArticles]
        @ParentId int
   AS
   BEGIN
        SET NOCOUNT ON;

        SELECT parent.Id, parent.Name,
parent.ParentId, parent.TreeLeft, parent.TreeRight,
COUNT(Articles.Id) as CountArticles
   FROM NestedCategory AS node,
           NestedCategory AS parent,
           Articles
   WHERE node.TreeLeft BETWEEN parent.TreeLeft AND
parent.TreeRight AND node.Id =
Articles.NestedCategoryId AND parent.ParentId =
@ParentId
   GROUP BY parent.Id, parent.Name, parent.ParentId,
parent.TreeLeft, parent.TreeRight
   ORDER BY parent.Id;
   END
```

The second procedure, entitled *sp_GetArticlesByNestedCategoryId* is used to return from the database a list of ads for the selected category including subordinate categories. Thus, *sp_GetArticlesByNestedCategoryId* procedure returns ads that belong to certain categories respectively subcategories within the given category.

```
   CREATE PROCEDURE
[dbo].[sp_GetArticlesByNestedCategoryId]
        @NestedCategoryId int,
        @FromArticleId int = null
   AS
   BEGIN
        SET NOCOUNT ON;

           IF(@FromArticleId is null) SET
        @FromArticleId = 0
        DECLARE @TreeLeft int
        DECLARE @TreeRight int

   select @TreeLeft=TreeLeft, @TreeRight=TreeRight
   from NestedCategory where id = @NestedCategoryId

        select top 10
   Id, CategoryId, NestedCategoryId, Title, Body
        from Articles
           where Id > @FromArticleId and
        NestedCategoryId in    (select id from
        NestedCategory where TreeLeft >= @TreeLeft
        and TreeRight <= @TreeRight)
        ORDER BY Id
   END
```

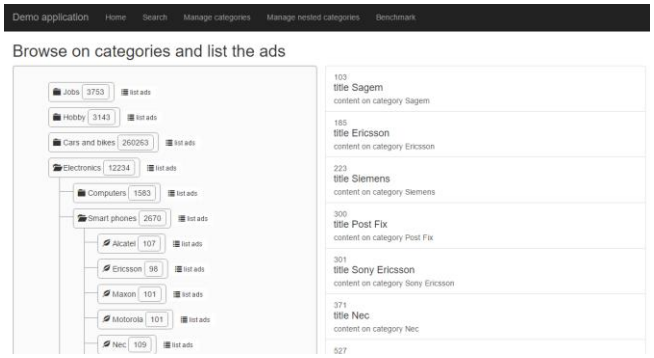


Fig. 8. Browse on categories and list the ads

B. *Categories management section*

In this section of the application shown in Fig. 9, we can add or delete categories from the two hierarchical data structures.

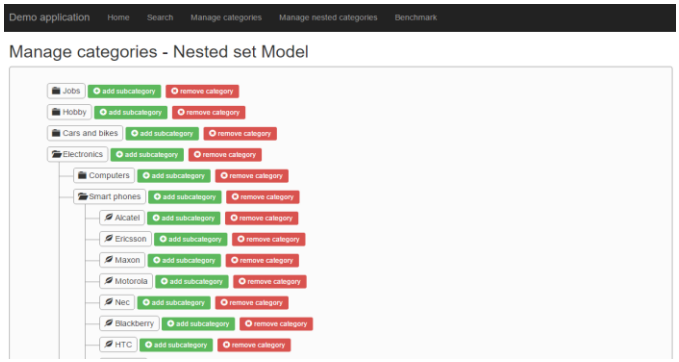


Fig. 9. Categories management page

The management interface is identical between the two hierarchical models, but at level database specific stored procedures are called for each model. In the adjacency list model the procedure that makes adding a new category has as input parameters the parent category name and the name of the new category, identify *ParentId* in the category table and then insert the new category. For the deleting a category in the adjacency list model, we used a stored procedure that has as input parameters the category name to identify the first *Id* of category and then delete the category and its children (all the categories which have *ParentId* identical with *Id* of category).

The procedure, entitled *sp_InsertNestedCategory* is used in the nested set model to add a category in the hierarchical data structure. The procedure has as input parameters the parent category name and the name of the new category, identify *ParentId* in the table, then left and right nodes updated with new values and then insert the new category.

```
CREATE PROCEDURE [dbo].[sp_InsertNestedCategory]
    @ParentName nvarchar(50),
    @Name nvarchar(50)
AS
BEGIN
    SET NOCOUNT ON;

    DECLARE @ParentId int
       SET @ParentId = (select Id from
    NestedCategory where Name = @ParentName)

       if (@ParentId is null or @Name is null or
    exists(select * from NestedCategory where
    Name = @Name))
     return;

    DECLARE @parentTreeLeft INT
    DECLARE @parentTreeRight INT
    DECLARE @countChilds int

       SET @parentTreeLeft = (SELECT TreeLeft
    FROM NestedCategory WHERE id = @ParentId)
       SET @parentTreeRight = (SELECT TreeRight
    FROM NestedCategory WHERE id = @ParentId)
       SET @countChilds = (select 2*count(*)
    from NestedCategory where TreeLeft >
    @parentTreeLeft and TreeRight <
    @parentTreeRight)

    BEGIN TRAN
    UPDATE NestedCategory
        SET TreeLeft = CASE WHEN TreeLeft >
    @parentTreeRight
                  THEN TreeLeft + 2
                  ELSE TreeLeft END,
           TreeRight = CASE WHEN TreeRight >=
        @parentTreeRight
                  THEN TreeRight + 2
                  ELSE TreeRight END
     WHERE TreeRight >= @parentTreeRight
       INSERT INTO NestedCategory(TreeLeft,
    TreeRight, ParentId, Name)
    VALUES(@parentTreeLeft + @countChilds + 1,
    @parentTreeLeft + @countChilds + 2,
    @ParentId, @Name);
    IF @@ERROR != 0
           ROLLBACK TRAN
    ELSE
           COMMIT TRAN
END
```

For the deleting a category in the nested set model we used a stored procedure entitled *sp_DeleteNestedCategory*, that has as input parameters category name to identify first *Id* of category, then the category will be deleted along with related subcategories, and left and right nodes updated by difference between *TreeRight* and *TreeLeft* + 1 of removed category.

```
CREATE PROCEDURE [dbo].[sp_DeleteNestedCategory]
    @Name nvarchar(50)
AS
BEGIN
    SET NOCOUNT ON;

    DECLARE @Id int
       SET @Id = (select id from NestedCategory
    where Name = @Name)
    if (@Id is null) return;

    DECLARE @treeLeft INT
    DECLARE @treeRight INT
    DECLARE @parentId INT
    DECLARE @treeWidth INT

       SET @treeLeft = (SELECT TreeLeft FROM
    NestedCategory WHERE id = @Id)
       SET @treeRight = (SELECT TreeRight FROM
    NestedCategory WHERE id = @Id)
    SET @treeWidth = @treeRight - @treeLeft + 1
       SET @parentId = (SELECT ParentId FROM
    NestedCategory WHERE id = @Id)

    BEGIN TRAN
       UPDATE Articles SET NestedCategoryId =
    @parentId WHERE CategoryId in (select id
    from NestedCategory where TreeLeft >=
    @treeLeft and TreeRight <= @treeRight)
       DELETE FROM NestedCategory where TreeLeft
    between @treeLeft and @treeRight
       UPDATE NestedCategory SET TreeLeft =
    TreeLeft - @treeWidth WHERE TreeLeft >
    @treeRight
       UPDATE NestedCategory SET TreeRight =
    TreeRight - @treeWidth WHERE TreeRight >
    @treeRight

    IF @@ERROR != 0
           ROLLBACK TRAN
    ELSE
           COMMIT TRAN
END
```

C. Performance analysis of the queries

For this analysis, we considered the stored procedures used for the ads listing from a specified category for the two models studied. The two stored procedures *sp_GetArticlesByCategoryId* and *sp_GetArticlesByNestedCategoryId*, are described below. Each of the two procedures returns ten ads that are from the specified category.

```
   CREATE PROCEDURE
[dbo].[sp_GetArticlesByCategoryId]

   @CategoryId int,
   @FromArticleId int = null
   AS
   BEGIN
        SET NOCOUNT ON;
        IF(@FromArticleId is null) SET
@FromArticleId = 0;

        with CategoryBranch as (
        select Id
        from Category
        where Id = @CategoryId
   union all
        select c.Id
        from Category c
        join CategoryBranch p on c.ParentId = p.Id
   )
   select top 10
      Id, CategoryId, NestedCategoryId, Title, Body
   from Articles
   where Id > @FromArticleId and CategoryId in
         (select Id from CategoryBranch)
   ORDER BY Id;
   END

   CREATE PROCEDURE
[dbo].[sp_GetArticlesByNestedCategoryId]

        @NestedCategoryId int,
        @FromArticleId int = null
   AS
   BEGIN
   SET NOCOUNT ON;
```

```
    IF(@FromArticleId is null) SET @FromArticleId = 0
    DECLARE @TreeLeft int
    DECLARE @TreeRight int

        select                    @TreeLeft=TreeLeft,
      @TreeRight=TreeRight   from       NestedCategory
      where id = @NestedCategoryId

        select top 10
        Id,   CategoryId,   NestedCategoryId,   Title,
Body
        from Articles
        where     Id     >     @FromArticleId     and
NestedCategoryId  in(select  id  from  NestedCategory
where  TreeLeft  >=  @TreeLeft  and  TreeRight  <=
@TreeRight)
        ORDER BY Id
    END
```

In the graphs from Fig 10 and Fig. 11, we can see the response time for the two stored procedures from runs with one iteration to runs with ten iterations. We noticed that the response time is affected by the total number of the ads from the sub-branch on which the search is made, but always the best execution time is obtained by the nested set model as shown in Fig 11.

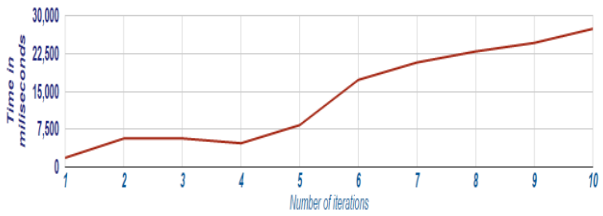


Fig. 10. Query performance for the adjacency list model

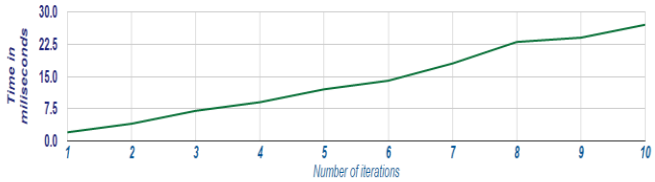


Fig. 11. Query performance for the nested set model

We can see that if we grow the number of iterations the difference of time is higher between the two models and the nested model has the best performance between the two.

V. Conclusions

When we work with hierarchical data structures with more than 2 levels, and the number of levels varies from one branch to another of the hierarchy, then it is better to store the hierarchical data as a nested set model in the database. In the nested set model is more difficult to do the adding, moving and deleting of nodes because we need to update every time the value of the left and right attributes to keep the integrity of the hierarchy. However, the advantage is pretty big because the number of the queries on the relational table is the same no matter the number of hierarchy levels from the nested set model, on the other hand, the number of queries for the adjacency list model is equal to the number of levels of hierarchy.

In case the hierarchical structure is large, it is suggested to break it into smaller hierarchical structures that are to be stored in separate tables, thus allowing a better administration of each hierarchical structure.

As an extension of our study, we would like to compare the performance of the nested set model with the hierarchical data model implemented in Microsoft SQL Server 2012.

References

[1] Joe Celko “Trees and Hierarchies in SQL for Smarties”, 2nd Edition Morgan-Kaufmann, 2012, ISBN 978-0-12-387733-8

[2] Joe Celko, “Trees in SQL”, Available: http://www.ibase.ru/devinfo/DBMSTrees/sqltrees.html (dec. 2015)

[3] T. Stryja, “Nested set model practical examples, part I”, Available: http://we-rc.com/blog/2015/07/19/nested-set-model-practical-examples-part-i (nov. 2015)

[4] T. Stryja, “Nested set model practical examples, part II”, Available: http://we-rc.com/blog/2015/07/19/nested-set-model-practical-examples-part-ii (nov. 2015)

[5] Mike Hillyer, “Managing Hierarchical Data in MySQL”, Available: http://mikehillyer.com/articles/managing-hierarchical-data-in-mysql/ (dec. 2015)

[6] Cornelia Győrödi, Robert Győrödi, George Pecherle, Andrada Olah, “A comparative study: MongoDB vs. MySQL”, IEEE - 13th International Conference on Engineering of Modern Electric Systems (EMES), 2015, Oradea, Romania, 11-12 June 2015, ISBN 978-1-4799-7649-2, pag. 1-6.

[7] Microsoft SQL Server 2014 Management Studio, Available: https://www.microsoft.com/en-us/download/details.aspx?id=42299 (dec 2015)

[8] Joe Celko's “SQL for Smarties”, 5th Edition, Morgan-Kaufmann, 2014, ISBN 9780128008300.

Content-Based Image Retrieval for Medical Applications with Flip-Invariant Consideration Using Low-Level Image Descriptors

Qusai Q. Abuein
Department of Computer Information Systems
Jordan University of Science and Technology

Radwan Batiha
Department of Computer Information Systems
Irbid National University

Mohammed Q. Shatnawi*
Department of Computer Information Systems
Jordan University of Science and Technology

Ahmad Al-Aiad and Suzan Amareen
Jordan University of Science and Technology

***Abstract*—Content-Based Image Retrieval (CBIR) has recently become one of the most attractive things in medical image research. Most existing research on CBIR focused on low-level image descriptors in image retrieval, which sometimes weaken the retrieval accuracy since many relevant images are not retrieved. Limited number of researches consider flipping the image on different sides. In order to fill the knowledge gap, this research focuses on considering the flipped images in retrieval based on previously implemented system that uses low-level image descriptors.**

Some improvements are made on the system considering the flipped images by extracting the features of the main and flipped images. The final results showed that the proposed system outperforms the existing system. The system has proven as a powerful method in helping medical staff, physician decision makers, and students to get better results by giving wide range of needed images, and helps in reasoning and building better decisions.

***Keywords*—CBIR; image retrieval; feature extraction; medical images; flipped images**

I. Introduction

The old-fashioned way of retrieving images is known as textual retrieval. This retrieval approach is basically applied by assigning specific keywords to each image in the dataset, and then retrieving the needed image or set of images according to some textual query. It is a time-consuming approach when a large number of images are searched for, and lacks accuracy due to the vocabulary mismatch since the image description is subjective.

Flipping image during the search process has been ignored by previous researchers especially in the medical field. In particular, none of the previous research in the medical field has applied image flipping in the search retrieval. Therefore, this research addresses this significant knowledge gap by enhancing the search algorithm of existing system to consider flipping images as part of the search process.

The new approach in retrieving images is Content Based Image Retrieval (CBIR). CBIR means that images can be searched for depending on nothing but their visual content. The visual content of the image is defined as the graph, the text, the image color [1], the local and global features [2], or any other content inside the image. CBIR in other words, is mainly describing the images based on their content [3]. In CBIR, the content of the image is represented as numeric measurement [4].

CBIR can be used in searching for and retrieving images from large database collections. To retrieve these images the images must go through indexing and feature extraction methods [3]. CBIR has many methods for analyzing images; each method represents different aspects of the visual information of an image. Image searching and archival can greatly reduce the time that is consumed by using automatic image analysis tools [1].

II. Related Work

A. CBIR for Medical Applications using Low-level Image Descriptors

A system that uses three low-level descriptors has been implemented by the authors of [5]. The system is objected to retrieve images based on their content (CBIR) for medical applications.

The system has already defined medical dataset, which contains many medical images from variant fields. The features of all images in the dataset are extracted using low-level image descriptors and the values of the features are stored in one main index file. The index of all images is consequently compared against each other. In this way, the related images will have the least distance between each other and will be retrieved together. [5]

The used features are CLD, EHD, and CEDD. The combination of these three descriptors has shown powerful results in Image retrieving, since they present color, texture, and shape, respectively.. Having such a comprehensive feature extraction system has improved the retrieval results.

B. An Efficient Iconic Indexing Strategy for Image Rotation and Reflection in Image Databases

Identify applicable sponsor/s here. If no sponsors, delete this text box (*sponsors*).

In their effort, the authors of [4] developed a new algorithm in CBIR that considers flipped and rotated images. They not only consider the flipped and rotated images, but also the flipped and rotated objects inside the images.

The images can be classified and distinguished in two ways: the first is the low level features color, texture, and shape features, the other includes the spatial relationships between objects which is known as Retrieval by Spatial Similarity (RSS).

CBIR retrieves the images according to the content, for example, an entered query by a user can be: "show all the images that contain a circle to the right of a rectangle".

The database indices stored in the databases do not consider the flipped and rotated images, it only gives the images that satisfy the condition, where the main database may contain many images that meet the users need, these images are flipped or rotated. Therefore, an algorithm that gives all the possibilities for any index has been proposed in this research [6].

The basic idea of this research is to use symbolic projection category [8].

To extract all possible indices for an image, several transformations are made in advance. These transformations include: the original image, the image flipped horizontally, flipped vertically, rotated by 90 degrees, rotated by 180 degrees, and finally rotated by 270 degrees. Next, all these indices are stored within the dataset.

To derive all the possible indices from the original image directly, an efficient iconic indexing strategy has been presented in this research. This is done by a unique bit pattern matrix (UBP matrix). In this way, the proposed strategy will not miss the qualified images in the main dataset when the query is issued in the different orientation ways. Using rules of transformations as shown in Table I, the matrix is derived and so all the indices are generated.

TABLE I. RULES OF TRANSFORMATION: X' AND Y' ARE THE TRANSFORMED OPERATORS TO X AND Y

Functions	**Operator**	
	x-axis	***y-axis***
	X	***Y***
Rotate 90°	Y	X'
Rotate 180°	X'	Y'
Rotate 270°	Y'	X
Flip Horizontally	X'	Y
Flip Vertically	X	Y'

Results have shown a 50% improvement compared to the traditional ways in searching. The more objects an image has, the better the results retrieved. Unfortunately, this approach is not applied on the medical images.

C. A rotation- and Flip-invariant algorithm for representing spatial continuity information of geographic Images in content-based Image Retrieval

This research proposes a rotation- and flip-invariant algorithm. The used images in this research are the high-resolution geographic images. This algorithm is applied by representing spatial continuity information in these images. This means CBIR is applied on the geographic images. [4]

The algorithm has three main steps. The first step is to start with variogram concept: one viogaram is taken as a sample and the basic shape is captured for it. The second step is to represent the spatial continuity anisotropy for the sample shape. The final step is to extract the rotation- and flip invariant as an output. This output is the new visual property which is represented in the form of a numeric index vector.

This vector consists of a set of semi-variances at selected lags and directions. By reordering these semi-variances the original images, flipped, and rotated can be extracted. The algorithm goes through a test to confirm if the reordering can align all the image representations.

Another test for the algorithm is measuring the retrieval precision. Seven types of typical geographic entities are retrieved from an Erie County ortho-photo database, and the precision is calculated to test the effectiveness of this algorithm [4].

D. Automatic Semantic Indexing of Medical Images

This paper uses a grid technology to medical CBIR. The technology is used to close the gap between monolithic CBIR systems for general image retrieval purpose and the programming tools. The programming tools are used to support the development of image processing algorithms and the automatic distributed execution of them. The Grid Concept (GC) provides resource sharing. Resource sharing incorporates data sharing, access to computers, access to software. GC uses processing techniques that extract image content information into MPEG-7 format automatically, and associate them to the existing domain ontology's [7].

Several works have been done before this work, Tsechpenakisetal [ ] built a system that automatically extracts images semantics by detecting and tracking moving objects in video sequences. The domain in his work is not medical field [9].

After that, Sjo¨bergetal proposed a method of content-based multimedia retrieval of objects with visual and textual properties. Objects that belong to a specific semantic class are associated with their low-level descriptors and textual features. For example, frequencies of significant keywords are extracted from audio tracks. The user provides the system with a set of sample objects that tells the system about the object he or she is looking for. The user selects the object samples from an existing database. The results showed that the retrieval performance has increased when the textual features are used. The results also showed that audio features perform very well. The domain here is not medical as well [10].

Perner [11].has developed architecture for image mining and learning semantic tagging rules. This architecture has a

domain of specific vocabulary obtained from a domain expert or given by the domain and does not use MPEG-7 features. This architecture aims to develop the right feature extraction procedures for describing the high level semantic terms

In the developed system, the process of image segmentation, feature extraction and annotation are performed on an image by image basis. An image is selected from the same category as the ontology term and uploaded to the Visual Descriptor Extraction (VDE) tool which is developed as a plug-in to automate annotizer and presents a graphical user interface for loading and processing visual content, extraction of visual features, and association with domain ontology concepts. This technique uses pen or mouse to select the region on the image corresponding to the ontological term. Once the region of interest is selected, all image features are extracted one by one using the VDE tool. The descriptors are available as XML files in MPEG-7 standard format. The ontology term is represented as prototype instance and the links to corresponding image feature descriptors are available as another XML file in Resource Description Format (RDF) [7].

The authors [12] generate what is called Rotation and scale invariant hybrid descriptor (RSHD). The authors found that the RSHD descriptor is inherently rotation-invariant and describes the image features more efficiently. The conducted experiments of [12] show promising results.

III. The Proposed System

The system, which was previously introduced [5], consists of three main phases shown in Figure 1.

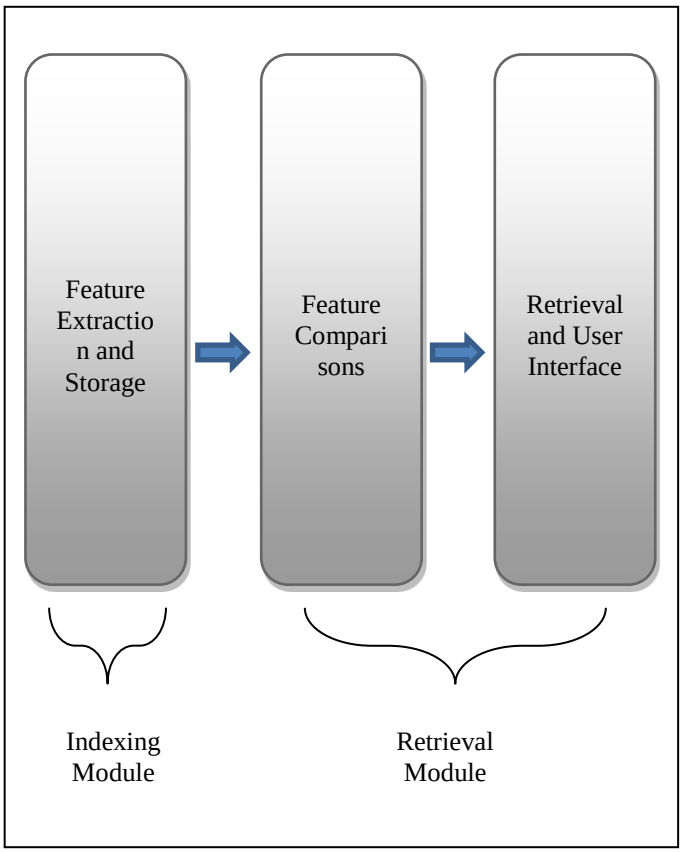


Fig. 1. Example of a figure caption. (figure caption)

The main motivation to conduct this research is the existing of the flipped images in the dataset used in the experiments introduced in [5]. One of the major limitations of the system [5] is that images that are considered as relevant are not retrieved by the introduced system in [5] because they are flipped. Figure 2 shows two relevant images each one is flipped to a different side.

Fig. 2. Flipped Images

Figure 2 shows two X-ray images, each representing an X-ray for an ankleL one is directed to the left side as shown in Figure 2 (D), and the other is directed to the right side as shown in Figure 2 (G). Using the introduced system in [5], when one of the images in Figure 2 is used as an input, the other image is not retrieved in the output images. This problem can be solved by flipping each image in the dataset, so that the features of the query image is compared against the original, and the flipped image, and if any one of them matches, the image is retrieved at the retrieval phase.

The work is mainly applied in the indexing phase, when the index dataset is created. Before the three low-level features are created, the image is flipped, and then the features of the original image and the flipped image are extracted.

Figure 3 shows how image (D) of Figure 2 looks like after flipping it.

The features of the two images of Figure 3 are extracted, the index for each image is computed, and then stored in the index data file. Note that the two index values belong to the same image, the main image in the dataset remains as is. The image is flipped only in the index application. So the number of images in the main dataset remains unchanged. Only the number of rows in the index data-file is duplicated.

Fig. 3. Image Before and After Flipping

Now if image (G) in Figure 2 is inserted as a query image, its features are extracted, and compared to the data-file which has the features' values of both images of Figure 3. Since Image (G) of Figure 2 and the flipped image of Figure 3 are relevant, the feature values of them will match, and the main image in the index data-file is retrieved even if it is flipped.

Another example of flipped and related images is shown in the next figures. Figure 4 shows two images for right- and left-section of a shoulder. The two images are relevant but are not retrieved in the same search issue due to the flipping issue. The same thing happens with Figure 5 that shows two images for a leg bent to the right and bent to the left.

Fig. 4. Two Images for Rigth and Left Shoulder Section

Fig. 5. Two Images for a Leg Bent to the Right and Left

IV. Experiments and Results

In order to measure the performance of the proposed system, an experiment is conducted using a test of 50 cases. Then, the search is applied using the two systems. The precision and recall is compared between the two systems.

Figure 6 shows the performance results of the proposed system that considers image flipping compared to the system introduced in [5] which does not consider image flipping.

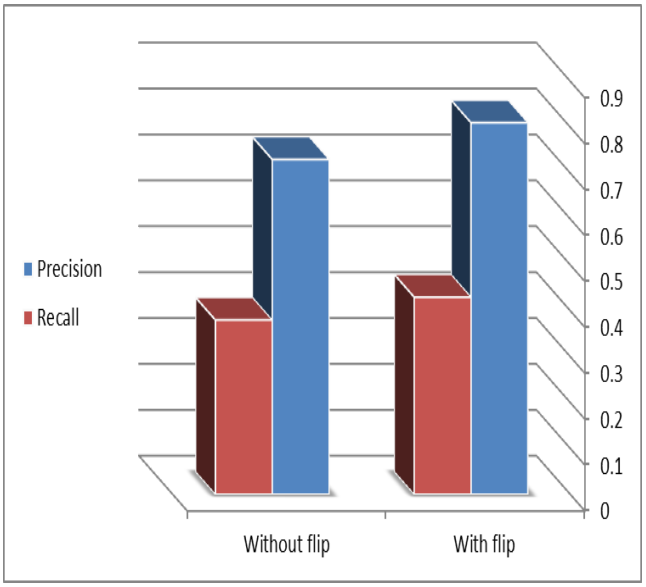


Fig. 6. Performance Comparison between the two Systems

Figure 6 shows that the improved system with flip-consideration has achieved 81% precision and 43% recall. While the previously introduced system without flipping-consideration has achieved only 73% precision and 38% recall.

V. Conclusion

The purpose of this research is to improve an existing CBIR system by incorporating image flipping during the search retrieval process. The results show a significant improvement on system. This improvement is made by considering the flipping issue in medical images, which has been ignored by previous research.

In the different medical datasets, many flipped relevant images are found. This flipping can lead to not retrieving all the related images. A system which considers the flipped images is introduced in this research. The results have shown more accurate image retrieval system with higher precision and recall values.

Based on the analysis and the experiments that are performed in this research, it is recommend to use image flipping as a pre-processing step in the future implementation of the whole system.

References

[1] Pavlopoulou C, Kak A, Brodley C. Content-based Image Retrieval for Medical Imagery. Proceedings SPIE Medical Imaging PACS and Integrated Medical Information Systems; 2003; San Diego CA.

[2] Müller H, Michoux N, Bandon D, Geissbuhler A. A Review of Content-based Image Retrieval Systems in Medical Application - Clinical Benefits and Future Directions. International Journal of Medical Informatics 2004; 73: 1-23..

[3] Katarina Trojacanec, Georgina Mirceva and Danco Davcev. Application of Edge Histogram Descriptor and Region Shape Descriptor to MRIs. The Faculty of Electrical Engineering and Information Technologies, Skopje, Macedonia; 2009.

[4] ZhixiaoXie. A Rotation- and Flip-Invariant Algorithm for Representing Spatial Continuity Information of Geographic Images in Content-Based Image Retrieval. Florida Atlantic University. Computers & Geosciences 30 (2004) 1093–1104. 2004.

[5] Suzan Amareen. Content-based Image Retrieval for Medical Applications using Low-level Image Descriptors, Master Thesis, Jordan University of Science and Technology, Defended on 31-12-2011.

[6] Wei-Horng Yeh, Ye-In Chang, An efficient iconic indexing strategy for image rotation and reflection in image databases. The Journal of Systems and Software 81 (2008) 1184–1195. National Sun Yat-Sen University, Kaohsiung, Taiwan, ROC, 2007.

[7] Gowri Allampalli-Nagaraj, Isabelle Bichindaritz. Automatic Semantic Indexing of Medical Images Using a Web Ontology Language for Case-Based Image Retrieval. Engineering Applications of Artificial Intelligence 22 p. 18– 25. 2009.

[8] Petrakis. Design and evaluation of spatial similarity approaches for image retrieval. Image and Vision Computing 20 (1), 59–76. 2002.

[9] Tsechpenakis, G., Akrivas, G., Andreou, G., Stamou, G., Kollias, S. Knowledgeassisted video analysis and object detection. In: EUNITE 2002—European Symposium on Intelligent Technologies, Hybrid Systems and their Implementation on Smart Adaptive Systems. Image Video and Multimedia Laboratory, Department of Electrical and Computer Engineering, National TechnicalUniversity of Athens Retrieved October 30, 2006

[10] Sjo¨ berg, M., Laaksonen, J., Po¨ lla¨ , M., Honkela, T. Retrieval of multimedia objects by combining semantic information from visual and textual descriptors. In: Kollias, S., Stafylopatis, A., Duch, W., Oja, E. (Eds.), Proceedings of 16th International Conference on Artificial Neural Networks. Springer, Berlin, pp. 75–83. 2006.

[11] Perner, P. Image mining: issues, framework, a generic tool and its application to medical-image diagnosis. Journal Engineering Applications of Artificial Intelligence 15 (2), 205–216. 2007.

[12] Shiv Ram Dubey, Satish Kumar Singh, Rajat Kumar Singh, “Rotation and scale invariant hybrid image descriptor and retrieval”, Computers and Electrical Engineering 46 (2015) 288–302.

A Study to Investigate State of Ethical Development in E-Learning

AbdulHafeez Muhammad [1,2]
[1]College of Computer Science, King Khalid University, Abha - Saudi Arabia
[2]Kulliya of ICT, IIUM, Kuala Lumpur, Malaysia

Mohd. Feham MD. Ghalib
Kulliya of Islamic Revealed Knowledge and Human Sciences, International Islamic University, Kuala Lumpur, Malaysia

Farooq Ahmad
College of Computer Science & Information Technology, University of AlBaha, Al Baha, Kingdom of Saudi Arabia

Quadri N Naveed [1,2]
[1]College of Computer Science, King Khalid University, Abha - Saudi Arabia
[2]Kulliya of ICT, IIUM, Kuala Lumpur, Malaysia

Asadullah Shah
Kulliya of ICT, International Islamic University Malaysia Kuala Lumpur, Malaysia

***Abstract*—Different researches evidenced that e-learning has provided more opportunities to behave unethically than in traditional learning. A descriptive quantitative enquiry-based study is performed to explore same issue in e-Learning environments. The factors required for ethical development of students were extracted from literature. Later, efforts are made to assess their significance and their status in e-Learning with 5-point Likert scale survey. The sample consisted of 47 teachers, 298 students, and 31 administrative staff of e-learning management being involved in e-Learning. The work also observed state of students on various ethical behaviors. The study emphasized that the physical presence of teacher, an ethically conducive institutional environment, and the involvement of the society members are among the main factors that help in the ethical development of a student which are missing in e-Learning. The results of the study showed that the moral behavior of e-Learners is at decline because of lack of these required factors in e-Learning. This work also suggested the need of a model indicating how these deficiencies can be addressed by the educational institutions for ethical development of higher education learners.**

***Keywords*—e-Learning; ethical development; ethics**

I. Introduction

The objective of education is not only to make a student knowledgeable but also add rational thinking self-sufficiency, knowledge ability, and to ethically develop a student [1]. Moreover, it is the responsibility of university education to produce graduates who use their competence for the welfare of the society. The written codes of ethics and making it binding on the members to follow for almost all the professional bodies (IEEE, ACM, Medical and others) is an indication of the necessity of ethical understanding [2]. Many studies show that ethical understanding and its application are declining in graduates [3]–[6] which is creating harmful situations for the society. There are many reasons for the gradual decline in the ethical development of graduates [6]–[9]. Some of them are attention to child at the early age in the family [10]–[12], disintegrating family systems [3], [12], no attention to ethical development in early education [8], [13], the philosophy that knowledge should be independent of religion and local social context [14]–[16], disappearance of explicit contents on ethics from the study material [3], [17], and the quick fix approach and economic push to teach market-oriented courses [3], [18].

This study is part of a bigger project to explore these ethical issues in e-Learning by highlighting the need of a model which supports in moral and ethical development of students. In the remaining part of the paper, statement of problem, objective of study, and literature review would be discussed first. Later, data analysis and its interpretation would be described. Finally, conclusion and future work would be highlighted.

II. Statement of Problem

Despite the many advantages of integrating technologies in education in the form of e-Learning, it has exacerbated the problems of cheating, plagiarism, procrastination, and violation of privacy [19]. Moreover temptation for unethical behavior among students is higher in e-Learning as compared with traditional learning, so the chances for learners to deviate from their academic objectives and behave unethically are higher in e-Learning[6], [7], [10], [20]–[22]. Many researchers have pointed out that although the ethical development of graduates and professionals in general is at decline, the situation is more alarming in case of e-Learning [4], [6], [7], [17], [19], [23]–[25]. It has been realized that ethical issues are on the rise in traditional learning, and specifically in the context of e-Learning due to impact of technology on teacher, society, study material, and academic institutions. As a result, both the society and professions are suffering. A similar perception that the ethical behavior of students in e-Learning is weaker as compared to the courses offered through traditional methods existed in King Khalid University of Kingdom of Saudi Arabia. The Ministry of Higher Education of the Kingdom of Saudi Arabia is supporting e-learning because of intrinsic needs of the country. Therefore, it is

necessary to study and investigate the state of ethical development in e-Learning at Saudi universities and suggests some ways, which can help to reduce the risks of factors involved in encouraging students on unethical behavior.

III. Purpose of Study

The purpose of this paper is to investigate state of ethical development as perceived by stakeholders by identifying the required factors of ethical development

The study aimed to achieve the following objectives:

a) To assess the significance of the factors required for ethical development of students.
b) To see the status of the factors required for ethical development of students in e-Learning.
c) To observe the status of ethical behavior of students in e-Learning.

IV. Literature Review

A. E-Learning

Creation, dissemination, managing data, and storing information are some of the activities performed in academic institution through different information and communication tools called e-Learning [26]. Authors of another study [27] believe that the procedure of acquiring knowledge and skills consists of five main features. These features are the teacher, the content, the learner, goals to be achieved, and the context or learning environment. Various changes have taken place in the contemporary world through advancements in information and communication technologies (ICTs). Elements of e-learning have also been revolutionized through the use of ICTs. Delivery of study material, evaluation of the abilities of students and improvement of students through teacher and student interaction is done through the use of ICTs in e-learning institutions. Most of the academic institutions are using e-learning as educational platform called Learning Management System (LMS) because of its advantages as synchronous and asynchronous learning, greater increase to information, more collaboration, better communication and lastly, improvement in pedagogical cost-effectiveness [28]. Blended and fully online modes are the two modes of e-learning used in academic institutions. The mode, in which entire contents are delivered through technology without using learner's physical interaction with the instructor, is the fully online mode. On the other hand, a combination of face-to-face and virtual environment in a traditional classroom is known as blended mode. One of the basic differences between a learning classroom and a traditional classroom is that it is not required by the students and instructors to be always physically present together in an e-learning classroom.

B. E-Learning in Kingdom of Saudi Arabia (KSA)

E-Learning has received much attention amongst the education sector and academic communities of Saudi Arabia as it offers access to many students who aspire to study at the universities but live in remote areas. In the case of female students, parents do not allow them to travel to cities for education. Ministry of Higher Education became more interested in integrating e-Learning in universities, opening new avenues of delivering instruction, making learning accessible and most importantly, improving students learning outcomes[29], [30].

The Ministry of Higher Education, aiming to facilitate the electronic environment and provisions for their universities, has established the National Centre for e-Learning and Distance Education (NCeL). Recently, almost all Saudi universities have integrated e-Learning in their degree program by choice and demand of county. King Khalid University (KKU), Abha, is one of the largest universities in KSA having more than 50,000 students and offering blended and online courses along with the traditional offerings in various disciplines [31]. The KKU is amongst the first higher education institutes in Saudi Arabia that employs technology in educating students. KKU has observed three stages of e-learning since 2009, which are: supplementary level, blended level, and entirely online level[31].

C. Ethical Development

Ethics can be defined as socially acceptable and moral behavior, which is contradictory with wrong doings or taboos of the society [32]. The term "ethics" is often used to describe the scientific study of moral behavior. Character, Morals, Values, and Ethics are the concepts of sociology which are interrelated with each other for ethical development [33]. There is a particular set of values, beliefs, and means through which, objectives and aims of educational institutions are delivered. These customs, attitudes and understandings are necessary to be understood by adults and youth. Through this, students are able to build their characters by learning self-competence, awareness, rational thinking and at like literate people in the society [34]. The main purpose of education is to give a direction for a successful life. The new generation is supported through the implementation of essential skills, knowledge, attitudes and understandings to make their personality and intellect useful for the welfare of the society. A teacher plays a significant role in this process because he/she grooms the already existing talents according to the requirement in every individual [35], [36].

D. Ethical Issue in e-Learning causing unethical behavior

Recent advancement of technologies has effected on the structure of academic institutes, family and societies which become the major reason for the deficiencies in the moral character of individual. There is an objection on ICT-based education that it is focusing only on advanced styles and methods of education but not caring for betterment and advancement of society [37], [38]. Other than this, even supporters of e-Learning accept that it is not offering social and extracurricular support which are important for instilling cultural and moral values[9], [39], [40]. Ethical issues have become more dominant due to the use of technology. In order to fight this challenge, UN and UNESCO started special "Ethics Education Program". A declaration was also put forward by the European Association and European universities so that a conference can be arranged to continue the progress of ethical development in higher education. Almost all the professional bodies in the world have developed their Code of Ethics, which indicates that there has been a violation of ethical codes in the past or present [41].

In his study [42] highlighted that data collection, security, privacy, trust, and ever-present technologies are some of the ethical issues, which have been a consequence of latest technologies. This has also given rise to less human contact due to which, human beings have become more technology-dependent.

There are six critical ethical issues, which are - access of intellectual property, privacy, protecting children, and securing information that occur due the use of electronic mean[43].

Due to the integration of technologies, academic institutions are facing direct or indirect ethical issues. In the last decade, students have become more ethically declined even after academic learning and education.

According to Brown [10], academic frauds and other kinds of electronic problems are faced by educational institutions through the use of technology and internet.

Different researchers also highlighted some of the other ethical issues in the educational world [44]–[46]. They mentioned that students often use unauthorized resources for the completion of their assignments because copy pasting is the common practice of almost all educational world. In addition, e-learners often work in groups to perform an individual task and they often take advantage of the possible misuse of ICTs.

Moreover, 59% of the U.S students enrolled in e-Learning programmes have admitted that they have been in some kind of academic fraud in their academic period. About 27% indicated very frequent involvement while 32% marked frequent involvement according to The National Survey of Student Engagement [47]. The use of mobile phones in academic institutions has also made students unethical. They usually divert their attention from studies towards playing games on mobile phones so they spend time on luxuries by using them. About 40% of students affirmed that they have used mobile phones for sending text messages or receiving text messages during a lecture according to a survey report. On the other hand, 70% of the students stated that their phones rang in the middle of the class. Unethical use of ICTs has promoted plagiarism in academic institutions, suggesting the need for further investigation[48], [49]. Other than these, authors [6]–[8], [25], [26], [50], [51] also highlight ethical issues in e-learning and observed following deficiencies which can become the cause of these issues.

- *Lack of direct or face-to-face interactions with students.*
- *Lack of counseling, which promotes flexible and independent learning environment.*
- *Lack of involvement in the social circle, which makes students isolated from their families and friends.*

V. Methods and Procedures

The study was descriptive quantitative based on a survey approach. The population of the study was students, teachers and the administrative staff of e-learning management working in e-learning environments of King Khalid University, Abha. Random sampling technique was use to select 47 teachers, 298 students and 31 administrative staff being involved in e-Learning. A self-developed questionnaire based on literature was developed to explore three area of study, first, to see the significance of the factors required for ethical development of student, second, to observe the status of these important factors in e-Learning environment, and finally to investigate the ethical behavior of students in e-Learning. The instrument was also pilot tested and verified through the experts of the field.

VI. Data Analysis and Interpretation

A. Demographic Information

Quantitative data were collected from 376 students, teachers, and administrative staff of King Khalid University all involved in e-Learning. The data were tabulated and analysed by using SPSS. The respondents belonged to various disciplines and had varying experience with e-learning as shown in table-1 in table-2 respectively. In the sample, the number of respondents (376) related to e-Learning were 47 teachers, 298 students, and 31 from staff to manage e-Learning and 06 were from other categories of support staff. The respondents were associated with the fields of study in e-learning as Computer Science (156), Science (75), Management Science (67), Sharia and Law (62),) and Medical Sciences (16). Among these 376 respondents, 99 were associated with e-Learning for a period of less than one year, 141 for less than three years but more than one year, and 136 were associated with e-Learning for more than three years.

TABLE I. Categories of Respondent W.R.T. Their Disciplines

Respondents	*Computer Sc*	*Science*	*Management Studies*	*Sharia & Law*	*Medical*	*Total*
Teachers	21	10	4	7	5	47
e-Learning Staff	14	6	6	2	3	31
Students	121	59	57	53	8	298
Total	156	75	67	62	16	376

TABLE II. Categories of Respondent W.R.T Experience in the Use of E- Learning

Respondents	*Less than 1 year*	*More than 1 year but less than 3 years*	*More than 3 year*	*Total*
Teachers	11	23	13	47
e-Learning Staff	8	3	20	31
Students	80	115	103	298
Total	99	141	136	376

The descriptive and inferential statistics were applied for data analysis to find significance of factors and their status in e-learning.

B. Significance of Factors Required for Ethical Development

The literature indicated that the main factors for ethical development are the teacher, members of society (including

parents, family and friends), physical environment provided to students, and explicit course contents on ethics. The statements mentioned in Table 3 were asked to the respondents to assess the significance of factors required for ethical development.

TABLE III. SIGNIFICANCE OF THE FACTORS REQUIRED FOR ETHICAL DEVELOPMENT OF STUDENTS (N=376)

Statements	Min.	Max.	Mean	Std. Deviation
Physical presence of teacher plays significant role in the ethical development of the students.	1	5	4.29	0.732
Members of society play a significance role in the ethical development of students.	1	5	4.15	0.928
Physical environment provided to student plays a significant role in the ethical development of students.	1	5	4.14	0.842
Study material in the curriculum plays a major role in the ethical development of students	1	5	3.53	1.070

Strongly agree=5, Agree=4, Neutral=3, Disagree=2, Strongly disagree=1

Table 3 indicates that most of the respondents strongly agree about the significance of factors required for ethical development of students i.e., physical presence of teacher (Mean= 4.29, Std. Deviation=0.732), member of society (Mean= 4.15, Std. Deviation=0.928), physical environment provided to student (Mean= 4.14, Std. Deviation=0.842) and Study material in curriculum (Mean= 3.53, Std. Deviation=1.070). All respondents agreed that these factors play a significant role in ethical development of students.

It can be concluded from the data in Table 3 that among all the respondents, factor, "physical presence of teacher" is having the highest mean value (4.29). Therefore, it is the most significant factor and study material in curriculum having least mean value (3.53) is the least significant factor for the ethical development of students.

Analysis of variance (ANOVA) test was applied to compare the opinion of teachers, students and e-learning management staff about significance of the factors required for ethical development of students. Significant level of $p < 0.05$ was adopted for the study. It can be seen in Table 4, that there is no significant difference among the opinion of teachers, students and e-learning management staff about different factors for ethical development of students. It can be concluded that all the respondents have the same opinion regarding the factors for ethical development of students.

TABLE IV. COMPARISON OF OPINION OF TEACHERS, STUDENTS AND E-LEARNING MANAGEMENT STAFF ON SIGNIFICANCE OF THE FACTORS REQUIRED FOR ETHICAL DEVELOPMENT OF STUDENTS (N=375)

Variables		Sum of Squares	Df	Mean Square	F	Sig.
Physical presence of teacher	Between Groups	2.414	2	1.207	2.267	0.105
	Within Groups	198.565	373	0.532		
	Total	200.979	375			
Members of society	Between Groups	0.637	2	0.318	0.368	0.692
	Within Groups	322.416	373	0.864		
	Total	3232.053	375			
Physical environment	Between Groups	0.697	2	0.349	0.490	0.613
	Within Groups	265.385	373	0.711		
	Total	266.082	375			
Study material	Between Groups	3.467	2	1.734	1.517	0.221
	Within Groups	426.150	373	1.142		
	Total	429.617	375			

C. Status of Factors Required for Ethical Development

Here, an effort is made to observe the status of factors, essential for ethical development in e-Learning environments. The statements related to same factors were asked to respondents to see the status of these factors in e-Learning environment as stated in Table 5.

TABLE V. STATUS OF THE FACTORS REQUIRED FOR ETHICAL DEVELOPMENT OF STUDENTS IN E-LEARNIN (N=376)

Statements	Min.	Max.	Mean	Std. Deviation
There is lack of physical presence of teacher in e-Learning, which is affecting ethical development of student.	1	5	4.03	0.861
Independent and flexible learning environment provided for e-Learning is not appropriate for the ethical development.	1	5	3.95	1.180
There is lack of involvement of members of the society in e-Learning, which is affecting ethical development.	1	5	4.14	0.796
There is lack of ethical contents in study material in e-Learning, which is negatively affecting the students.	1	5	3.94	1.063

Strongly agree=5, Agree=4, Neutral=3, Disagree=2, Strongly disagree=1

Table 5 shows the opinion of respondents about the status of factors required for ethical development of students. In the opinion of all respondents, the factors, that is,, "lack of physical presence of teacher in e-learning" (Mean= 4.03, Std. Deviation=0.861), "independent and flexible learning environment" (Mean=3.95, Std. Deviation= 1.180), "lack of involvement of members of society" (Mean= 4.14, Std. Deviation=0.796), "lack of ethical contents in study material" (Mean= 3.94, Std. Deviation=1.063), are affecting negatively in e-learning environment for ethical development of students. The conclusion that can be drawn is that "lack of involvement of society member in e-learning" having highest mean value (4.14) is the most deficient factor, which can influence more for ethical development of students. While "lack of ethical contents in study materials in e-learning" having mean value (3.94) is the least affecting factor for ethical development of students.

TABLE VI. COMPARISON OF OPINION OF TEACHERS, STUDENTS AND E-LEARNING MANAGEMENT STAFF ABOUT THE STATUS OF THE FACTORS REQUIRED FOR ETHICAL DEVELOPMENT OF STUDENTS (N=375)

Variables		Sum of Squares	df	Mean Square	F	Sig.
lack of physical presence of teacher	Between Groups	2.182	2	1.091	1.477	0.230

in e-Learning	Within Groups	275.552	373	0.739		
	Total	277.734	375			
Independent and flexible learning environment	Between Groups	4.043	2	2.021	1.456	0.235
	Within Groups	517.997	373	1.389		
	Total	522.050	375			
lack of involvement of members of the	Between Groups	0.261	2	0.130	0.205	0.815
	Within Groups	237.269	373	0.636		
	Total	237.529	375			
Lack of ethical contents in study material.	Between Groups	1.737	2	0.869	0.768	0.465
	Within Groups	422.090	373	1.132		
	Total	423.827	375			

ANOVA results in Table 6 show that there is no significant difference among the opinion of teachers, students and e-learning management staff about status of factors for ethical development of students in e-learning system. It can be concluded that all the respondents have same opinion regarding the status of factors for ethical development of students in e-learning system.

D. Status of Ethical Behavior of Students in e-learning courses

To observe the ethical status of students, only three values that is, plagiarism, punctuality (attendance/attention) and miscommunication (false representation/telling lies) are taken. Many researchers [52], [53] have taken these as the general indicators of unethical behavior of e-learners. Table 7 shows the ethical behavior of students in e-learning by asking statements of three values. The participants were asked to compare the status of the students' behavior for these values as in their conduct in online courses as compared to the normal courses.

TABLE VII. THE STATUS OF ETHICAL BEHAVIOR OF STUDENTS IN E-LEARNING (N=376)

Statements	Min.	Max.	Mean	Std. Deviation
Plagiarism, copy/paste and cheating are common among students in e-Learning.	1	5	4.37	0.704
Students are not attentive in e-Learning	1	5	3.72	1.062
Students miscommunicate while with the teachers and other students in e-Learning	1	5	3.63	1.145

Strongly agree=5, Agree=4, Neutral=3, Disagree=2, Strongly disagree=1

The values in Table 7 show that Plagiarism, copy/paste and cheating are common among students (Mean= 4.37, Std. Deviation=0.704), Students are not attentive (Mean= 3.72, Std. Deviation=1.062), and students miscommunicate with the teachers and other students (Mean= 3.63, Std. Deviation=1.145). It can be seen that in the opinion of the respondents, in online courses most of the students behave unethically because of more opportunities for cheating and being less serious.

TABLE VIII. COMPARISON OF OPINION OF TEACHERS, STUDENTS AND E-LEARNING MANAGEMENT STAFF ABOUT THE STATUS OF ETHICAL DEVELOPMENT OF STUDENTS (N=376)

Variables		Sum of Squares	df	Mean Square	F	Sig.
plagiarism, copy/paste and cheating	Between Groups	0.850	2	0.425	0.85	0.42
	Within Groups	184.765	373	0.495		
	Total	185.614	375			
Attention of students	Between Groups	1.034	2	0.517	0.45	0.63
	Within Groups	421.5106	373	1.130		
	Total	422.551	375			
Miscommunication	Between Groups	1.592	2	0.796	0.60	0.54
	Within Groups	490.022	373	1.314		
	Total	491.614	375			

ANOVA results in Table 8 show that there is no significant difference among the opinion of teachers, students and e-learning management staff about the three different status of ethical behavior of students in e-learning. Therefore, it can be concluded that all the respondents have same opinion regarding the status of ethical behavior of students in e-Learning.

VII. CONCLUSION RECOMMENDATIONS AND FUTURE WORK

The study was carried out to investigate the status of ethical behavior of students in an e-Learning environment from a sample of 376 stockholders of an e-Learning environment, that is, students, teachers and administrative staff.

From the findings of this study, it may be concluded that teacher, society members (family and friends etc.), environment and policies of academic institution, and teaching of explicit content on ethics are the essential factors for the ethical development of students. It was also tested statistically that, there is no difference among the opinions of respondents about the importance of factors required for ethical development. When respondents were enquired about the status of these essential factors in e-Learning environments, they agreed that lack of physical presence of teacher, no involvement of society members (family and friends etc.), no ethically conducive environment and policies of academic institution, and absence of explicit contents on ethics in courses is the cause of ethical decline of the e-learner. It was also tested statistically that, there is no difference among the opinions of respondents about there is lack of these factors required for ethical development in e-learning. Students' ethical behavior was also tested on three values of plagiarism, attendance/attention, and miscommunications/lies. It was seen that all the respondents agreed that the behavior of e-Learners is more unethical. It was also tested statistically that there is no difference among the opinions of respondents about status of ethical behavior of students in e-Learning.

Following are some recommendations for the institutions offering e-Learning based on the findings of the study.

- Institutions should provide proper alternatives to compensate for the absence of face to face interaction between students and teachers. For example, there must be interactive sessions, virtual class rooms and at least some physical face to face meetings in the start of the course.
- Teachers should adopt multimedia tools available in learning as technology in appropriate way to compensate for the lack of physical presence of teacher.
- There must be focused and supervised learning environment for students where exams should be conducted by proper monitoring tools like *Lockdown browsers* or *Netopt.*
- There must be proper polices and guidelines for students and teachers involved in e-learning. These policies should be properly implemented and monitored.
- The curriculum of professional studies should contain content related to the inculcation of ethical values in the students.
- Parents and the community leaders should be involved in the ethical building mechanism for the university students.

Finally, the study strongly recommends and highlights the need of a model for future work that may help the educators to equip the individuals with moral values and life skills. This study can be considered as base for future researchers because of lack of existing literature about this topic in the Arab countries. Authors of the papers are working to find ways and models which can compensate for the deficiencies identified in the ethical development of e-learners.

REFERENCES

[1] V. Campbell and R. Bond, "Evaluation of a character education curriculum," Educ. Values. New York Irvingt. Publ., 1982.

[2] G. Wood, "A cross cultural comparison of the contents of codes of ethics: USA, Canada and Australia," J. Bus. Ethics, vol. 25, no. 4, pp. 287–298, 2000.

[3] F. Ahmad, "Computer Science & Engineering Curricula and Ethical Development," in Teaching and Learning in Computing and Engineering (LaTiCE), 2014 International Conference on, 2014, pp. 220–225.

[4] C. Anitha and T. S. Harsha, "Ethical Perspectives in Open and Distance Education System.," Turkish Online J. Distance Educ., vol. 14, no. 1, pp. 193–201, 2013.

[5] P. Brey, "Ethical issues for the virtual university," Rep. cEVU Proj. (EuroPACE/European Comm. To Appear online www. cevu. org, pp. 1–25, 2003.

[6] M. AbdulHafeez, Farooq A, and S. Asadullah, "Resolving Ethical Dilemma in Technology Enhanced Education through smart mobile devices," Int. Arab J. e-Technology, vol. 4, no. 1, pp. 25–31, 2015.

[7] A. H. Muhammad, H. A. Wahsheh, A. Shah, and F. Ahmad, "Ethical perspective of learning management system a model to support moral character of online learner," in Information and Communication Technology for The Muslim World (ICT4M), 2014 The 5th International Conference on, 2014, pp. 1–6.

[8] M. Abdulhafeez, A. Farooq, and S. Asadullah, "Integration of ethical perspective in e-education," EDULEARN12 Proc., pp. 3759–3765, 2012.

[9] M. AbdulHafeez, S. Asadullah, M. Rosydi, and a. Farooq, "Inculcating ethical values in the students through e-Learning platform," 2013 5th Int. Conf. Inf. Commun. Technol. Muslim World, pp. 1–6, Mar. 2013.

[10] T. Brown, "Ethics in eLearning," Rev. Educ. do Cogeime, pp. 211–216, 2008.

[11] Su. Couch and S. Dodd, "Doing the Right Thing," J. Fam. Consum. Sci. Sep, vol. 97, p. 3, 2005.

[12] P. K. Rono and A. A. Aboud, "The role of popular participation and community work ethic in rural development: the case of Nandi District, Kenya," J. Soc. Dev. Afr., vol. 18, no. 2, pp. 77–104, 2003.

[13] A. Colby and W. M. Sullivan, "Ethics teaching in undergraduate engineering education," J. Eng. Educ., vol. 97, no. 3, pp. 327–338, 2008.

[14] M. H. Khamis and M. J. Salleh, "The philosophy and objectives of education in Islam," 2010.

[15] Z. Al Zeera, Wholeness And Holiness In Education An Islamic Perspective. IIIT, 2001.

[16] K. Musbahtiti, M. Ramadan Saady, and A. Muhammad, "Comprehensive e-Learning system based on Islamic principles," in Information and Communication Technology for the Muslim World (ICT4M), 2013 5th International Conference on, 2013, pp. 1–5.

[17] M. Abdulhafeez, A. Farooq, and S. Asadullah, "INNOVATIVE ROLES AND RESPONSIBILITIES OF TEACHERS AND LEARNERS IN E-LEARNING," EDULEARN12 Proc., pp. 4121–4127, 2012.

[18] O. C. Ferrell, J. Fraedrich, and others, Business ethics: Ethical decision making & cases. Cengage Learning, 2014.

[19] J. Trushell, K. Byrne, and R. Simpson, "Cheating behaviours, the Internet and Education undergraduate students," J. Comput. Assist. Learn., vol. 28, no. 2, pp. 136–145, 2012.

[20] M. I. FARISI, "Academic Dishonesty In Distance Higher Education: Challenges And Models For Moral Education In The Digital Era," Turkish Online J. Distance Educ., vol. 14, no. 4, 2013.

[21] A. H. Mohammad, A. Farooq, and S. Asadullah, "Integration of ethical perspective in e-education," no. July, pp. 3759–3765, 2012.

[22] M. E. Brown and L. K. Treviño, "Ethical leadership: A review and future directions," Leadersh. Q., vol. 17, no. 6, pp. 595–616, 2006.

[23] R. G. Ledesma, "Academic dishonesty among undergraduate students in a Korean university," Res. World Econ., vol. 2, no. 2, p. p25, 2011.

[24] M. Nejati, R. Jamali, and M. Nejati, "Students' ethical behavior in Iran," J. Acad. Ethics, vol. 7, no. 4, pp. 277–285, 2009.

[25] A. H. Muhammad and K. T. Musbah, "Improvement Quality of LMS Through Application of Social Networking Sites.," Int. J. Emerg. Technol. Learn., vol. 8, no. 3, p. pp–48, 2013.

[26] A. Hafeez, M. Muhammad, A. Shah, and others, "Integration of ICT in education: an Islamic perspective.," pp. 61–71, 2011.

[27] T. Sepic, I. Pogarčić, and S. Raspor, "eLearning: The influence of ICT on the style of teaching," in MIPRO, 2010 Proceedings of the 33rd International Convention, 2010, pp. 995–1000.

[28] A. Nawaz and G. M. Kundi, "Demographic implications for the user-perceptions of e-learning in higher education institutions of NW. FP, PAKISTAN," Electron. J. Inf. Syst. Dev. Ctries., vol. 41, 2010.

[29] A. Aljabre, "An exploration of distance learning in Saudi Arabian universities: current practices and future possibilities," Int. J. Instr. Technol. Distance Learn., vol. 9, no. 2, pp. 21–28, 2012.

[30] A. Clementking and A. Muhammad, "Technology Based Learning Analysis of CBCS Model at KKU.," Int. J. Emerg. Technol. Learn., vol. 8, no. 3, 2013.

[31] K. K. U. E-Learning Deanship, "KKU," 2016. [Online]. Available: www.elc.kku.edu.sa. [Accessed: 10-Jan-2016].

[32] P. Bowden and V. Smythe, "Theories on teaching & training in ethics," 2008.

[33] H. Lee, H. Chang, K. Choi, S.-W. Kim, and D. L. Zeidler, "Developing character and values for global citizens: Analysis of pre-service science teachers' moral reasoning on socioscientific issues," Int. J. Sci. Educ., vol. 34, no. 6, pp. 925–953, 2012.

[34] I. N. George and U. D. Uyanga, "Youth and Moral Values in a Changing Society," IOSR J. Humanit. Soc. Sci., vol. 19, no. 6, pp. 40–44, 2014.

[35] Z. A. Shaikh and S. A. Khoja, "Personal learning environments and university teacher roles explored using Delphi," Australas. J. Educ. Technol., vol. 30, no. 2, 2014.

[36] M. Abdulhafeez, A. Farooq, and S. Asadullah, "Innovative roles and responsibilities of teachers and learnersin E-Learning," EDULEARN12 Proc., pp. 4121–4127, 2012.

[37] P. Brey, "Ethical issues for the virtual university," Rep. cEVU Proj. (EuroPACE/European Comm. To Appear online www. cevu. org, 2003.

[38] P. Brey, "Social and ethical dimensions of computer-mediated education," J. Information, Commun. Ethics Soc., vol. 4, no. 2, pp. 91–101, 2006.

[39] K. K. Ali, R. Salleh, and M. Sabdin, "A study on the level of ethics at a Malaysian private higher learning institution: comparison between foundation and undergraduate technical-based students," Int. J. Basic and, Appl. Stat., vol. 10, no. 8, pp. 35–49, 2010.

[40] J. L. Cordova and P. Thornhill, "Academic honesty and electronic assessment: tools to prevent students from cheating online---tutorial presentation," J. Comput. Sci. Coll., vol. 22, no. 5, pp. 52–54, 2007.

[41] H. Ten Have, "Promoting and applying bioethics: the ethics programme of UNESCO," 2010.

[42] B. C. Stahl, S. Rogerson, and K. J. Wakunuma, "Future Technologies: The Matter of Emergent Ethical Issues in Their Development," in Future Computing, Service Computation, Cognitive, Adaptive, Content, Patterns, 2009. COMPUTATIONWORLD'09. Computation World:, 2009, pp. 603–607.

[43] B. Kracher and C. L. Corritore, "Is there a special e-commerce ethics?," Bus. Ethics Q., pp. 71–94, 2004.

[44] R. A. Fess, "Cheating and plagiarism. Ethics and Higher Education. May, WW editor." New York: Macmillan Publishing Company and American Council on Education, 1990.

[45] W. L. Kibler, "Academic Dishonesty: A Student Development Dilemma.," NASPA J., vol. 30, no. 4, pp. 252–267, 1993.

[46] D. L. McCabe, K. D. Butterfield, and L. K. Trevino, "Academic dishonesty in graduate business programs: Prevalence, causes, and proposed action," Acad. Manag. Learn. Educ., vol. 5, no. 3, pp. 294–305, 2006.

[47] G. D. Kuh, "The national survey of student engagement: Conceptual and empirical foundations," New Dir. Institutional Res., vol. 2009, no. 141, pp. 5–20, 2009.

[48] T. Andrews, L. Dyson, R. Smyth, and R. Wallace, "The Ethics of M-Learning: Classroom Threat or Enhanced Learner Agency?," 2011.

[49] A. Berglund, A. Pears, A. Nylén, F. Ahmad, B. Alghamdi, K. Alghamdi, A. Alhabish, A. Aljoufi, E. Alzahrani, R. Alzahrani, and others, "Teaching and Learning Computer Science at Al Baha University, Saudi Arabia: Insights from a Staff Development Course," in Learning and Teaching in Computing and Engineering (LaTiCE), 2015 International Conference on, 2015, pp. 1–6.

[50] K. Musbah, M. R. Saady, and A. Muhammd, "Comprehensive E-Learning System Based on Islamic Principles," p. 4799, 2013.

[51] A. H. MUHAMMAD and A. SHAH, "CHAPTER FIFTEEN E-LEARNING: INCULCATION OF VALUES AND ETHICS IN HIGHER EDUCATION LEARNERS MUHAMMAD SHAHID FAROOQ," Creat. Learn. MOOCs Harnessing Technol. a 21st Century Educ., p. 205, 2015.

[52] K. Nagi, "Solving ethical issues in e-learning," Spec. Issue Int. J. Comput. Internet Manag., vol. 14, no. SP1, pp. 1–7, 2006.

[53] L. M. Hinman, "Academic integrity and the world wide web," ACM SIGCAS Comput. Soc., vol. 32, no. 1, pp. 33–42, 2002.

[54]

Improved Tracking Using a Hybrid Optcial-Haptic Three-Dimensional Tracking System

[1,2]M'hamed Frad , [1]Hichem Maaref, [1] Samir Otmane, [2]Abdellatif Mtibaa
[1]IBISC Laboratory, Université d'Evry Val d'Essonne, France
[2]EμE Laboratory, University of Monastir, Tunisia

***Abstract*—The aim of this paper is to asses to what extent an optical tracking system (OTS) used for position tracking in virtual reality can be improved by combining it with a human scale haptic device named Scalable-SPIDAR. The main advantage of the Scalable-SPIDAR haptic device is the fact it is unobtrusive and not dependent of free line-of-sight. Unfortunately, the accuracy of the Scalable-SPIDAR is affected by bad-tailored mechanical design. We explore to what extent the influence of these inaccuracies can be compensated by collecting precise information on the nonlinear error by using the OTS and applying support vector regression (SVR) for calibrating the haptic device reports. After calibration of the Scalable-SPIDAR we have found that the average error in position readings reduced from to 263.7240±75.6207 mm to 12.6045±8.4169 mm. These results encourage the development of a hybrid haptic-optical system for virtual reality applications where the haptic device acts as an auxiliary source of position information for the optical tracker.**

Keywords—virtual reality; Scalable-SPIDAR; support vector regression; hybrid tracking system

I. Introduction

Optical trackers provide a reliable and accurate position tracking for virtual reality applications. The optical tracking relies on measurements of reflected or emitted light [1]. It is therefore, evident, that there must be a clear free line-of-sight between the light source and camera assembly. This requirements turns out to be difficult to maintain at all times and a partial occlusion may turn out to be the biggest problem as it results in a tracking-loss.

Scalable SPIDAR haptic device is not dependent on free line of sight, is suitable for a large-scale immersion and is significantly cheaper than commercial optical trackers. This device was used in virtual reality to track and measure motion of user's hand as well as to enable large scale immersion. Unfortunately, previous experiments with Scalable-SPIDAR [2] have revealed significant inaccuracies between caused by design structure shortcomings.

Our work focus on combining a human scale haptic device and an optical tracker in hybrid tracking system; this approach aims to overcome obstructions of line-of-sight while maintaining an interrupted and accurate tracking for applications in virtual reality.

This paper deals with an important point in the feasibility of such a system. For efficient application of a hybrid tracking system sufficient registration accuracy between the two components has to be achievable. This essential since the Scalable-SPDIAR is intended to serve as secondary source of position information for the optical tracker and therefore have to be reported in the coordinate system of the optical tracker.

Furthermore, it has to be assessed to what extent systematic errors in the Scalable-SPIDAR reading positions that stem from the bad-tailored mechanical design can be calibrated.

II. Previous Work

In the last decade several hybrid tracking systems have been proposed in the literature. The suggested methods attempt to compensate for the shortcomings of each tracking technology by using multiple measurements to provide robust tracking. State et al's [3] work developed a hybrid tracking scheme that has the registration accuracy of vision-based tracking systems and the robustness of magnetic tracking systems. Similar to this, in [4] the authors built a hybrid tracking system integrating optical and magnetic tracking. The built system is faster than a standalone optical tracker and outperforms a magnetic system in term of accuracy.

You et al [5] presented a hybrid approach with integrated inertial and vision tracking technologies. They use the complementary nature of these two tracking technologies to overcome the shortcomings in each separate component. In [6], an image-based system is coupled with an inertial in order to provide robust and accurate tracking. In fact, in cases when the image based system fails due to abrupt movements, the inertial system takes over.

Birkfellner et al [7] developed a hybrid tracking system that combines an electromagnetic tracking system and optical tracker in order to avoid obstructions of the line-of-sight necessary for the operation of the OTS while maintaining an interrupted tracking and the accuracy needed in computer aided-surgery. More recently Harders et al [8] introduced a hybrid tracking method that combines the IR optical tracker with a vision based tracking approach

III. Virtual Reality Setup

For our experiment, we used an infrared 6DOF optical tracking system (ARTrack1/Dtrack) with an accuracy from 0.4 to 1.4 mm. The infrared (IR) cameras ARTrack1 illuminate the measurement volume by an IR flash. They are able to recognize retro reflective markers and they compute the marker positions in image coordinate (2D) with high precision [9].

The IR optical tracker is combined with a human scale haptic device called Scalable-SPIDAR [10] for Space Interface Device for Artificial Reality. The device is derived from the original desktop SPIDAR which was developed by Hirata and Sato [11]. The scalable-SPIDAR is composed of a cubic frame that encloses a cave-like space, where the use can move around to perform large scale movements. The front side of the device holds a large screen where a generated virtual world is displayed. The device has 8 couples of DC motor/rotary encoder mounted in the corners of the cubic frame. Position of the user's hands can be measured by the length of the strings. The length of a string is known by reading the values from the rotary encoder. The Scalable-SPIDAR haptic device is however is subject to inaccuracies due shortcomings in the mechanical structure design.

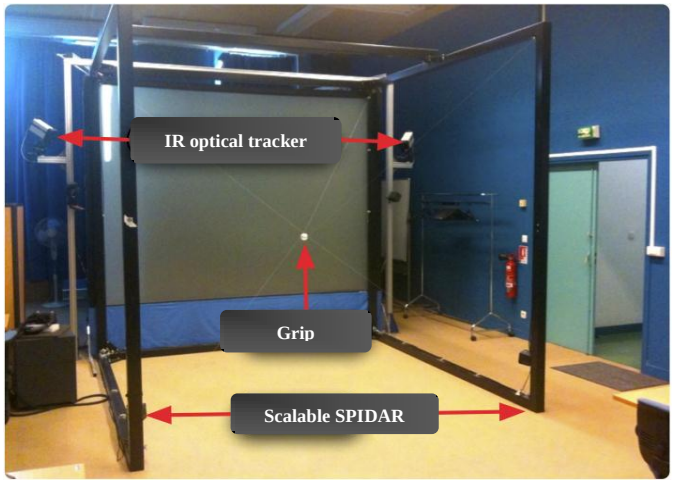


Fig. 1. View depicting all the components of our VR setup. The IR optical tracker (ARTrack1) was mounted below the Scalable-SPIDAR haptic device

IV. METHODS

In order to access the achievable accuracy from this hybrid tracking setup, we had to perform the following steps:

- Register the coordinate systems Ref_{Haptic} and $Ref_{Optical}$ in order to merge the data from the different systems modalities such that the position data from the Scalable SPIDAR are reported in $Ref_{Optical}$.
- Compensate erroneous position readings from the Scalable-SPIDAR caused by bad-tailored design.

A. Registration of tracking systems

The IR optical tracker determines the coordinates of retro reflective markers within the measurement volume. In order to acquire 3D point measurements of the haptic point in the world (i.e. Optical tracker) coordinate system an IR marker is attached at the end of the Scalable-SPIDAR grip.

A custom procedure allows the user to accomplish the Scalable-SPIDAR initialization in precise and repeated manner. Using the IR optical tracker, the user can accurately place to fixed known location within the working volume: the Scalable-SPIDAR origin.

Following his, we are able to report the same location in a common world coordinate system. Let $p_i^{RefOptical}$ and $p_i^{RefHaptic}$ represent the position of the grip, to which attached an IR marker, respectively in the haptic and optical coordinate systems. The position of the grip is determined by: $\vec{p}_{Ref Optical} = \vec{p}_{Ref Haptic} + \vec{t}$ where $\vec{t}$ is the translation vector between the Optical Tracker and the Scalable-SPIDAR origins.

B. Calibration of the Scalable-SPIDAR haptic device

Scalable-SPIDAR is a multi-modal haptic device for large scale virtual environment. It provides a workspace that is large enough to cover almost the measurement volume seen by the infrared cameras. However, due to various limitations, it is subject to inaccuracies and therefore cannot provide faithful data rendering. Problems are mainly caused by shortcomings in mechanical structure design. The calibration method involves several steps: characterizing the Scalable-SPIDAR haptic device and then applying methods to correct errors in the reported position. With our setup, we calibrated the Scalable-SPIDAR for a working volume of 1m x 1m x 1m.

1) Characterization Protocol

In order to characterize the Scalable-SPIDAR haptic device, we collect tuples that consist of the tracked data and the "truth values" what a convenient reference should an accurate reference report. The accuracy of these truth values is crucial of the calibration method. Therefore, we use the IR optical tracker available within our setup. To provide for well-distributed data that can be collected, a volumetric calibration protocol is proposed. To this end, the screen of our setup is filled by a virtual grid divided into a sequence of a small cubes. Each small cube corresponds to a sub-space of the Scalable-SPIDAR workspace.

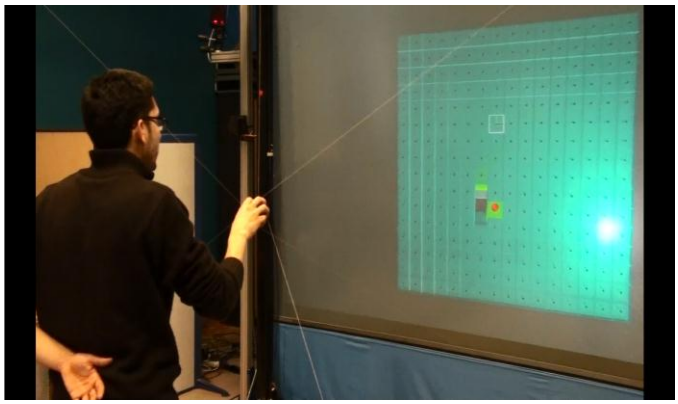


Fig. 2. Display in the Scalable-SPIDAR of boxes to collect. The virtual grip (the red sphere) is moved inside colored box to collect point

A user holds the device's grip s reasonably straight and moves it until the virtual grip ranges inside the cube. Programs record then the position given by the IR optical tracker and the haptic device. Once these positions recorded, the cube vanishes ensuring that only one measurement was associated to this sub-space. In practice a volume of 1 m^3 is considered. This volume is divided into 4096 small cubes. Hence 4096 point measurements are sampled inside the virtual grid. The proposed protocol sounds well for gathering a large number of data points with pretermitted distribution using a quasi-static collection mode.

At each of the resulting 4096 points, measurements reported by the two devices were taken, and the position error was determined as the distance between the tracked position by the Scalable-SPIDAR and the corresponding reference position as reported by the optical tracking system (OTS):

$$err_{pos}=\sqrt{(x_{OTS}-x_{S\_SPIDAR})^2+(y_{OTS}-y_{S\_SPIDAR})^2+(z_{OTS}-z_{S\_SPIDAR})^2} \quad (1.1)$$

Where $(x_{S\_SPIDAR}, y_{S\_SPIDAR}, z_{S\_SPIDAR})$ is the position reported by the Scalable-SPIDAR and $(x_{OTS}, y_{OTS}, z_{OTS})$ is the reference position.

Fig 3 represents the position errors spatially at each reference position $(x_{OTS}, y_{OTS}, z_{OTS})$, with the error magnitudes proportional to the corresponding circle diameters. The plot clearly shows that errors are more pronounced when the grip is manipulated away from the center of the workspace toward the edges of the cubic frame.

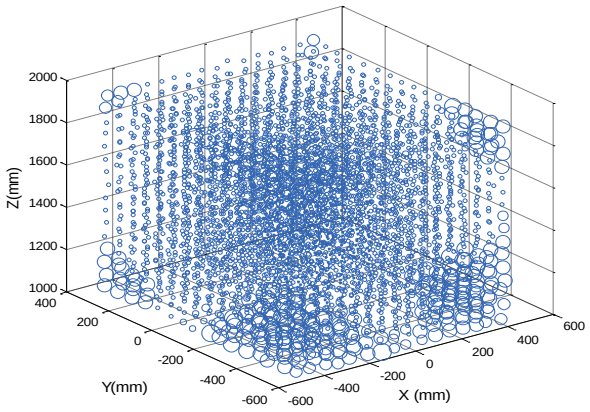


Fig. 3. Distance errors represented spatially

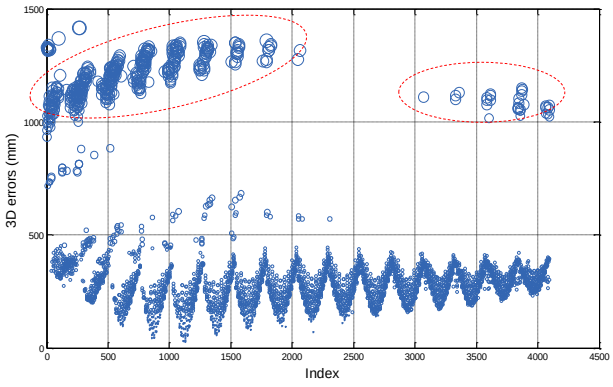


Fig. 4. Distance errors represented in Sequence

In figure 4, position errors are plotted as a function of the sequence in which they were collected from the front of the volume to the back. Plotting distance errors as 1D plot results in the loss of a lot of spatial information, but still shows the same trend, and from the plot's periodicity we can infer that errors increase at the volume edges. Note the clusters of large error in the upper left and right corner of plot, these groups of outliers are removed using boxplot method.

2) Calibration technique: Support Vector regression (SVR)

A function approximation problem can be expressed as to find a function f from a set of observations, $X=\{(\vec{x_1}, y_1), (\vec{x_2}, y_2), ..., (\vec{x_N}, y_N)\}$ with $\vec{x_i} \in R^m$ and $y_i \in R$ where N is the number of training data, $\vec{x_i}$ is the i input vector, and y_i is the desired output for the input $\vec{x_i}$. Based on the support vector machine theory, SVR is to approximate the given observations in an m-dimensional space by a linear function in another feature space. The function in SVR is determined as [12]:

$$f(\vec{x}, \vec{\theta}) = \langle \vec{\theta}.\Phi(\vec{x}) \rangle + b \quad (1.2)$$

Where $\langle .,. \rangle$ is an inner product defined on F, $\Phi(.)$ is nonlinear mapping function from R^m to F, $\vec{\theta} \in F$ is a weight vector to be identified in the function, and b is a threshold. Generally, the considered cost function is [31] [32]:

$$R_{SV}[f] = R_{emp}[f] + C.\|\vec{\theta}\|^2 \quad (1.3)$$

where $R_{emp}[f] = \frac{1}{N}\sum_{i=1}^{N} L(y_i - f(\vec{x_i}, \vec{\theta})) = \frac{1}{N}\sum_{i=1}^{N} L(e_i)$

[13], [14], $L(y_i - f(\vec{x_i}, \vec{\theta}))$ is the loss function measuring the error between y and the estimated output $f(\vec{x_i}, \vec{\theta})$ for a given $\vec{x}$, and $C \succ 0$ is a regular constant. The goal of adding the regularization term is to maintain the weight vector as small as possible in the approximation process. When over fitting phenomena happens, some undesirable information, has been modeled in the function. Those undesirable signals usually are not smooth, and as result, some parameters may become large to accommodate such behaviors. Therefore, in (1.3) the cost function has incorporated the intention to minimize $\vec{\theta}$, which in turn, reduces the model complexity. In other terms, the regularization term in (1.3) controls the tradeoff between the approximation accuracy and the model complexity in order to provide good generalization performance accuracy [15].

In classic SVR, the $\varepsilon-$ insensitive function is used as the loss function (1.3). It was first presented in the original SV algorithm [16], [17]. The $\varepsilon-$ insensitive function is defined as

$$L(e) = \begin{cases} 0, & \text{for } |e| \le \varepsilon \\ |e| - \varepsilon, & \text{otherwise} \end{cases} \quad (1.4)$$

It was mentioned in [16] that the solution of the above problem can be formulated in terms of support vectors, $\vec{\theta} = \sum_{i=1}^{N} \beta_i \Phi(\vec{x_i})$ and the function f is then written as:

$$f(\vec{x},\vec{\theta}) = \sum_{i=1}^{N} \beta_i \left\langle \Phi(\vec{x_i}), \Phi(\vec{x}) \right\rangle + b \qquad (1.5)$$

In (1.5) , the inner product $\left\langle \Phi(\vec{x_i}), \Phi(\vec{x}) \right\rangle$ in the feature space is considered as kernel function $K(\vec{x_i}, \vec{x})$ [18]. The choice of the kernel function is usually left for users. The kernel function chosen in our work is Gaussian and is defined as:

$$K(\vec{x}, \vec{x_i}) = \exp[\frac{-\left\|\vec{x} - \vec{x_i}\right\|^2}{2\tau^2}] \qquad (1.6)$$

Where τ is a constant. The coefficients β_i in (1.5) can be solved by quadratic programming methods with suitable transformation of the above problem into constraint optimization problems and properly rearranging the equation into a matrix form [19],[20] .

According to SVM theory, Support vector regression has the advantage of self-determining its structure. Therefore, there are no initialization problems for SVR. For training data with certain noise distributions the $\varepsilon-$ insensitive function [21]. Nevertheless, the robust effects against training data sets with outliers are not obvious in SVR. In this study, we use informal box plots to pinpoint the outlying points in the current training data.

V. RESULTS

A. Data preprocessing

Before applying the regression technique, input data need to be prepared. One essential task is to eliminate outliers. Training data without awareness may lead to unwanted data and may jeopardize function approximation. A close examination of the 1D plot shows that observations with large errors are inconsistent with the majority of other observations.

Thus, we need a way to detect these observations and deflate their influence. To this end, we use informal box plots [22] to pinpoint the outlying observations. Taking advantage we eliminate around 700 observations. The distance error distribution without outliers is plotted as frequency histogram. Some of the representative statistics that describe much of this distribution error are given in Table1.

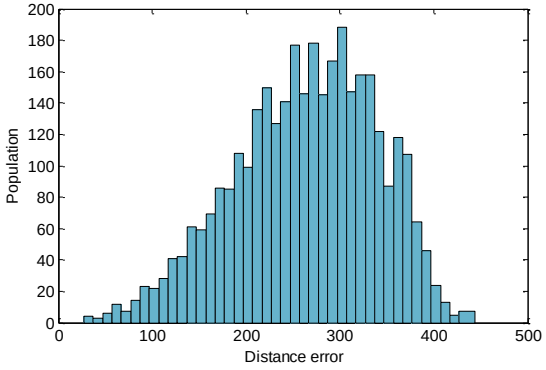


Fig. 5. Distance errors represented as a frequency histogram

TABLE I. ERROR STATISTICS FOR UNCALIBRATED SCALABLE-SPIDAR

Position error	***Scalable-SPIDAR***
Mean (mm)	263.7240
Standard derivation (mm)	75.6207
Maximum (mm)	444.2114

The other data preprocessing consideration is scaling. The task of training of regression algorithm is significantly simplified if data lie within a small range. We scale all inputs to have mean zero and standard deviation one.

$$\begin{cases} N_{S\_SPIDAR} = ((P_{S\_SPIDAR} - \Gamma_{S\_SPIDAR})./\sum_{S\_SPIDAR}) \\ N_{OTS} = ((P_{OTS} - \Gamma_{OTS})./\sum_{OTS}) \end{cases} \qquad (1.7)$$

Where

$\Gamma_{S\_SPIDAR} = \gamma_{S\_SPIDAR}.J$ and $\sum_{S\_SPIDAR} = \sigma_{S\_SPIDAR}.J$

$\Gamma_{OTS} = \gamma_{OTS}.J$ and $\sum_{OTS} = \sigma_{OTS}.J$

γ_{S_SPIDAR} : Matrix of means of components of the matrix of positions given by the Scalable-SPIDAR haptic device.

γ_{OTS} : Matrix of means of components of the matrix of positions given by the Optical Tracking System.

P_{S_SPIDAR} : Positions returned by the Scalable-SPIDAR haptic device.

P_{OTS} : Positions returned by the Optical Tracking System

J: Identity matrix

N_{S_SPIDAR} : Standardized positions of the Scalable-SPIDAR haptic device.

N_{OTS} : Standardized positions of the Optical Tracking System.

B. Calibration using Support Vector Regression

In the following section, we applied support vector regression (SVR) for calibrating the Scalable-SPIDAR haptic device. The convergence of support vector machine depends on the selection of a kernel function. Kernel functions projects the data into high dimensional feature space. This work uses the Gaussian kernel to perform mapping between Scalable-SPIADR and optical tracking system (OTS) data. In the following, trials Gaussian kernel were applied:

$$K(\vec{x}, \vec{x_i}) = \exp[\frac{-\left\|\vec{x} - \vec{x_i}\right\|^2}{2\tau^2}]$$

Where τ is a constant.

A search was performed for the most effective capacity parameter C to improve generalization accuracy of the regression technique. The capacity measures the flexibility or richness of regression functions and gives the protection against over fitting. In our experiments, the capacity was set to values between 30 and 100. Another parameter used in the training of support vector regression is epsilon, which checks

the insensitivity of the regression. The algorithm assumes that estimations that lie within epsilon distance of their true values are enough accurate. Epsilon was chosen to be equal to 0.05.

To assess the performances of the support vector regression technique, we plot the distribution of position errors but after calibration using support vector regression method. Figure 6 shows that SVR is quite efficient and exhibits lower errors in the overall workspace.

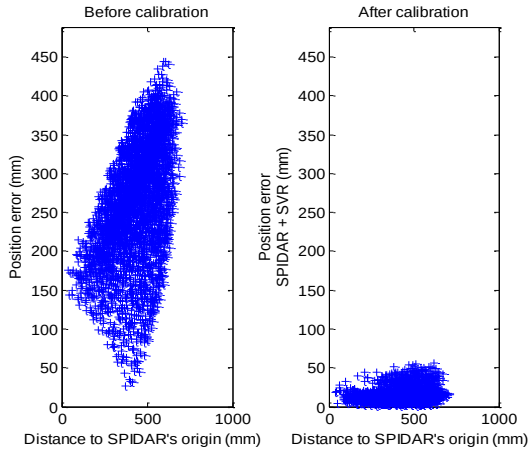


Fig. 6. Distribution of position errors as a function of distance to the Scalable-SPIDAR's origin (before and after SVR calibration)

We can see from the table below that the SVR calibration was successful in reducing the mean error in tracked position form 263.7240±75.6207 mm to 12.6045±8.4169 mm. The maximum possible error is about 57.5928 mm

TABLE II. ERROR STATISTICS BEFORE AND AFTER APPLYING SUPPORT VECTOR REGRESSION TECHNIQUE

Position error	***Raw***	***SVR***
Mean (mm)	263.7240	12.6045
Standard derivation (mm)	75.6207	8.4169
Maximum (mm)	444.2114	57.5928

After testing support vector regression technique with training data, we need to measure its ability to handle unseen data. The driving idea is to build a testing dataset. To this end, the grip of the Scalable-SPIDAR haptic device was moved in random trajectories of sequential points. These paths represent groups of data points within our predetermined working space. Following this, 512 corresponding point measurements of both the Scalable-SPIDAR and the IR optical tracker can be obtained, therefore allowing evaluating the generalization capability of our technique.

TABLE III. ERROR STATISTICS IN GENERALIZATION

Position error	***Raw***	***SVR***
Mean (mm)	267.2019	11.6649
Standard derivation (mm)	75.5131	3.8349
Maximum (mm)	436.1809	22.1889

Table 3 shows that SVR keeps convenient performance when handling unseen data. The mean and standard derivation error position values guaranteed by the SVR are good

VI. HYBRID TRACKING SYSTEM

Due to obstructions of line-of-sight, tracking data reported by the IR optical tracking could be incorrect which leads to a loss of tracking. As a sequence, the consistency between physical and virtual environments is affected, which is revealed by breaks in presence. To overcome this problem, we combine IR optical tracking system (OTS) with a human scale haptic device. The main goal is to maintain an interrupted tracking. In the following section, we provide details on our hybrid system.

Fig 7 illustrates of the pipeline of our hybrid system.

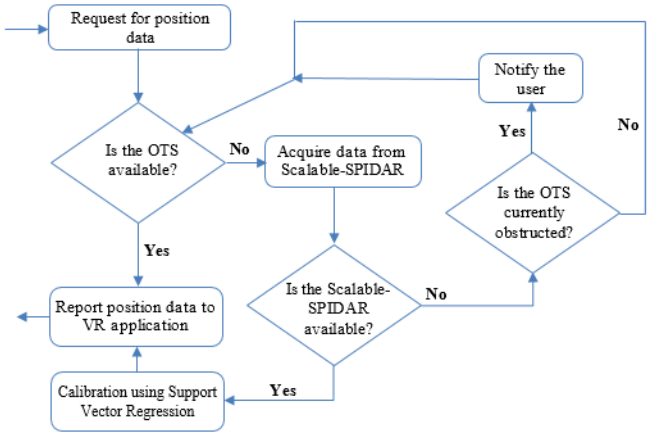


Fig 7. A flow diagram of the hybrid system

Once a request is being sent from the virtual reality application, the system checks whether it can provide data from the OTS. If this optical component loses track then the hybrid system requests position from the haptic component. In this case, the system checks the validity of data and calibrates the Scalable-SPIDAR using the Support vector regression technique. Switching between the two components depends on the optical tracking activity.

VII. CONCLUSION

In this paper, we have presented a hybrid tracking system for virtual reality applications. The use of support vector regression technique allows for compensation for nonlinear errors in Scalable-SPIDAR position readings. The results show that is possible to reduce the average error between the expected true position and the calibrated position from 263.7240 ±75.6207 mm to 12.6045 ± 8.4169 mm

These results encourage the development of a hybrid optical- large-scale haptic system for virtual reality applications where the haptic device acts as an auxiliary source of position information for the optical system.

REFERENCES

[1] G. Welch and E. Foxlin, "Motion tracking: no silver bullet, but a respectable arsenal," Computer Graphics and Applications, IEEE, vol. 22, no. 6. pp. 24–38, 2002.

[2] M. Frad, H. Maaref, S. Otmane, and A. Mtibaa, "SPIDAR calibration based on regression methods," Networking, Sensing and Control (ICNSC), 2014 IEEE 11th International Conference on. pp. 679–684, 2014.

[3] A. State, G. Hirota, D. T. Chen, W. F. Garrett, and M. A. Livingston, "Superior Augmented Reality Registration by Integrating Landmark Tracking and Magnetic Tracking," in Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques, 1996, pp. 429–438.

[4] T. Auer and A. Pinz, "Building a hybrid tracking system: integration of optical and magnetic tracking," Augmented Reality, 1999. (IWAR '99) Proceedings. 2nd IEEE and ACM International Workshop on. pp. 13–22, 1999.

[5] S. You, U. Neumann, and R. Azuma, "Hybrid inertial and vision tracking for augmented reality registration," Virtual Reality, 1999. Proceedings., IEEE. pp. 260–267, 1999.

[6] M. Aron, G. Simon, and M. O. Berger, "Handling uncertain sensor data in vision-based camera tracking," Mixed and Augmented Reality, 2004. ISMAR 2004. Third IEEE and ACM International Symposium on. pp. 58–67, 2004.

[7] W. Birkfellner, F. Watzinger, F. Wanschitz, R. Ewers, and H. Bergmann, "Calibration of tracking systems in a surgical environment," IEEE Transactions on Medical Imaging, vol. 17, no. 5. pp. 737–742, 1998.

[8] M. Harders, G. Bianchi, B. Knoerlein, and G. Szekely, "Calibration, Registration, and Synchronization for High Precision Augmented Reality Haptics," Visualization and Computer Graphics, IEEE Transactions on, vol. 15, no. 1. pp. 138–149, 2009.

[9] G. ART Advanced Realtime Tracking, "ARTtrack1 & DTrack - Manual."

[10] L. Buoguila, M. Ishii, and M. Sato, "Multi-Modal Haptic Device For Large-Scale Virtual Environment," ACM Multimed., pp. 277–283, 2000.

[11] Y. Hirata and M. Sato, "3-dimensional Interface Device For Virtual Work Space," Intelligent Robots and Systems, 1992., Proceedings of the 1992 lEEE/RSJ International Conference on, vol. 2. pp. 889–896, 1992.

[12] C.-C. Chuang, S.-F. Su, J.-T. Jeng, and C.-C. Hsiao, "Robust support vector regression networks for function approximation with outliers," IEEE Transactions on Neural Networks, vol. 13, no. 6. pp. 1322–1330, 2002

[13] V. Vapnik, Estimation of Dependences Based on Empirical Data: Springer Series in Statistics (Springer Series in Statistics). Secaucus, NJ, USA: Springer-Verlag New York, Inc., 1982.

[14] R. O. Duda and P. E. Hart, Pattern Classification and Scene Analysis. New Yotk: John Willey & Sons, 1973.

[15] A. J. Smola, B. Schölkopf, and K.-R. Müller, "The Connection Between Regularization Operators and Support Vector Kernels," Neural Netw., vol. 11, no. 4, pp. 637–649, 1998.

[16] Vladimir N. Vapnik, "The nature of statistical learning theory," IEEE Trans. Neural Netw., 1997.

[17] V. Vapnik, S. E. Golowich, and A. J. Smola, "Support Vector Method for Function Approximation, Regression Estimation and Signal Processing," in Advances in Neural Information Processing Systems 9 -- - Proceedings of the 1996 Neural Information Processing Systems Conference (NIPS 1996), 1997, pp. 281–287.

[18] A. Aizerman, E. M. Braverman, and L. I. Rozoner, "Theoretical foundations of the potential function method in pattern recognition learning," Autom. Remote Control, vol. 25, pp. 821–837, 1964.

[19] H. Drucker, C. J. C. Burges, L. Kaufman, A. Smola, and V. Vapnik, "Support vector regression machines," Adv. Neural Inf. Process. Dystems, vol. 1, pp. 155–161, 1997.

[20] a. J. Smola and B. Schölkopf, "On a Kernel-Based Method for Pattern Recognition, Regression, Approximation, and Operator Inversion," Algorithmica, vol. 22, pp. 211–231, 1998.

[21] A. J. Smola, B. Sch, and B. Schölkopf, "A Tutorial on Support Vector Regression," Stat. Comput., vol. 14, no. 3, pp. 199–222, 2004.

[22] J. ~W. Tukey, Exploratory Data Analysis. Reading, Mass.: Addison-Wesley, 1977.

Physiologically Motivated Feature Extraction for Robust Automatic Speech Recognition

Ibrahim Missaoui and Zied Lachiri
Signal, Image and Information Technology Laboratory
National Engineering School of Tunis, University of Tunis El Manar, BP. 37 Belvédère,1002 Tunis, Tunisia

***Abstract*—In this paper, a new method is presented to extract robust speech features in the presence of the external noise. The proposed method based on two-dimensional Gabor filters takes in account the spectro-temporal modulation frequencies and also limits the redundancy on the feature level. The performance of the proposed feature extraction method was evaluated on isolated speech words which are extracted from TIMIT corpus and corrupted by background noise. The evaluation results demonstrate that the proposed feature extraction method outperforms the classic methods such as Perceptual Linear Prediction, Linear Predictive Coding, Linear Prediction Cepstral coefficients and Mel Frequency Cepstral Coefficients.**

Keywords—Feature extraction; Two-dimensional Gabor filters; Noisy speech recognition

I. Introduction

Over the last years, numerous feature extraction methods have been developed for noise robust Automatic Speech Recognition (ASR) to improve performance and robustness of the recognition task. Several of these methods exploit the principles of speech processing of human speech perception to overcome the lack of robustness against the variability of speech signals. The traditional feature extraction methods such as Mel-frequency cepstral coefficients (MFCC) [1], Linear Prediction coding (LPC) [2] and Perceptual Linear Prediction (PLP) [3] were been based on the use of auditory filter modeling. Further improvements were made by using various auditory modeling in other methods [4][5][6].

Recent physiological and psychoacoustic studies have additionally shown that the primary auditory cortex neurons responsive to spectro-temporal modulations which referred as the Spectro-Temporal Receptive Fields (STRFs) have an important role in speech perception. Two-dimensional spectro-temporal Gabor filters have successfully used for modeling STRFs [7][8]. This has led to various extraction approaches of spectro-temporal features that achieve good performance in ASR noise robustness compared to traditional features [9][10][11]. In [12], Gabor features was obtained by processing a log Mel-spectrogram by a number 2D Gabor filters which were organized in a filterbank while these features were calculated from time-frequency representation derived from Power-Normalized Cepstral Coefficients (PNCCs) [15] in [16].

In this study, a physiologically motivated extraction method of Gabor features for noisy speech recognition is presented. The proposed method was based on the use of a set of 41 two-dimensional Gabor filters organized in a filter bank. It was applied to recognition of the TIMIT isolated words in the noisy environments. The recognition task is performed using Hidden Markov Models, which have been built using HTK toolkit [15].

This paper was organized as follows: Section 2 describes the proposed Gabor features extraction method. The experimental framework and results were detailed in section 3. Section 4 provides conclusions of this paper.

II. The Proposed Feature Extraction Based on Two-Dimensional Gabor Filters

A novel method based on two-dimensional Gabor filters is proposed to extract robust speech features for recognition of isolated speech words. The various steps were illustrated in Figure 2.

After pre-emphasizing the input speech signal, the power spectrum of signal is calculated by performing a windowing operation using a Hamming window (20 ms length with 10 ms overlap) and the square of Discrete Fourier Transform. It is then passed into a Bark-scale filter bank which aims to simulate the critical-band-masking curves, in order to obtain a critical-band power spectrum [3].

Subsequently, the equal loudness pre-emphasis and the intensity loudness conversion (third root amplitude compression) are performed to reproduce the two psychoacoustic properties of human hearing system; the non-equal sensitivity increase across frequency and the power law of hearing, which represents the simulation of the relation between the speech signal intensity and the perceived loudness of speech [3]. These two steps allow the reduction of spectral amplitude variation of the obtained spectrum.

Finally, the proposed features named as Gabor Bark Power Spectrum features or GBPS features were extracted by applying a set of two-dimensional Gabor filters organized in a filter bank to the representation of the obtained spectrum. This filterbank is composed of 41 two-dimensional Gabor filters [12]. These filters represent one of the most recent states of the art methods that were been successfully applied as front-end to noise robust speech recognition [12][16][18]. The Gabor features were obtained by calculating the 2D convolution of the filter and a time-frequency representation of speech to capture spectro-temporal modulations. Each two-dimensional Gabor filter is the product of two function terms: a complex sinusoid term denoted as $s(n,k)$ and a Hanning envelope $h(n,k)$ (with the time and frequency window lengths are W_n and W_k) [12][13][14].

$$s(n,k) = \exp\left(i\omega_n(n-n_0) + i\omega_k(k-k_0)\right) \quad (1)$$

$$h(n,k) = 0.5 - 0.5\cos\left(\frac{2\pi(n-n_0)}{W_n+1}\right)\cos\left(\frac{2\pi(k-k_0)}{W_k+1}\right) \quad (2)$$

The two terms ω_n and ω_k are time modulation frequency and the spectral modulation frequency. These terms determine the periodicity of the Gabor function and allow it to will be being tuned to a wide range of directions of spectro-temporal modulation.

The used bank of 41 Gabor filters were selected to get transfer functions of these filters having a constant overlap in the modulation frequency domain and covering a broad interval, which aimed to offer an approximated orthogonal filter and a limitation of redundancy of the filter output signal. The temporal and spectral modulation frequencies of the used bank of 41 Gabor filters were illustrated in Figure 1.

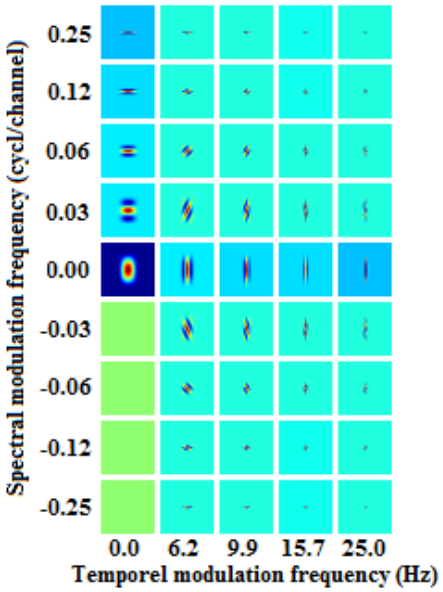


Fig. 1. The real components of a set of 41 Gabor filters employed in the proposed method

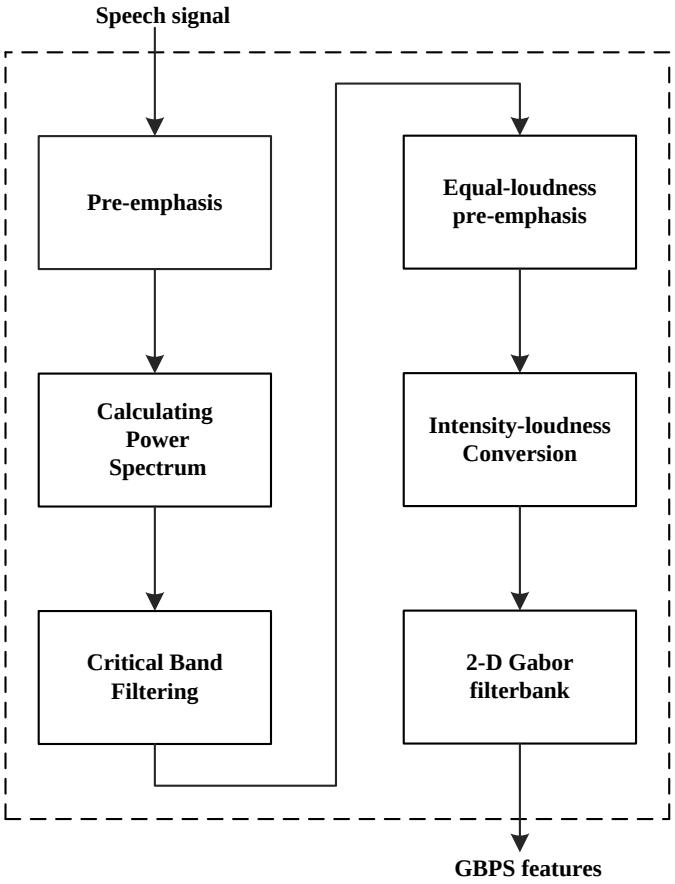


Fig. 2. Overview of the proposed feature extraction method based on two-dimensional Gabor filters

III. EXPERIMENTAL FRAMEWORK

A. The used Databases

The TIMIT database [19] was used for all ASR experiments reported in this paper. It is one of the standard databases used to evaluate the robustness and performance of any new method on an ASR task because it has a wide range of speakers and dialects. This database consists of speech signals with sampling frequency equal to 16 kHz of 630 (192 female and 438 male) different speakers from eight different major dialects of The United States, ten sentences spoken by each one of these speakers

In our experimental study, we used isolated words speech extracted from TIMIT database. A total of 9240 isolated speech words were exploited in the learning phase and 3294 isolated speech words were used for the recognition phase.

Furthermore, six background noises (restaurant, exhibition, babble, Car) drawn from the AURORA database [20] are used to evaluate the robustness of the proposed method under additive noise. The noisy isolated words used in this work were obtained by combining clean isolated words by each noise for various noise levels SNR.

B. The used Speech recognizer

The speech recognizer used in our experiments was based on HMM which have been built using the Hidden Markov Model Toolkit (HTK 3.4.1) [17]. This portable toolkit is developed by Cambridge University and used to construct and manipulate HMM optimized for speech recognition. An HMM is used to model a series of acoustic vectors. It represents a collection of stationary states which are connected by transition of Markov chain. At each state change, an observed acoustic vector o_t which described by an emitting probability distribution density $b_j(o_t)$ is generated. The transition between state s_i and state s_j is also probabilistic and has a discrete probability a_{ij} associated with it [21][22]. An example of an HMM consisting of five states with non-emitting entry and exit states is showed in Figure 3.

In the case of continuous density HMM, the most widely used output probability density $b_j(o_t)$ is the Gaussian mixture density which was defined as [17]

$$b_j(o_t) = \sum_{k=1}^{K} c_{jk} N(o_t; \mu_{jk}, \vartheta_{jk}) \quad (3)$$

Where $N(o_t; \vartheta_{jk}; o_{jk})$ is the multivariate Gaussian density with ϑ_{jk}, μ_{jk} and c_{jk} are the covariance matrix, the mean vector and weight associated with, the k^{th} Gaussian component at state j. "n" is the dimension of the vector o_t.

$$N(o_t; \vartheta_{jk}; o_{jk}) = \frac{1}{\sqrt{(2\pi)^n |\vartheta_{jk}|}} e^{(-\frac{1}{2}(o_t-\mu_{jk})^T \vartheta_{jk}^{-1}(o_t-\mu_{jk}))} \quad (4)$$

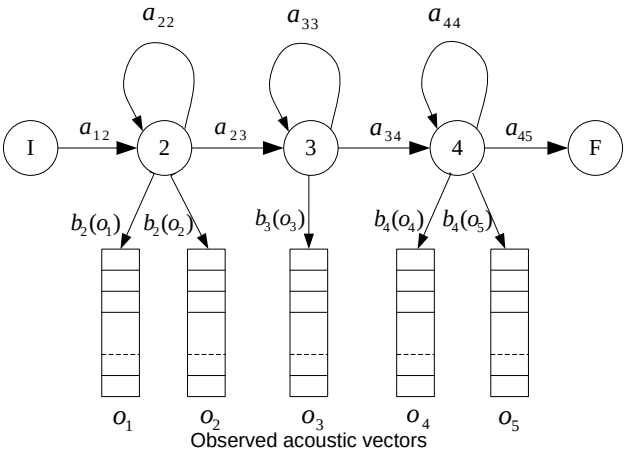


Fig. 3. Illustration of Hidden Markov models with five left-to-right states

The HMM topology exploited in our experiments is the left-to-right five-state HMM with Gaussian Mixture density and diagonal covariance matrix. Each HMM state is represented by four Gaussian Mixtures (HMM-4-GM).

C. Results and discussion

For all of our experiments, the proposed Gabor Bark Power Spectrum features or GBPS features are compared to four classic features combined with energy (E) such as Perceptual Linear Prediction (PLP_E), Linear Predictive Coding (LPC_E), Linear Prediction Cepstral coefficients (LPCC_E) and Mel Frequency Cepstral Coefficients (MFCC_E).

TABLE I. THE RECOGNITION RATE OF THE PROPOSED FEATURES, MFCC, PLP, LPC, AND LPCC OBTAINED USING HMM-4-GM IN THE RESTAURANT NOISE CASE

Restaurant noise	**features**				
SNR level	***GBPS features***	***PLP_E***	***MFCC_E***	***LPCC_E***	***LPC_E***
0 dB	48.15	15.24	14.63	15.42	14.63
5 dB	75.41	31.88	31.12	27.35	18.73
10 dB	91.26	60.35	60.53	48.15	27.41
15 dB	94.35	80.94	81.51	72.50	41.59
20 dB	95.60	88.77	89.13	82.91	54.07
25 dB	95.96	91.04	92.11	87.13	61.87

TABLE II. THE RECOGNITION RATE OF THE PROPOSED FEATURES, MFCC, PLP, LPC, AND LPCC OBTAINED USING HMM-4-GM IN THE EXHIBITION NOISE CASE

Exhibition noise	**features**				
SNR level	***GBPS features***	***PLP_E***	***MFCC_E***	***LPCC_E***	***LPC_E***
0 dB	44.44	5.65	6.34	5.98	5.04
5 dB	71.98	16.58	18.09	12.48	7.95
10 dB	88.37	38.49	18.09	30.42	17.30
15 dB	93.69	55.83	58.32	47.33	24.89
20 dB	95.23	73.95	74.13	62.96	33.24
25 dB	95.87	84.58	86.00	78.96	44.23

TABLE III. THE RECOGNITION RATE OF THE PROPOSED FEATURES, MFCC, PLP, LPC, AND LPCC OBTAINED USING HMM-4-GM IN THE BABBLE NOISE CASE

Babble noise	**features**				
SNR level	***GBPS features***	***PLP_E***	***MFCC_E***	***LPCC_E***	***LPC_E***
0 dB	45.87	18.94	18.94	15.66	13.36
5 dB	69.73	35.22	36.04	26.62	17.58
10 dB	87.80	60.60	60.50	50.00	24.74
15 dB	93.87	81.24	81.88	74.38	41.20
20 dB	95.29	88.92	89.53	83.58	53.64
25 dB	95.75	91.17	91.77	87.95	61.60

TABLE IV. THE RECOGNITION RATE OF THE PROPOSED FEATURES, MFCC, PLP, LPC, AND LPCC OBTAINED USING HMM-4-GM IN THE CAR NOISE CASE

Car noise	**features**				
SNR level	***GBPS features***	***PLP_E***	***MFCC_E***	***LPCC_E***	***LPC_E***
0 dB	49.73	11.63	13.11	14.85	8.23
5 dB	72.19	20.16	21.40	20.67	12.57
10 dB	89.74	37.37	38.40	37.28	24.32
15 dB	94.02	60.23	60.99	55.43	35.40
20 dB	95.39	80.66	82.48	73.62	42.53
25 dB	95.87	88.95	89.95	84.06	51.21

The result rates of recognition experiments with proposed Gabor features and the four classic features obtained using HMM-4-GM are summarized in the Tables I, II, III, and IV. Six noises (restaurant, exhibition, babble and car noises) drawn from the AURORA database and six specific signal-to-noise ratios (SNR) ranging from 0 dB to 25 dB in 5 dB steps were considered.

As illustrated in these tables, the proposed Gabor features outperform PLP_E, LPC_E, LPCC_E and MFCC_E features in the different cases. It can be observed that the highest percentage of the recognition rates is obtained using our Gabor features at almost all SNR levels, particularly at low SNR values. For example, in the car-noise case at SNR equal to 5 dB, the recognition rate of our Gabor features is higher than that of PLP_E, LPC_E, LPCC_E and MFCC_E features by 52.03, 59.62, 51.52 and 50.79 respectively. As can also be seen in the different tables, when decreasing the value of SNR level, the performance of all features degrade, but the proposed features remain robust and more performing than the classic features.

IV. CONCLUSION

A new physiologically motivated feature extraction method based on Gabor filterbank for isolated-word speech recognition under noisy conditions is presented in this paper. The proposed method takes into consideration the extraction of spectro-temporal modulation frequencies and the limitation of the redundancy on the feature level. The robustness of our Gabor Bark Power Spectrum features or GBPS features was evaluated on isolated speech words taken from TIMIT database using HMM. The obtained results show that our Gabor features have given the best results at all SNR levels compared to four classical features combined with energy: PLP_E, LPC_E, LPCC_E and MFCC_E features.

REFERENCES

[1] S.B. Davis, and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences", IEEE T Acoust Speech, vol. 28, pp. 357–366, August 1980.

[2] D. O'Shaughnessy, "Linear predictive coding", IEEE Potentials, vol. 7, pp. 29–32, February 1988.

[3] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech", J Acoust Soc AM, vol. 87, pp. 1738–1752, April 1990.

[4] R.P. Lippmann, "Speech recognition by machines and humans", Speech Commun, vol. 22, pp.1–15, July 1997.

[5] B.T. Meyer, "Kollmeier B. Robustness of spectro-temporal features against intrinsic and extrinsic variations in automatic speech recognition", Speech Commun., vol.53, pp. 753–767, May 2011.

[6] Y. Zouhir, and K. Ouni, "A bio-inspired feature extraction for robust speech recognition", SpringerPlus, vol. 3, pp.651, November 2014.

[7] N. Mesgarani, and S. Shamma, "Speech processing with a cortical representation of audio", IEEE International Conference on Acoustics, Speech and Signal Processing; 22-27 May 2011; Prague, Czech Republic: IEEE. pp. 5872–5875.

[8] N. Mesgarani, S. David, and S. Shamma, "Representation of phonemes in primary auditory cortex: how the brain analyzes speech", IEEE International Conference on Acoustics, Speech and Signal Processing, 15-20 April 2007; Honolulu, Hawaii, USA: IEEE. pp. 765–768.

[9] M. Kleinschmidt, and D. Gelbart, "Improving word accuracy with Gabor feature extraction", International Conference on Spoken Language Processing; 16–20 September 2002; Denver, Colorado, USA: ISCA. pp. 25–28.

[10] H. Lei, B.T. Meyer, and N. Mirghafori, "Spectro-temporal Gabor features for speaker recognition", IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP; 25-30 March 2012; Kyoto, Japan: IEEE. pp. 4241–4244.

[11] S.V. Ravuri, and N. Morgan, "Using spectro-temporal features to improve AFE feature extraction for ASR", Proceedings of Annual Conference of the International Speech Communication Association INTERSPEECH, 26-30 September 2010; Makuhari, Chiba, Japan: ISCA. pp. 1181–1184.

[12] M.R. Schädler, B.T. Meyer, and B. Kollmeier, "Spectro-temporal modulation subspace-spanning filter bank features for robust automatic speech recognition", J Acoust Soc AM, vol. 131, pp. 4134–4151, May 2012.

[13] C. Kim, and R.M. Stern, "Feature extraction for robust speech recognition using a power law nonlinearity and power-bias subtraction", Proceedings of Annual Conference of the International Speech Communication Association INTERSPEECH; 6–10 September 2009; Brighton, United Kingdom: ISCA. pp. 28–31.

[14] I. Missaoui, and Z. Lachiri, An Extraction Method of Acoustic Features for Speech Recognition, Res. J. Appl. Sci. Eng. Technol, vol. 12, no. 9, 2016.

[15] I. Missaoui, and Z. Lachiri, Histogram equalization based front-end processing for noisy speech recognition, Journal of Theoretical and Applied Information Technology, 2016. in press.

[16] B.T. Meyer, and C. Spille, B. Kollmeier, and N. Morgan, "Hooking up spectro-temporal filters with auditory-inspired representations for robust automatic speech recognition", Proceedings of Annual Conference of the International Speech Communication Association INTERSPEECH, 9-13 September 2012; Portland, Oregon, USA: ISCA. pp. 1259–1262.

[17] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, The HTK book (Revised for HTK version 3.4.1). Cambridge University Engineering Department, 2009.

[18] B.T. Meyer, S.V. Ravuri, M.R. Schädler, and N. Morgan, "Comparing Different Flavors of Spectro-Temporal Features for ASR", Proceedings of Annual Conference of the International Speech Communication Association INTERSPEECH; 27-31 August 2011; Florence, Italy: ISCA. pp. 1269–1272.

[19] J.S. Garofolo, L.F. Lamel, W.M. Fisher, J.G. Fiscus, and D.S Pallett, "TIMIT acoustic-phonetic continous speech corpus CD-ROM", NIST speech disc 1-1., NASA STI/Recon Technical Report N 93, 27403, 1993.

[20] H. Hirsch, and D. Pearce, "The Aurora Experimental Framework for the Performance Evaluation of Speech Recognition Systems Under Noisy Conditions", Proceedings of the ISCA Workshop on Automatic Speech Recognition: Challenges for the New Millennium, 18–20 September 2000; Paris, France: ISCA. pp. 181–188.

[21] Y. Ephraim, and N. Merhav, "Hidden markov processes", IEEE T Inform Theory, vol. 48, pp.1518–1569, June 2002..

[22] L.R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition", P IEEE, vol. 77, pp. 257–286, February 1989.

An Informational Model as a Guideline to Design Sustainable Green SLA (GSLA)

Iqbal Ahmed
Graduate School of Science and Engineering, Saga University, Japan

Hiroshi Okumura
Graduate School of Science and Engineering, Saga University, Japan

Kohei Arai
Graduate School of Science and Engineering, Saga University, Japan

***Abstract*—Recently, Service Level Agreement (SLA) and green SLA (GSLA) becomes very important for both the service providers/vendors and as well as for the users/customers. There are many ways to inform users/customers about various services with its inherent execution functionalities and even non-functional/Quality of Service (QoS) aspects through SLAs. However, these basic SLAs actually do not cover eco-efficient green issues or IT ethics issues for sustainable development. That is why green SLA (GSLA) already came into play for achieving sustainability in the industry. Nevertheless, the current practice of GSLA in the industry do not respect the sustainability at all. GSLA defined as a formal agreement incorporating all the traditional commitments respecting some green computing parameters such as carbon footprint, energy consumption etc. Therefore, there are still gaps for achieving sustainability through existing GSLA. To reach the goal of achieving sustainability and getting more customers, many IT (Information Technology) and ICT (Information and Communication Technology) business are looking for a real GSLA which would meet the ecological, economical and ethical aspects (3Es) of sustainability. This research discovers the missing parameters and introduce new parameters under sustainability hoods. In addition, it defines GSLA of sustainability with new green performance indicators and their measurable units. It also discovers the management complexity of proposed new GSLA through designing a general informational model and identifies various new entities and their effects with other entities under three pillars of sustainability. The ICT engineer could use the informational model as a guideline to design a sustainable GSLA for the industry. Therefore, the proposed model could help different service providers/vendors to define their future business strategies for the upcoming sustainable society.**

***Keywords*—SLA; GSLA; Green Computing; Sustainability; IT ethics; Informational model**

I. INTRODUCTION

Nowadays, cloud and grid computing and many data centers acts as most promising service providers. These computing and communication industry provides different services in compare to traditional computing with some scalability benefits. At the same time, cloud services are offered at various levels: Infrastructure, Platform and Software as a Service [1, 2]. At each level, they maintain a SLA with respect to their parties. SLA is defined as a formal document between an IT service provider and one or more customer outlining Service Commitment [3]. The main issue is that most of these traditional/basic SLA actually do not cover eco-efficient green issues. The growth rate of SLA in recent time is increasing as well as the need of GSLA for actual sustainability achievement in IT industry [4]. Presently, the revolution of ICTs and ITs in daily average life has also resulted in the increase of Green House Gas (GHG), due to continual increase in global "carbon footprint''. In 2007, the ICT sector produced as much GHG as the aero industry and is projected to grow rapidly [5, 6]. If ICT has a negative impact on environment, it can be also be used for greening the other human activities (logistic, city, industry etc) in this new society. Indeed, the dimensions of green informatics and green computing contributions are: the reduction of energy consumption, the rise of environmental awareness, the effective communication for environmental issues and the environmental monitoring and surveillance systems, as a means to protect and restore natural ecosystems potential [7]. At the same time, many IT and ICT industries or service providers need to think about their business scope in the light of green perspective to achieve sustainability. However, the IT and ICT sectors mostly concern about energy or power consumption, carbon, recycling and productivity issues under greening computing lens whereas the practicing of sustainability is still far away from reality. In addition, most of the recent IT and ICT industries overlooked many green parameters under sustainability lens due lack of proper guidance to identify new parameters. Therefore, with the increase attention that green informatics and sustainability practice within our society, it is timely to not only conduct SLAs for traditional/basic computing performance metrics or only on energy or carbon footprint issues, but also to relate the effort of conducting green computing with respect to 3Es of (Ecology, Economy and Ethics) sustainability pillars [4]. Therefore, the journey of sustainable GSLA is getting importance in ICT business world. This research digs down for finding more new green performance indicators for developing a new sustainable GSLA under three pillars. In addition, a new GSLA proposed, which covers all the existing green performance indicators as well as some other missing indicators covering three pillars of sustainability. Finally, GSLA research demonstrates the management complexity of interactions between all the performance indicators using an informational model and also tries to analyze all relationships and different level of effects for most of the green parameters to achieve sustainable GSLA.

The rest of the work is organized as follows- the next background section discusses briefly about existing theory and practical works on basic SLA and GSLA's form different service providers. The existing GSLA subsection actually shows currents trends of the industry to practice sustainability under greening lens. The identification of new green

indicators section discovers the most important missing performance indicators for future sustainable GSLA under 3Es of sustainability pillars. Moreover, defining new GSLA using informational model section helps the ICT engineers to define and manage future sustainable GSLA with newly identified missing green parameters to respect true sustainability. Finally, the conclusion gives brief discussion about few challenges and future plan of this sustainable GSLA research.

II. Background Study

The details literature review and analysis are based on existing work in the field of SLA, GSLA, green computing, energy optimization in IT industry, impact of ICT on environment and natural resource, IT ethics issues, IT for Sustainability etc. In the findings, GSLA research divides its work based on basic SLA and then existing GSLA for different types of services from their providers. The rigorous literature review and their details analysis for SLAs found in [4].

A. Basic SLAs

In the findings on existing empirical work, firstly this research splits its outcomes based on basic SLAs for different types of domain from various providers such as Network, Compute, Storage and Multimedia [8]. Most of the performance indicators in basic SLA sections are quantitative parameters and they are simple to evaluate, control and monitor. Thus, it is easy to respect most of the basic performance indicators from both customer and providers side. In addition, there is no eco-efficient green parameters included in basic/traditional SLA parameters.

SLA for Network, Compute, Storage and Multimedia domain:

The basic SLAs for network specifies service level commitments which are applied to measure and evaluate network performance and give proper support for their clients. Usually, from different network service provider, the following performance indicators found in their SLAs are [4]- *Network Availability, Delay, Latency, Packet Delivery Ratio, Jitter, Congestion, Flow Completion time, Response time, Bandwidth, Utilization, MTBF (Mean Time Between Failure), MTRS (Mean Time to Restore Services), Solution time, Resolution time, LAN/WAN period of operation, LAN/WAN Service Time, Internet access across Firewall, RAS (Remote access Services).* Some indicators like *Bandwidth, Utilization, and Congestion* are related to link capacity whereas *Availability, Delay, Jitter, Response Time* etc. associated with time related information for different network service providers.

Most the cloud, grid service companies provides computing service to their consumers. The basic SLA parameter and their measurement unit for computing domains are [4], -: *Broad Network Accessibility, Multi-tenancy, Rapid Elasticity, Scalability, Resource Pooling Time, Solution Time, Response Time, Availability (MTBF & MTTR), Capacity, Virtualization, Delay, Resolution Time and Logging & Monitoring.* Here, *Broad Network Accessibility, Multi-tenancy* and *Logging & Monitoring* are informative indicators presented in their SLAs.

The storage domains are typically handled by cloud storage provider. Interestingly, today's cloud storage SLAs just ensure uptime guarantee but not data availability and data protection. In some case, traditional SLAs just mention about data storage security and backup but there is no proper authority or standard to check their commitments. Some common basic SLA performance indicator [4] for storage services are as follows-: *Availability, Response Time, Maximum Down Time, Uptime, Failure Frequency, Period of Operation, Service Time, Accessibility, Backup, Physical Storage Backup, Transportation for Backup, Size, Data Accessibility, Security.*

Multimedia service domain SLAs are classified into three broad application areas- Audio, Video and Data. It is challenging to monitor and evaluate some qualitative indicator such as *Mean Opinion Score (MOS)* and *Lip Synchronization* for one-way video, conferencing or in videophone. Most of the SLA indicators for multimedia domain for different applications are *Information Loss (PLR), Jitter, One-way Delay, MOS, Lip Synchronization,* and *Security Policy* [4].

B. Existing GSLAs

Currently, several IT and ICT industries, cloud providers provide their GSLAs under green computing practice. Here, recent GSLAs are mainly focused only on energy/ power, carbon footprint, green energy, recycling issues. Additionally, several existing GSLA also demonstrates their productivity issues with necessary monitoring unit. In addition, various research draws attention only on minimizing energy consumption while improving networking performance on wireless connection under green computing hood [9, 10]. All these performance indicators (Table I) help various service providers and consumers either to design or to choose services mainly with respect to energy consumption, renewable energy usages, carbon emission issues and productivity issues in recent time. However, the IT industry needs to find out new parameters/indicators for achieving sustainability as current trends of the society shows that people are much more concerned about sustainability in this scope.

Table I depicts the performance indicators and their unit for different services considering green computing practices. The table has several headings. *Green Computing Domain* mentions the category of green computing practices in IT industry; *Performance Indicator Name* is the notion which used an evaluating, monitoring metric for defining performance in GSLAs, and then their measurable unit as *Unit* column.

TABLE I. Performance Indicator for Different Services Considering Existing Gsla [4]

Green Computing Domain	Performance Indicator Name	Unit
	Total Power Consumption	kW-h (Kilowatt-hour)
	PUE (Power Usages Effectiveness)	Number (1.0 to ∞) Or Dimensionless
	DCiE (Data Center Infrastructure Efficience)	% (Percentage)
	CPE (Compute Power Efficiency)	Watts

Energy/ Power	SPECPower	Watt
	JouleSort	kW/J
	WUE (Water Usages Effectiveness)	Liter/kW-h
	TDP (Thermal Design Power)	Watts
	ERF (Energy Reuse Factor)	Number [0 to 1.0]
	ERE (Energy Reuse Effectiveness)	Number [0 to ∞]
	GEC (Green Energy Co-efficient)	Number [0 to 1.0]
	ITEE (IT Equipment Energy Efficiency)	% (Percentage)
	ITEU (IT Equipment Utilization)	Number
	HVAC (Heating, Ventilation, Air-conditioning) Effectiveness	Dimensionless
	Cooling System Efficiency	kW/ton
Carbon footprint	CUE(Carbon Usages Effectiveness)	KgCO2 per kW-h
	DPPE (Data Center Performance Per Energy)	Number [0 to 1]
Recycling	e-Wastage Or IT Wastage	Gm (Gram)
	Recycling	% (Percentage)
Productivity	DCP (Data Center Productivity)	Not Available
	DCeP (Data Center Energy Productivity)	Not Available
	Analysis Tool	Not Known
	EnergyBench	Numeral Rating
	ScE (Server Compute Efficiency)	% (Percentage)
Costing Information	Energy/Power Cost	Currency [according to country]
Others	SWaP (Space, Wattage and Performance)	Not Available
	Air Management Metric	F (Fahrenheit)
	UPS System Efficiency	% (Percentage)

III. Identification of New Green Indicators

Fig.1 depicts the overall idea to define new sustainable GSLA as well as gives the idea of this research. In current focus, the traditional performance based parameters and few green parameters resides together to provide current GSLA in the industry. In existing GSLAs, most of the performance indicators mainly concentrate on energy consumption issues and productivity concern in cloud and grid computing industry (Table I). In addition, the ICT engineers could easily evaluate and monitor these parameters at hardware level or software level. However, the existing GSLA do not consider recycling, radio wave, toxic material usage, noise, light pollution for sustainable development. Moreover, people's interaction and IT ethics issues, such as user satisfaction, intellectual property right, user reliability, confidentiality etc are also missing in current GSLA. Therefore, the new focus part discovers the concepts of 3Es relationship with current GSLA, which could be used as a guideline for the ICT engineer to design and respect all the parameters of sustainable GSLAs. Next section discusses the proposed new performance indicators of GSLA for achieving sustainability from 3Es perspectives (Ecological, Economical and Ethical).

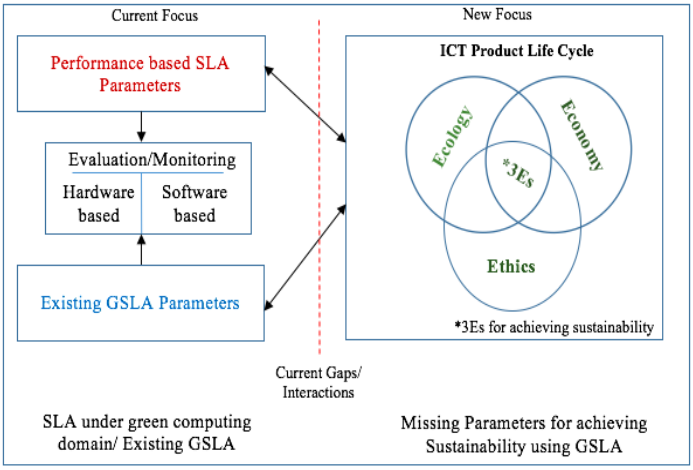


Fig. 1. Overview of GSLA (Green SLA) for achieving Sustainability

A. Ecological Point

Recycling- The recycling of ICT equipment impose into their whole life cycle. This is a very complex indicator and need to be sub divide as reuse, refurbish, sub-cycling and up cycling. According to [8,11], the Recyclability Rate of an equipment ranges from 0 to 1. Again, at each stage of recycling, it needs to be considers the *CUE*, *GEC*, *Energy Cost* (Table I) because recyclability includes energy consumption and carbon emission simultaneously. Recycling has direct relationship with eWastage and pollution as it helps to reduce global magnitude of e-waste. On the other hand, if recycling procedure is not well managed then it could pollute air, water and soil. Recycling information should put into new GSLA according to government laws, directives such as Waste Electrical and Electronic Equipment (WEEE) Directive (2012/19/EU) by European Union (EU). There are also some voluntary recycler standards in US like *e-Steward and Responsible Recycling (R2) Practices*.

Toxic Material Information- Electric and Electronic products contain several toxic materials such as Beryllium, Cadmium, Lead, and Mercury etc. These chemical elements and their compound both cause serious health hazards and also make environment polluted [12, 13]. Beryllium is used in manufacturing computer motherboard and is acutely & chronically toxic to humans mainly affecting their lungs [12]. Cadmium and its compounds is used in some switches, many laptop's batteries and in some older CRTs monitor as phosphor coating. These materials and its compounds are also toxic to human, which affects kidneys in the long run [12]. Lead is usually used for primary electric solder on printed circuit boards. Lead could damage to the nervous system and blood system in human body [12, 13] and also causes severe air pollution. In some switching devices and batteries, mercury could be used which is highly toxic. Mercury has a high level impact on human nervous system [12, 13]. All these toxic material should have a safety limit and needs to be defined or restricted by third party or governing body such as Directive on the Restriction of the Use of Certain Hazardous Substances in Electrical and Electronic Equipment (2002/95/EC) from EU commonly known as RoHS Directive. The information about

the usage of these toxic materials in IT and ICT field should be stated clearly for making SLA greener.

Obsolescence Indication- The services, process, product or technology used or produced by a company for profit will become obsolete after certain period. Therefore, it is a matter of urgency in ICT industry to indicate or label product's life time with obsolescence indicators [14] according to product's raw materials scarcity, demands, usages limit etc at different stage of product's life cycle. These indicators should be stated into new GSLA to create awareness for both customer levels and company levels for achieving sustainability. It might be complex to indicate or determine the obsolescence of ICT equipment because it depends on different variables associated with equipment's production cost, raw material scarcity, energy issues and user's interaction. Additionally, Optimum Obsolescence [15] or obsolescence management model [16] might help to decide when a product needs to be reused, recycled or land filled. There is no standard to indicate this parameter in SLAs till now but it might be related with product life cycle costing, recyclability rate indicators as well.

Radio Wave Information- The electromagnetic radiation emitted by electronic equipment in IT industry, is a controversial topic in scientific community. The health effects of radio waves were also studied and most of these studies found that the EMF (electromagnetic field) effects on the human body are not only depends on their field level but on their frequency ranges and energy [8, 11]. All studies claim that the unique non controversial effect of non ionizing EMF is thermal [11]. To avoid this electromagnetic effect, the government of each country defines maximum level of EMF generated by wireless antenna and their maximum *Specific Absorption Rate (SAR)* value [8]. The EMF levels and safety used following measuring units. *Gauss (G), Tesla (T)* for EMF values; *Gray (Gy)* and *Sievert (Sv)* for measuring radiation effects on human tissues [17]. These radio waves information should state in GSLA according to government's defined level clearly and precisely.

Noise Pollution- The network engineer who works in Data Centre might need guidelines and regulations to control noise pollution in his/her workplace. The noise generated from data center causes hearing loss permanently [18]. OSHA and NIOSH- these two US government agencies look after the limit of noise level in work places. The noise pollution level might be stated on a GSLA using decibel (dB) measuring unit. Moreover, the noise created by ICT equipment such as Ringtone of a cell phone might also responsible for some sort of pollution as it become disturbing and irritating for other peoples. This type of pollution might be subjective and easily prevented by increasing awareness among the cell phone users.

Visual Pollution- The aesthetic aspects of ICT industry, for example- installing an antenna in a beautiful landscape or on a roof top. This could create hypersensitivity affect [8] and these might be very much subjective to human being such as Perception of Affective Quality (PAQ) [19] is an individual's perception of an object's ability to change his/her neurophysiologic states as feeling either good or bad.

Light Pollution- Computer Screen generates light pollution affecting health [8]. According to American Optometric Association, Computer Vision Syndrome (CVS) causes headache, blur, dry eye, eyestrain, sleep disorder etc [20]. The safe computing practice and awareness might help to decrease CVS. There is still no standard or measurable unit for light pollution level but it should be mentioned in proposed new GSLA. The next Table II demonstrates the new GSLA indicators from ecological point of view and their proposed measurable units.

TABLE II. GSLA PROPOSAL UNDER ECOLOGICAL PILLAR OF SUSTAINABILITY

Performance Indicator Name		**Description**	**Unit**
Recycling Rate (RR)	Reuse	Amount of ICT product reuse/ percentage of ICT equipment refurbished/ percentage of IT equipment sub cycled or up cycling;	gm (gram) OR % (Percentage) OR RR[0~1] [8]
	Refurbish		
	Recycle		
Toxic material limit/ Toxic material Usage Level		Information about using toxic material in ICT product and their limit level;	Preferred/ Acceptable
Obsolescence Indication Labeling		Indication about the perfect time to change an ICT equipment;	Labeling according to laws
EMF Level/ Radiation Effect Level		Amount of electromagnetic energy radiation; usually the strength is measured by frequency;	T (Tesla) / G (Gauss) OR Sv (Sievert) / Gy (Gray)[17]
Noise Pollution Level		The noise emitted from ICT equipment e.g. Ringtone of Cell phone, noise in data center;	µdB/dB (micro decibels)
Visual Pollution Level		The aesthetic aspect of ICT industry e.g. installing an antenna in a beautiful landscape or roof top;	Subjective OR *PAQ* [19]
Light Pollution Level		The light pollution generated by ICT equipment e.g. Computer Screen;	Subjective

B. Economical Point

Carbon Taxation- A number of countries has implemented carbon taxes [21] or energy taxes and *Cap and Trade System* [22] that is very much effective to reduce Green House Gas (GHG) emissions while stimulating technological innovation and economic growth. The taxation may create political or social unrest in some cases; therefore, it may be difficult to impose. In 1990s, a carbon/energy tax was proposed at the EU level but failed due to industrial lobbying but in 2010 the European Commission implemented a Pan-European minimum tax on pollution under the European Union Greenhouse Gas Emissions Trading Scheme (EU ETS) [21] which is quite successful. According to this new plan, 4 to 30 euro would be charged per ton of carbon emission. On the other hand, in US, the *Cap and Trade* gave more assurance to decline GHG emission and also has some political advantages [22]. Therefore, according to different country's economic, social or political culture, carbon taxation or *Cap and Trade* policy should need to be established and this information need to put into the new GSLA.

Civil Engineering Cost- The cost of civil engineering includes building cost, antenna setup cost, digging tranches for

cabling etc. The building costing also need to consider designing cost, manufacturing cost, renovation cost and finally dismantling cost of an IT facility or data center. All these costing information should come into proposed new GSLA. The cost of civil engineering is also associated with carbon emission indicators in each step. It is important to note here that; the new green datacenter have an environmental impact in their lifespan. For example, most of the green datacenter uses natural resources (air, water) for cooling purpose but also at the same time it dissipates heat directly to the atmosphere, which might create imbalance in the surrounding eco-system of that datacenter.

Table III showed the economic performance indicators and their measuring unit for evaluating new GSLA.

TABLE III. GSLA PROPOSAL UNDER ECONOMIC PILLAR OF SUSTAINABILITY

Performance Indicator Name		Description	Unit
Carbon Tax		Tax for carbon content on fuel in most case; this should be charged according to government laws;	Currency (dollar)
Civil Engineering Cost		Information about costing related building, antenna installation, digging for cabling etc.;	Currency (dollar)
Cooling Cost		Amount of cooling cost in a data center or percentages of renewable energy usage for cooling;	Currency (dollar)
ICT Product Life Cost	Manufacturing	Considering the whole life cycle of an ICT product and their costing; LCA assessment need to consider here;	Currency (dollar)
	Purchasing		
	Delivery		
	Operational		
	Dismantling		

Cooling Cost- The cooling system costing information need to be mentioned into the new GSLA. It includes energy (electric power, renewable energy) costing, infrastructure (humidity, temperature monitoring) and transportation costing for cooling the whole site. This indicator might become complicated because of *HVAC, Air Management Metric* and *Cooling System Efficiency* indicators in existing GSLA (Table I) and these might need to define newly.

ICT Product Life Cost- ICT product life costing consider the whole life cycle of a product cost including mainly manufacturing from raw materials, purchasing, delivery, operational and dismantling. Operational cost has association with utility cost such as energy and maintenance costing and dismantling cost also has association with recycle or refurbishment costing. Again, the life cycle assessment LCA [23] need to be considered in this parameter. ICT Product life cost indicators, thus become very complex to assess and monitor in GSLA.

C. Ethical Point

Mostly, the green computing practice focuses on the ecological, economical point but usually neglect human's interaction and ethical aspects [8]. The use of ethics in IT and ICT field covers many indicators such as *Satisfaction level, Intellectual Property Right, Reliability, Confidentiality, Security and Privacy, Gender/Salary/Productivity Information*. All of these indicators are usually subjective and informative, thus making new GSLA assessment difficult. For example, *Customer Satisfaction Index (CSI)* could be used for evaluating satisfaction level of a customer through designing and analyzing survey. In addition, User satisfaction could be rated from 0 to 5, where 0 indicates worst level of satisfaction and 5 is the preferred level. Moreover, the ICT Company should analyze social responsibilities towards *Customers, Employee and Community* [8, 24].

Table IV gives the idea of these responsibilities as performance indicators with respect to ethics for greening SLA to achieve sustainability.

TABLE IV. GSLA PROPOSAL UNDER ETHICS PILLAR OF SUSTAINABILITY

Performance Indicator Name		Description	Unit
Satisfaction level [Customer, Employee, Community]		Whether the customer, employee and community are satisfied with; [usually defined by third party or community]	*CSI* Rating [25,26,27]
Intellectual Property Right [Customer, Employee, Community]		IPR means copyright, patents of user's data; no hacking; royalty etc. ;	YES/NO
User Reliability		Whether customer reliability preserved by the company ; reliability between employee and company;	Test based Rating
Confidentiality		Information should be kept confidentially and also available for customer, employee or for community;	Test based Rating
Security & Privacy	Authentication & Authorization	Rules regarding security and privacy should clearly state and defined or not; usually it could be defined third party or government law.	High / Medium / Low OR Preferred/ Acceptable
	Access Control & Privilege Management		
	Data Geographic		
	Data Integrity		
	Transparency		
	Physical Security		
	Termination Management		
Gender Balance Information (only industry oriented)		The information about gender balance in an organization;	YES/NO
Salary Balance Information (only industry oriented)		The salary balance of an organization in IT industry;	YES/NO

IV. DEFINING GSLA USING INFORMATIONAL MODEL

In the previous section, this research found most of the important performance indicators with respect to three pillars of sustainability and this will definitely help ICT and IT service providers to develop and design their existing GSLA more greener for achieving sustainability as well as making more profit in their businesses. However, ICT engineer would face some challenges to incorporate, manage and finding the relationship between all new performance indicators for GSLA under three pillars of sustainability in future. This GSLA research tries to help ICT engineers to define new GSLA using an informational model. The general global view of GSLA indicators with respect to three pillars are shown in Fig.3 and then the relationships, interdependencies and management complexity among the new indicators and

existing indicators are depicted with discovering some important new entities under sustainability lens.

A. General Model of Future GSLA

To achieve sustainability, the proposed *GSLA* entity should aggregate and satisfy all three entities in general model- *Ecology Pillar*, *Economy Pillar*, *Ethics Pillar*. Now, it the matter of urgency that, to achieve sustainability the ICT industry need to identify more new parameters from user's perspective under this three pillar too. It is important to indicate that, the *ICT Product Life Cycle* must need to include at the first level of GSLA model as this entity have direct relationship to calculate existing ecological, economical and ethical indicators, such as carbon emission, energy consumption, recycling, energy cost etc. The ICT product life cycle and its relationships with sustainability pillars coexist while developing future GSLA. Therefore, ***ICT Product Life Cycle*** also needs to define as new entities for achieving sustainability in the industry. The whole life cycle of an ICT product consists of following entities, - manufacturing, transportation, usage and dismantling entities (Fig.2). All these entities should directly connect to *GSLA* entity to respect global analysis of proposed model. The total GHG emission, total energy consumption and total costing of energy could not be estimated without considering all these product life cycle entities. Additionally, an environmental closed-loop supply (ECLS) [28] chain would need to be added with the proposed relationships as currently ICT products remanufacturing are getting importance in the industry. The ECLS chain would be helpful to improve economic and environmental performance of every product [28]. The interaction between ICT product life cycle and GSLA are shown first (Fig.2) and then the general global model of future GSLA is proposed (Fig.3) using UML class diagram notation.

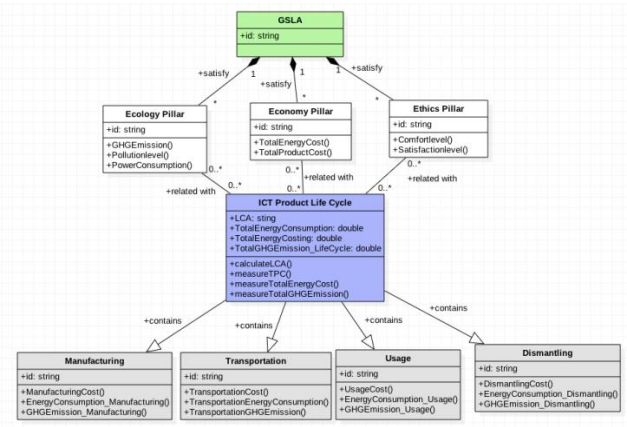


Fig. 2. Relationship between GSLA and ICT Product Life Cycle [4]

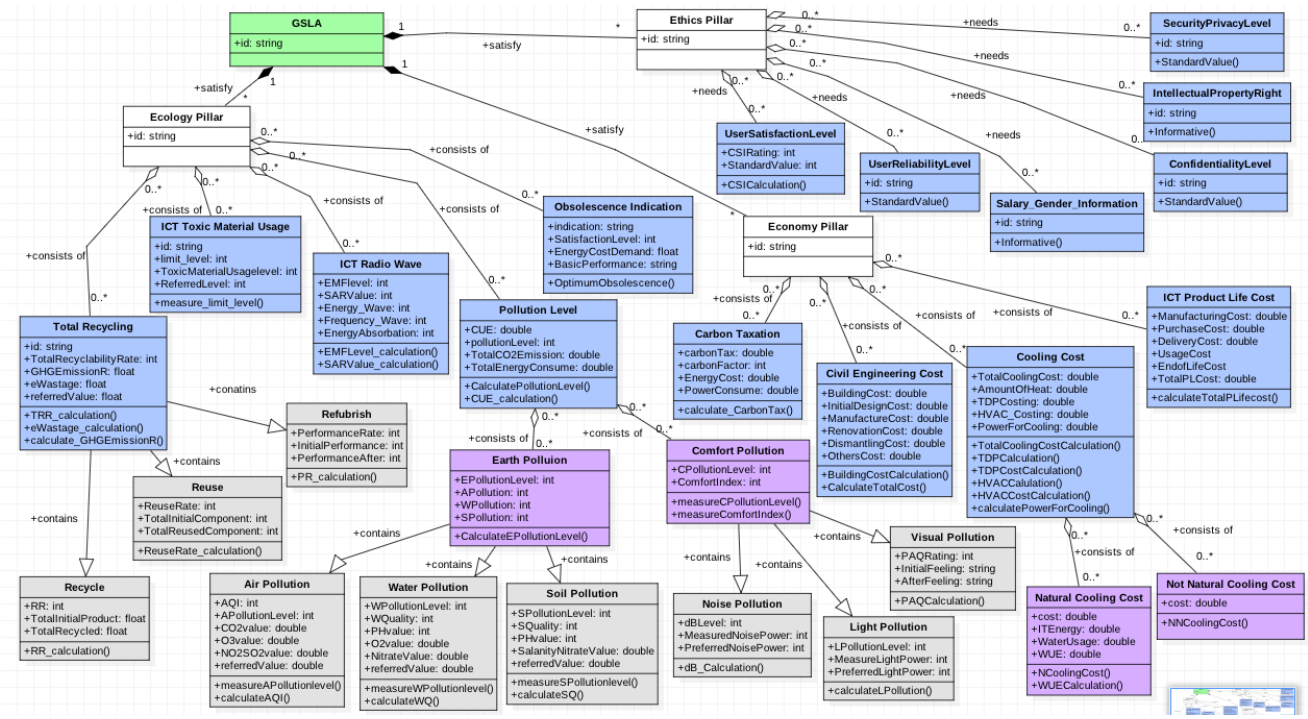


Fig. 3. General Model to define proposed GSLA

B. Discussion about the genral model of sustaianble GSLA

In the proposed model (Fig.3), the *GSLA* entity should aggregate and satisfy all three main entities of sustainability achievement- *Ecology Pillar*, *Economy Pillar*, and *Ethics Pillar*.

Ecology Pillar entity is consisting of following missing indicators from existing GSLA work, Total Recycling, ICT Toxic Material Usage, ICT Radio Wave, Pollution level and Obsolescence Indication. Total Recycling is a complex indicator and it could compose of three other sub metrics- reuse, refurbish or recycling of an ICT product or equipment. However, for simplicity, it depicts as one entity in the proposed model. Moreover, Total Recycling entity might have direct relation with some already defined existing green indicators from, such as e-Wastage and Recycling (table I) under green computing domain. For example, recycling helps

to reduce global magnitude of e-waste as metals, plastics, glass and other materials could be recovered from ICT product through recycling procedure. In addition, eWastage entity has direct impact on existing carbon emission and energy consumption of green computing domain. Again, in the general model, Pollution level consists of two other sub-entities, - Earth Pollution, which have direct impact on environment and composed of three other entities (air, water and soil) and Comfort Pollution have direct relations with people's comfort (noise, light and visual pollution entity). Therefore, **Total Recycling** could be defined as new entities in the proposed GSLA. The main challenges to define this new Total Recycling entities is to gather all necessary information from other entities and monitoring their effect. Most of non-technical parameters under sustainability pillars in this entity need some laws and directives to derive exact information for the users. Again, **Obsolescence Indication** could be another entity in the ecological pillar of proposed GSLA. Obsolescence is relative information estimated from other useful existing criteria. It could be calculated from cost of energy, carbon/GHG emission, ICT product life cycle assessment and or pollution level. There is an interesting relation between obsolescence and people. Therefore, Obsolescence Indication entity has indirect relationship with Ethics Pillar entity in proposed GSLA model. There is also an interesting relationship between existing User Satisfaction indicator with this entity. For example, people often change their mobile phone frequently because it might become old fashioned to use it. However, there is still no available standard to define obsolescence indication. Obsolescence management of an ICT product could be defined according to some regulatory lever, education/training for user behaviors and recycling practice in the society. Next, the **Pollution Level** entity might take into account for future GSLA. There is an interesting relationship between Comfort Pollution sub-entities with Ethics pillar in proposed GSLA as ethical pollution is mostly concerned with people's comfort in their daily life. The Earth Pollution level entity consists of three other entities in general model, - Air, Water and Soil pollution and Air Pollution is directly responsible for GHG emission in the atmosphere. Air, water and soil pollution might have direct relations with Recycling, ICT Toxic Material Usages and ICT Radio Wave entity.

Economy Pillar entity for sustainable GSLA is composed of *Carbon Taxation, Civil Engineering Cost, Cooling Cost* and *ICT Product Life Cost* entities. Here, the cooling cost entity is an important indicator for data center and these costing could be estimated based on either natural cooling facility (water, air) or not natural cooling facility. Moreover, *Cooling Cost* entity need to evaluate and defined accurately with the help of existing indicators from Table I, such as *TDP, HVAC Effectiveness, Cooling system efficiency* etc. In the proposed model, *Cooling Cost* entity is actually consisting of two other sub-entities, - *Natural Cooling Cost* and *Not Natural Cooling Cost.* In addition, *Natural Cooling Cost* helps to derive existing *WUE* (table I) in the model as *WUE* indicator is the ratio between annual usage of water for cooling and the total energy used by IT equipment. In addition, ***Energy Cost*** might be defined as new entity under economic pillar for achieving sustainability. *Energy Cost* entity has direct relations with *ICT Product Life Cycle, Carbon Taxation etc.* The costing of energy depends on the types of energy sources used in the IT facility. Recently, there are two types of energy is considered, - renewable (solar, wind, tidal etc) and non-renewable (gas, natural gas, fuel etc) energy. However, different types of energy costing actually depend on different country's government policies, economic conditions, political culture, and industry growth etc.

Ethics Pillar entity could be the important parameter for sustainable GSLA development model as it has direct relationships and interactions with people and society. IT ethics needs following parameters to be associated in proposed model, *User Satisfaction Level, Reliability Level, Confidentiality Level, Intellectual Property Right, Security Privacy Level, Salary and Gender Balance Information.* Here, most of the parameters under ethics pillar are very much subjective and non-technical. Thus, it could be the most challenging part for ICT engineers to monitor, manage and assess these parameters in future. *User Satisfaction Level* could be measured and evaluated by using standard method of survey for specific services and then possible to calculate standard CSI [25,26] for that services. *Security Privacy Level, Intellectual Property Right* could be monitored by third party using government defined rules and regulation. Third party could monitor and update information periodically regarding *User Reliability Level, Confidentiality Level, Salary Gender Information* in future GSLA. In the model, salary and gender balance information are shown in one entity for simplicity and they might carry same type of informative attributes. Moreover, still there is no standard authority or third party to evaluate these ethical parameters. The ICT companies should also analyze their social responsibilities towards their customers, employee and community through developing IT Ethics program and guideline.

C. Identification of new entities for future GSLA

Fig.3 shows the general view of proposed new GSLA definition and now the complexity of managing future GSLA discussed briefly in previous sub-section. In future, this research could elaborate the sustainability achievement in the IT and ICT industry by taking all important entities from the discussion of general model. These newly identified entities might have direct and indirect relationships and different level of effects with some existing and new green performance indicators. Therefore, this research identifies following central entities for sustainable future GSLA, - *ICT Product Life Cycle, Total Recycling, Obsolescence Indication, Pollution level, GHG Emission, Energy Consumption,* and *Energy Cost.* Among these, *GHG Emission, Energy Consumption* and *Energy Cost* are already defined entities in exiting GSLA work under green computing practice in the industry. All these newly identified entities would cover all the dependencies and respect all other existing and new indicators under three pillars of sustainability (Fig.1).

TABLE V. RELATIONSHIPS BETWEEN NEW AND EXISTING PARAMETERS FOR SUSTAINABLE GSLA

Central Entity	Relationships		
	Direct	**Indirect Important Effects**	**Indirect Small Effects**
Total Recycling	ICT Product Life; eWastage; Earth Pollution; Energy Consumption; GHG Emission; Energy Cost; Dismantling;	ICT Radio Wave; ICT Toxic Material Usage; Manufacturing	Comfort Pollution
Obsolescence Indication	ICT Product Life; ICT Performance; ICT Product Cost;	Pollution Level; Energy Consumption; GHG Emission	Ethics Pillar
GHG Emission	Total Recycling; Air Pollution; Carbon Taxation; Dismantling; Energy Consumption;	Obsolescence Indication; ICT Toxic Material Usage	Comfort Pollution
Energy Consumption	Total Recycling; ICT Product Life; ICT Product Cost; Energy Cost; Carbon Taxation; GHG Emission; ICT Radio Wave.	Obsolescence Indication; Cooling Cost	Building Cost
Pollution Level	ICT Product Life; Total Recycling; Energy Consumption; GHG Emission.	ICT Toxic Material; ICT Radio Wave; Obsolescence Indication	Ethics Pillar
ICT Product Life	Energy Consumption; GHG Emission; Pollution Level; Total Recycling; Energy Cost;	Obsolescence Indication;	ICT Product Cost;
Energy Cost	ICT Product Life; ICT Product Cost; Carbon Taxation; Energy Consumption.	Cooling Cost; Building Cost	Total Recycling.

V. CONCLUSION

This sustainable GSLA research discovers most of the recent day green indicators and their measurable unit (Table I) from various cloud service providers and as well as from some data centers. In addition, it discovers today's concerns are mainly on energy issues and productivity under greening lens. Missing performance indicators and their influences on GSLA with respect to 3Es are discussed and also identified in this research. Table II to Table IV lists all new proposed performance indicators and their measurable units for developing a new sustainable GSLA. Thus incorporating all new and existing indicators for future GSLA might be difficult and cumbersome work for the ICT engineers. The management complexity of all identified indicators in future sustainable GSLA would be the most challenging task. Therefore, the definition of GSLA section thus proposes an informational model to help ICT engineers to understand the interactions and important effects of various performance indicators. The informational model also helps to design a new sustainable GSLA and to derive new parameters under sustainability lens. Therefore, it could use a guideline for the ICT engineers in future. Still some challenges exist for designing sustainable GSLA research such as, some performance indicators need to be defined accurately which has association with other indicators; most of the subjective, qualitative indicators related with ethics issue need standardization or governed and authorized by proper laws and directives. The standardization of green indicators is one of the main issues as mentioned by ITU-T report (2012). Also, further research is necessary on monitoring and evaluating the indicators for a viable sustainable GSLA in the industry. The next steps of this research is to design all newly identified entities from the global model and finds out the evaluation procedure of this new sustainable GSLA.

ACKNOWLEDGMENT

The authors would like to show their gratitude and thanks to PERCCOM program (European Union) for giving the idea of SLA research. Part of this work published in IJACSA, Vol.6. No.12, 2015 and this is an extension of previous work.

REFERENCES

[1] R. Buyya, J. Broberg, and A. Goscinsk, "Cloud Computing: Principles and Paradigm," A John Wiley & Sons, Inc. Publication, ISBN: 978-0-470-88799-8, February 2011.

[2] SLA@SOI, Source: http://sla-at-soi.eu/, retrieved on April 2015.

[3] L. Wu, and R. Buyya, "Service Level Agreement (SLA) in Utility Computing Systems," Performance and Dependability in Service Computing: Concepts, Techniques and Research Directions, V. Cardellini et. al. (eds), ISBN: 978-1-60-960794-4, IGI Global, Hershey, PA, USA, July 2011, pp.1-25.

[4] I. Ahmed, H. Okumura, and K. Arai, "Analysis on existing Basic SLAs and green SLAs to define new sustainable Green SLA", International Journal of Advanced Computer Science and Applications, Vol.6, No. 12, December 2015, pp. 100-108.

[5] J. Mankoff, R. Kravets, and E. Blevis, "Some Computer Science Issues in Creating a Sustainable World," Computer, Vol. 41, No. 8, 2008.

[6] SMART 2020 Report, "Enabling the low carbon economy in the information age," The Climate Group, GeSI, 2008.

[7] Z. S. Andreopoulou, "Green Informatics: ICT for Green and Sustainability," Journal of Agriculture Informatics (EIFTA), Vol. 3, No. 2, 2012.

[8] E. Rondeau, F. Lepage, J. P. Georges, and G. Morel, "Measurements and Sustainability," Chapter 3, Green Information Technology, 1st Edition, A Sustainable Approach, Dastbaz & Pattinson & Akhgar, ISBN: 9780128013793, Elsevier Book, 304 pages, March 2015.

[9] D. Jiang, X. Zhengzheng, and L.V Zhihan, "A multicast delivery approach with minimum energy consumption for wireless multihop networks" Journal of telecommunication Systems, 2015, pp. 1-12.

[10] D. Jiang, X. Ying, Y. Han Y, and L.V Zhihan, "Collaborative multi-hop routing in cognitive wireless networks" Journal of Wireless Personal Communications, 2015, pp. 1-23.

[11] N. Drouant, E. Rondeau, J. P. Georges, and F. Lepage, "Designing green network architectures using the Ten Commandments for a mature ecosystem," Computer Communications, Vol. 42, April 2014, pp. 38-46.

[12] White Paper, "The Dangerous Chemical in Electronic Products," Toxic Tech, Source: http://www.greenpeace.org/usa/PageFiles/58525/toxic-tech-chemicals-in-elec.pdf, Greenpeace, Retrieved on March 2015.

[13] A. Chen, K. N. Dietrich, X. Huo, and S.-M. Ho, "Development Neurotoxicants in E-Waste: An Emerging Health Concern," Environmental Health Perspectives, Vol.119, No.4, April 2011.

[14] M. Dastbaz, C. Pattinson, and B. Akhgar, "Green Information Technology: A Sustainable Approach," ISBN: 9780128013793, Elsevier Book, 348 pages, March 2015.

[15] C. Tuppen, "Circularity and the ICT Sector," Advancing Sustainability LLP @ Ellen MacArthur Foundation, United Kingdom, September 2013.

[16] P. Sandborn, "Software Obsolescence- Complicating the Part and Technology Obsolescence Management Problem," IEEE Transaction on Components and Packaging Technologies, Vol.30, No.4, December 2007, pp. 886-888.

[17] White paper, "EMF and RF safety Levels- A comparative Guide," ScanTech Report, Source: www.scantech7.com, 214.912.4691, Australia, March 2015.

[18] M. Kassner, "Data Center may be hazardous to your hearing," Online Article of Tech Republic U.S, February 2014.

[19] P. Zhang, "Theorizing the Relationship between Affect and Aesthetics in ICT Design and Use Context," International Conference on Information Resource Management, Dubai, UAE, May 2009.

[20] T. R. Akinbinu, and Y. J. Mashalla, "Impact of Computer Technology on Health: Computer Vision Syndrome," Journal of Medical Practice and Review, Vol. 5 (3), November 2014, pp. 20-30.

[21] L. H. Goulder, and A. R. Schien, "Carbon Taxes versus Cap and Trade: a critical review," Journal of Climate Change Economics, Vol. 4, No. 3 (2013), November 2013.

[22] R. Repetto, "White paper: Cap and Trade is Better Climate Policy than a Carbon Tax," United Nations Foundation, May 2013.

[23] A. S. Williams, "Life Cycle Analysis: A Step by Step Approach," ISTC Reports, Illinois Sustainable Technology Center, Urbana Champaign, USA, December 2009.

[24] M. B. Uddin, M. R. Hassan, and Kazi M. Tarique, "Three Dimensional Aspects of Corporate Social Responsibility," Daffodil Intenational University Journal of Business and Economics, Bangladesh, Vol.3, No.1, January 2008.

[25] Presentation on Customer Satisfaction Index (CSI), Institute for Choice, University of South Australia, January 2014.

[26] B. Angelova, and J.Zekiri, "Measuring Customer Satisfaction with Service Quality Using American Customer Satisfaction Model (ACSI Model)," International Journal of Academic Research in Business and Social Sciences, Vol.1, No.3, October 2011.

[27] E. Ciavolino, and J. J. Dahlgaard, "ECSI- Customer Satisfaction Modelling and Analysis: A Case Study," Journal of Total Quality & Business Excellence, Vol. 18, Issue. 5, July 2007, pp. 545-554.

[28] M. A. Ruimin, Y. A. O Lifei, J. I. N Maozhu, R. E. N Peiyu, and L. V Zhihan, "Robust environmental closed-loop supply chain design under uncertainty" Journal of Chaos, Solitons & Fractals, Elesevier, November 2015.

AUTHORS PROFILE

Kohei Arai, He received BS, MS and PhD degrees in 1972, 1974 and 1982, respectively. He was with The Institute for Industrial Science and Technology of the University of Tokyo from April 1974 to December 1978 also was with National Space Development Agency of Japan from January, 1979 to March, 1990. During from 1985 to 1987, he was with Canada Centre for Remote Sensing as a Post Doctoral Fellow of National Science and Engineering Research Council of Canada. He moved to Saga University as a Professor in Department of Information Science on April 1990. He was a councilor for the Aeronautics and Space related to the Technology Committee of the Ministry of Science and Technology during from 1998 to 2000. He was a councilor of Saga University for 2002 and 2003. He also was an executive councilor for the Remote Sensing Society of Japan for 2003 to 2005. He is an Adjunct Professor of University of Arizona, USA since 1998. He also is Vice Chairman of the Commission "A" of ICSU/COSPAR since 2008. He wrote 33 books and published 510 journal papers. He is now Editor-in-Chief of IJACSA and IJISA.

A Novel Approach to Detect Duplicate Code Blocks to Reduce Maintenance Effort

Sonam Gupta
Research Scholar, Suresh Gyan Vihar University,
JAIPUR (RAJASTHAN), INDIA

Dr. P. C Gupta
Associate Professor, Department of Computer Science & Informatics, University of Kota
KOTA (RAJASTHAN), INDIA

***Abstract*—It was found in many cases that a code might be a clone for one programmer but not the same for another one. This problem occurs because of inaccurate documentation. According to research, the maintainers are not aware of the original design and thus, face the difficulty of agreeing on the system's components and their relations or understanding the work of the application. The problem also occurs because of the different team of development and maintenance resulting in more effort and time during maintenance. This paper proposes a novel approach to detect the clones at the programmer side such that if a particular code is a clone then it can be well documented. This approach will provide both the individual duplicate statements as well as the block in which they appear. The approach has been examined on seven open source systems.**

Keywords*—*Clones; Program Dependence Graph (PDG); Control Flow Graph (CFG); Abstract Syntax Tree (AST)

I. INTRODUCTION

Detecting duplicate code occurrence requires a powerful understanding of the clones. Studies[1,2] connected with clone investigation conveys with its assessments of partner degree question inside one of the supply code, as in an exceedingly net application [3], inside the identification of clone-related bugs [4], partner degreed in a product bundle piece [5]. Distinctive studies measure different previews upheld the modification history of the code to see the genealogic nature of clones [6], to watch the consistency of clones being looked after [7], and to spot real refactoring of clones [8].

Alongside this, there are difficulties connected with the understanding of clones at the clone group level and potential monstrous measures of data that may be recovered from clone identification instruments. Also, difficulties were found inside the upkeep of clones in light of the fact that it identifies with the evacuation of their related duplication through the strategy for refactoring.Redundant code is additionally typically deceptively referred to as cloned code within the literature—although that means that one piece of code springs from the other one within the original sense of this word. In step with the Merriam-Webster lexicon, a clone is one that seems to be a replica of a resourceful kind. It's an equivalent word to duplicate. Despite the fact that exploration winds up in repetitive code, not every excess code could be a clone. There are additional cases inside which two code fragments that aren't any duplicate of each option just happen to be comparative or perhaps indistinguishable all of a sudden. Likewise, there is additionally repetitive code that is semantically proportional. However, it consolidates an absolutely different usage. There is no understanding of the investigation group on the exact thought of repetition and cloned code. The meaning of clones communicates this dubiousness as clones square measure portions of code that square measure comparative in venture with some meaning of likeness. There are various exact studies on the advancement of clones that depict some consideration getting perceptions. Antoniol, *et al*. propose measurement got from clones over numerous arrivals of a framework to watch and foresee the development of clones [9]. Their study for the data base framework mSQL demonstrated that the forecast of the normal assortment of clones every work is genuinely solid. In an alternate detailed analysis for the UNIX working framework piece, they found that the extent of cloned codes is confined. Exclusively few clones are regularly found crosswise over frameworks; most clones are completely contained inside a subsystem. Inside the framework building design, constituting the equipment plan reflection layer, more up-to-date equipment architectures have a tendency to display marginally higher clone rates. The explanation behind this improvement is that more current modules are normally gotten from existing comparative ones.

Reusing code pieces by reiteration and sticking with or while not minor adjustment may be a typical movement in programming bundle advancement. Therefore, programming bundle frameworks ordinarily contain areas of code that are awfully comparable, alluded to as programming bundle clones. Prior exploration demonstrates that a real part of the code in an extremely commonplace programming framework has been cloned , and in one great case it had been even five hundredth [10]. While such duplicate is generally purposeful and may be useful from multiple points of view [11], it might be unsafe in programming bundle support and development. Case in point, if a bug is located in a code piece, all similar parts should be checked for a comparative bug [12]. Copied pieces might impressively expand the work to be carried out once upgrading or adjusting code [13]. A late study that worked inside the setting of business frameworks demonstrates that conflicting changes to codes are successive and bring about extreme astounding conduct [14]. Numerous other options moreover demonstrate that product bundle frameworks with code clones will be harder to keep up [15,16] and may present refined blunders [12,17]. In this way code clones are thought about one amongst the undesirable "odours" of a PC code [18] and it is wide accepted that cloned code will make programming bundle upkeep and advancement

impressively a great deal of troublesome. Accordingly, the location, viewing and evacuation of code clones is a crucial subject in programming bundle support and development [19]. Thus, the algorithm proposed in this paper will help the programmer to locate the exact position of the cloned code. Along with the position it will also inform about the percentage of clone within the system. If the percentage is more than that of threshold value then the programmer either documents it correctly or else removes the clone. This reduces the maintenance effort.

II. Algorithm to Detect the Duplicate Codes

The algorithm proposed is based on the hybrid technique which combines program dependence graphs (PDG), control flow graphs (CFG), and abstract syntax tree (AST) based approach. This approach is better than other approaches [21] since by combining three techniques it will work well at the complier level and AST will help to construct the nodes of duplicate codes. The algorithm will work as follows.

First, the statements will be inserted in a function named InsertStatements()

Step 1: Generate a key for entering statement into navigable collection (which is the value of the statement)

Step 2: increment the no of statements present in navigable collection

Step 3: if identity is already present, add this value into that key value pair

Step 4: else add a new item into the navigable collection and terminate

After making the pool of all statements, the matrix will be used for setting the values as true in a square matrix of size equal to no. of statements in an item of the navigable collection and the matrix is updated only in the upper half portion, that is, above the upper right principal diagonal only with no operation on lower half of matrix. The working of matrix will be done in putAMarkOnMatrix() function :

Step 1: initialize index1 with starting index of statement1

Step 2: initialize index2 with starting index of statement2

Step 3: if index1 < index2

Set all the cells true or 1 from index1 to index2

Step 4: else

Set all the cells true or 1 from index2 to index1

Now the duplicity of the statements will be checked in StatementGroup.java function. For checking the duplicity, it uses a matrix to store all the different statements in the class

Step1: if (index1 < index2) then collector.setTrueToCell(index1, index2);

Step2: else collector.setTrueToCell(index2, index1);

Whenever a duplicate statement is found, the cell in the matrix for showing the result is highlighted, and a mark is put on the code which is duplicated again using the above given code as given in following code:

```
Step 1: for (inti = 0; i< N - 1; i++) {  //N Array size
          Statement s1 = statArray.get(i);
Step2: for (int j = i + 1; j < N; j++) {
          Statement s2 = statArray.get(j);
          //Match or not
            if (s1.equals(s2)) {       //insert the result to matrix
                putAMarkOnMatrix(collector, s1, s2);
Step 3: Repeat above steps till N
```

Above code is checking the code for the duplicity. If it is duplicate code, then it puts the mark on the duplicate code using above pseudocode. Now the AST structure will be constructed so that the clones can be identified.

The working of construction of AST will be as follows:

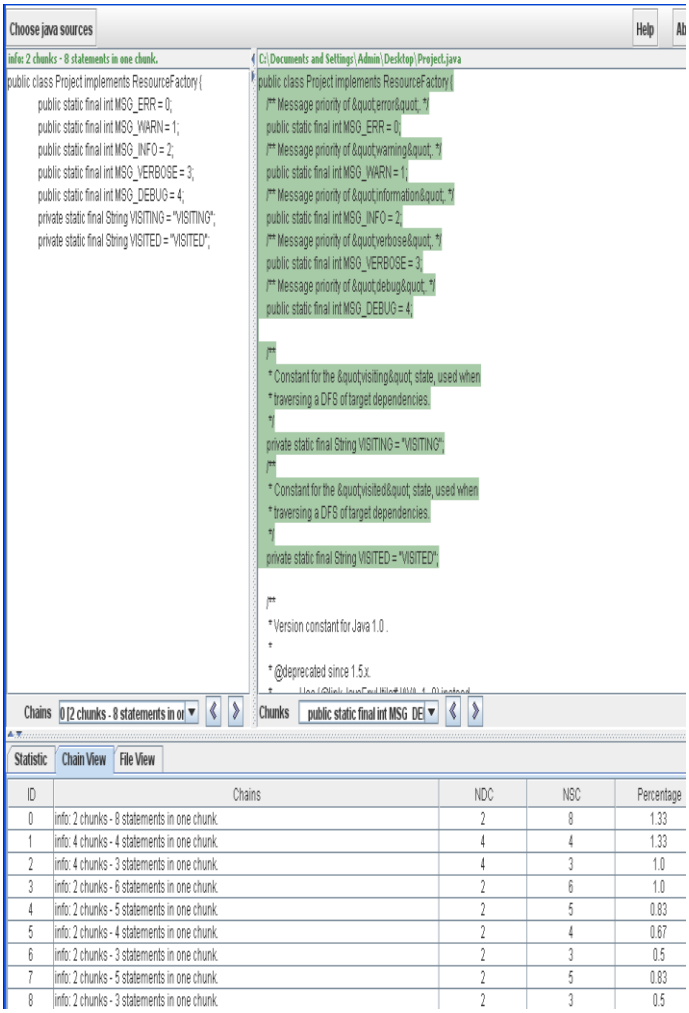
```
Step 1: Initialize token = new CodeReviewToken(i, s);
Step 2: Make type= astType; //token.getType();
Step 3: if (Configuration.anonymizeType(type)) then
 return  JavaRecognizer._tokenNames[type];
Step 4: return token.getText();
Step 5; if (token != null) then
         index = (token).getStartIndex();
Step 6:Initialize tid = this.token.getID();
 Step 7: if (tid > -1) then tl.addToken(tid)
Step 8: if (getFirstChild() != null) then
         if (!Configuration.anonymizeType(astType)) then
            ts += ((CodeReviewAST)
getFirstChild()).toStringList(tl);
Step 9: Display ts as the duplicated code block.
```

The AST constructed will also be able to identify non-contiguous clones as now if any extra line is inserted or deleted, it will form a block of code.

By using the above-mentioned pseudocode every token will be assigned a unique id so that when converting each text into tokens all of them can be grouped on the basis of their numbers and then each duplicate code can be categorized under single head.

III. Results and Experimental Study

The above proposed algorithm has been developed in Java. The tool has been analysed on seven open source systems namely Apache Ant 1.7.0, Columbia 1.4, EMF 2.4.1, JMeter 2.3.2, JEdit 4.2, JFreeChart 1.0.10, and JRuby 1.4.0. The result will be viewed as given in Figure 1.



ID	Chains	NDC	NSC	Percentage
0	info: 2 chunks - 8 statements in one chunk.	2	8	1.33
1	info: 4 chunks - 4 statements in one chunk.	4	4	1.33
2	info: 4 chunks - 3 statements in one chunk.	4	3	1.0
3	info: 2 chunks - 6 statements in one chunk.	2	6	1.0
4	info: 2 chunks - 5 statements in one chunk.	2	5	0.83
5	info: 2 chunks - 4 statements in one chunk.	2	4	0.67
6	info: 2 chunks - 3 statements in one chunk.	2	3	0.5
7	info: 2 chunks - 5 statements in one chunk.	2	5	0.83
8	info: 2 chunks - 3 statements in one chunk.	2	3	0.5

Fig. 1. Screenshot of proposed tool

As shown in Figure 1, the tool will tell the programmer about the number of duplicate blocks and number of statements in each block along with its location. Now the programmer can make the changes in the documentation or in the code as required. This approach will reduce the maintenance effort to a much lower level. The number of clones found in each system by the help of this tool were compared with the JDeodorant tool[20] as well as with manual detection[20]. The comparison between the proposed approach and JDeodorant is shown in Figure 2. The results clearly show that the proposed tool extracted more number of bad smells. Similarly. the cloned blocks detected by the proposed tool are compared with the manual approach as shown in Figure 3. It clearly shows that the proposed tool is able to find almost same number of cloned blocks as that were actually present in the systems.

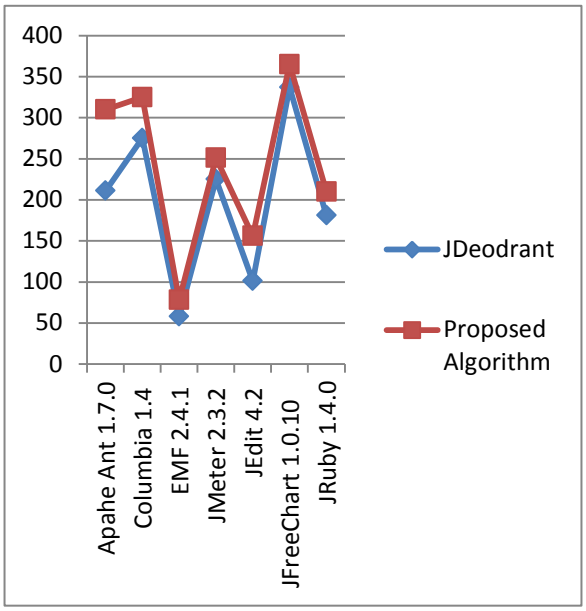


Fig. 2. Comparison of detected clones

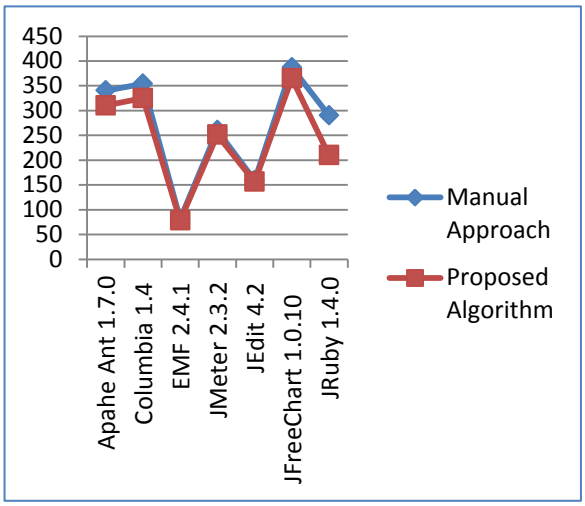


Fig. 3. Comparison of detected clones with manual approach

IV. Conclusion

The proposed tool has been developed to reduce the maintenance effort, as it has been proved [5] that a lot of cost and effort is wasted in maintenance due to clones. This tool consists of hybrid techniques of PDG, CGF and AST thereby overcoming the disadvantages of each. The tool has been experimented of seven open source systems. Along with it the results have also been compared with JDeodorant tool as well as with manual extraction. The results clearly show that the tool is able to find more number of clones which actually have

bad smell. Moreover. the tool also provides the location of each duplicate block along with its percentage. Now the programmer can set the threshold value based on the percentage. If he finds that the percentage is higher, then he will either remove the duplicity or will document it correctly so as to reduce maintenance time and cost.

V. Future Work

The tool has been experimented on just open source systems. In future. the work will be extended to study on licensed systems. Along with this, the work will be extended to convert the duplicated block into functions so that a single change can be reflected in all versions.

References

[1] Basili, V. R. and B. T. Perricone (1984). "Software errors and complexity: an empirical investigation." Commun. ACM 27(1): 42-52.

[2] Parnas, D. L. (1994). Software aging. Proc. Int'l Conf. on Software Engineering (ICSE). Sorrento, Italy, IEEE Computer Society Press**:** 279-287.

[3] Damith Rajapakse and Stan Jarzabek, "Using Server Pages to Unify Clones in Web Applications: A Trade-Off Analysis," *International Conference on Software Engineering*, Minneapolis, Minnesota, May 2007, pages 116 - 126.

[4] Lingxiao Jiang, Zhendong Su, and Edwin Chiu, "Context-Based Detection of Clone-Related Bugs," *Joint Meeting of the European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, Dubrovnik, Croatia, September 2007, pages 55 - 64.

[5] Robert Tairas and Jeff Gray, "An Information Retrieval Process to Aid in the Analysis of Code Clones," *Empirical Software Engineering*, Volume 14, Number 1, February 2009, pages 33 - 56.

[6] Miryung Kim, Vibha Sazawal, David Not kin, and Gail Murphy, "An Empirical Study of Code Clone Genealogies," *Joint Meeting of the European Software Engineering Conference and Foundations of Software Engineering*, Lisbon, Portugal, September 2005, pages 187 - 196.

[7] Lerina Aversano, Luigi Cerulo, and Massimiliano Di Penta, "How Clones are Maintained: An Empirical Study," *European Conference on Software Maintenance and Reengineering*, Amsterdam, The Netherlands, March 2007, pages 81 - 90.

[8] Robert Tairas and Jeff Gray, "Sub-clones: Considering the Part Rather than the Whole," *International Conference on Software Engineering, Research, and Practice*, Las Vegas, Nevada, July 2010.

[9] Antoniol, G., Casazza, G., Penta, M.D., Merlo, E.: Modeling clones evolution through time series. In: International Conference on Software Maintenance, IEEE Computer Society Press (2001) 273–280.

[10] St´ephane Ducasse, Matthias Rieger, Serge Demeyer. A Language Independent Approach for Detecting Duplicated Code. In Proceedings of the 15th International Conference on Software Maintenance (ICSM'99), pp. 109-118, Oxford, England, September 1999.

[11] Cory Kaiser and Michael W. Godfrey. "Cloning Considered Harmful" Considered Harmful: Patterns of Cloning in Software. Empirical Software Engineering, Vol. 13(6):645–692 (2008).

[12] Zhenmin Li, Shan Lu, Suvda Myagmar, and Yuanyuan Zhou. CP-Miner: Finding Copy-Paste and Related Bugs in Large-Scale Software Code. In IEEE Transactions on Software Engineering, Vol. 32(3): 176-192, March 2006.

[13] Jean Mayrand, Claude Leblanc, Ettore Merlo. Experiment on the Automatic Detection of Function Clones in a Software System Using Metrics. In Proceedings of the 12th International Conference on Software Maintenance (ICSM'96), pp. 244-253, Monterey, CA, USA, November 1996.

[14] E. Juergens, F. Deissenboeck, B. Hummel and S. Wagner. Do Code Clones Matter? In Proceedings of the 31st International Conference on Software Engineering (ICSE'09), pp. 485–495, Vancouver, Canada, May 2009.

[15] Xian, Y., Angler, D.: Using redundancies to find errors. In: Proceedings of the 10th ACM SIGSOFT symposium on Foundations of software engineering, ACM Press (2002) 51–60.

[16] S. Schleimer, D. S. Wilkerson, and A. Aiken. Winnowing: local algorithms for document fingerprinting. In Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data (SIGMOD'03), pp. 7685, San Diego, California, June 2003.

[17] A. Chou, J. Yang, B. Chelf, S. Hallem and D. R. Angler. An Empirical Study of Operating System Errors. In *Proceedings of the 18th ACM symposium on Operating systems principles (SOSP'01)*, pp. 73–88, Banff, Alberta, Canada, October 2001.

[18] Martin Fowler. *Refactoring: Improving the Design of Existing Code*, Addison-Wesley, 1999.

[19] B. Lagu¨e, D. Proulx, J. Mayrand, E. Merlo and J. Hudepohl. Assessing the Benefits of Incorporating Function Clone Detection in a Development Process. In Proceedings of the 13th International Conference on Software Maintenance (ICSM'97), pp. 314– 321, Bari, Italy, October 1997.

[20] Krishnan, Giri Panamoottil, and Nikolaos Tsantalis. "Unification and refactoring of clones." Software Maintenance, Reengineering and Reverse Engineering (CSMR-WCRE), 2014 Software Evolution Week-IEEE Conference on. IEEE, 2014.

[21] C.K. Roy and J.R Cordy, A Survey on Software Clone Detection Research, Queen's School of Computing Technical Report: 541 pp., September 2007.

An Adaptive Key Exchange Procedure for VANET

Hamza Toulni
GITIL Laboratory, LIAD. Faculty of Sciences Aïn Chock
Hassan II University of Casablanca
Casablanca, Morocco

Benayad Nsiri
GITIL Laboratory, LIAD. Faculty of Sciences Aïn Chock
Hassan II University of Casablanca
Casablanca, Morocco

Mohcine Boudhane
GITIL Laboratory, LIAD. Faculty of Sciences Aïn Chock
Hassan II University of Casablanca
Casablanca, Morocco

Mounia Miyara
LIAD Laboratory. Faculty of Sciences Aïn Chock
Hassan II University of Casablanca
Casablanca, Morocco

***Abstract*—VANET is a promising technology for intelligent transport systems (ITS). It offers new opportunities aiming at improving the circulation of vehicles on the roads and improving road safety. However, vehicles are interconnected by wireless links and without using any infrastructure, which exposes the vehicular network to many attacks. This paper presents a new solution for the exchange of security keys to protect information exchanged between vehicles. In addition to securing the inter-vehicular communication, the proposed solution has considerably decreased the time for the exchange of keys, thus improving the performance of VANET.**

Keywords—ITS; Vehicular ad-hoc networks; public key exchange; Security

I. INTRODUCTION

Vehicular Ad Hoc NETworks (VANETs) are a wireless communication technology applied to transportation; it is specially designed to solve the problems caused by the increasing number of vehicles and urban sprawl. VANET permits communication between vehicles themselves or between vehicles and the road infrastructure to improve the intelligent transport systems by the major benefits that can be gained from wireless technology, such as the improvement of road safety and traffic fluidity.

VANET is a subclass of MANET (Mobile Ad Hoc Networks), in which the mobile nodes are replaced by vehicles. So vehicles inheriting all properties associated with the nodes in MANET, but with some special characteristics, such as high speed of nodes which makes the environment of VANET highly dynamic, and this leads to frequent network topology changes. And unlike traditional wireless networks where limited power is a major constraint, nodes of vehicular networks have large capacities of energy they derive from vehicle power system, which ensures better performance in the computations.

However, in addition to the problems inherited from MANET, there are other challenges [1] that must be overcome to enable communication between vehicles by VANET. One of the most critical and important problems is security and privacy, due to the importance of information exchanged within VANET, and each change in the alerts constitutes a serious threat to people's lives.

VANET is an ideal target for various attacks [2] because vehicles share among themselves all kinds of information via wireless links without any administration by a centralized infrastructure, this facilitates attackers to intercept the information exchanged or to inject wrong information in the network. Hence the importance of securing VANET to protect the exchanged information, but adding the security mechanism involves an additional computation cost in the network, thereby influencing the transmission performance.

To address this critical issue of security, a new procedure for the exchange of security keys is present in this paper to protect information exchanged between vehicles, in addition to ensuring the inter-vehicle communication; the proposed solution has significantly reduced the time for key exchange, improving the performance of VANET.

The rest of the paper is organized as follows; the following section provides an overview of security in VANET. In Section III, we present an overview of cryptography. In Section IV, we present the proposed key exchange procedure. In Section V, we present the simulation results and an analysis of the proposed procedure. Finally, we conclude the paper in Section VI.

II. RELATED WORK

VANET is a promising technology that provides several advantages to supply value added services to improve safety and traffic, but the nature of the transmission medium makes VANET more vulnerable to attack. Therefore, network security is an essential element to support the implementation and operation of applications and services in VANET.

A. The security threats

As each network VANET is exposed to several attacks:

- Sniffing: The malicious vehicles listening to the transmission medium in order to extract information exchanged in its neighborhood; it may want to spy on personal information or collect information and to perform then other types of attacks.
- Unauthorized access: The malicious vehicles are accessing to network services without having the rights or privileges.

- Denial of Service: The goal is to make the different resources and services unavailable to users in the network; it is usually caused by other attacks on the bandwidth or energy resources of other nodes. The most naive technique to cause a denial of service in a wireless network is Jamming, another method of attack which consists of requesting a service that provides by a node in a repetitive manner in order to waste his resources.
- Spoofing: The malicious vehicles attempting to impersonate another node in order to receive their messages or have the privileges that are not granted.
- Falsifying information: Malicious vehicles are attempting to change the information contained in a message or even remove messages during their trip.
- Therefore, the security mechanisms in VANET must necessarily reach a number of general security requirements, such as:
- Authentication: This security required allows network members to ensure the correct identification of vehicles with which they communicate, and thus know more information about the issuer vehicle as its identifier, address, properties, and its geographical position.
- Integrity: This security required helps to ensure that the data exchanged are not subjected to voluntary or accidental tampering. Thus, it allows recipients to detect data manipulation by unauthorized entities and to reject the packages.
- Confidentiality: This security required guarantees that only authorized entities can access to data transmitted across the network. However, the confidentiality of information in VANET depends on the application and the communication scenario, especially in the case of warning messages of an emergency that must be read by any entity in VANET.
- Non-repudiation: This security required ensures that no required sender cannot deny being at the origin of a message, this objective is essential in sensitive communications. So the overall purpose of non-repudiation is to collect, maintain and make available all the evidence about an event or action, to resolve disputes about an occurrence and not an action. Non-repudiation depends on authentication, and the system can identify the author of a malicious message.
- Availability: This security is required to guarantee entities authorized to access network resources with adequate quality of service. The resources must remain available even in the case of failure in the network. This not only secures the system but also makes it fault tolerant. And resources should remain available until the fault is repaired.

To satisfy these requirements and overcome the threats of attacks, many researchers have proposed solutions to ensure secure communication within VANET.

B. Proposed solutions

In the literature, the security issue VANET attracted the attention of many researchers, and several solutions have been proposed to overcome the threats of attacks.

In [3] Raya, *et al.* propose a detailed analysis of threats that endanger VANET and propose a security architecture. This architecture is based on the use of private keys and also included a certification authority, they also proposed a method for the management and conservations of the keys.

In [4], Karl, *et al.* have proposed the Security-Requirements Engineering using Cluster Analysis (SECA). This is an approach which allows the analysis of a large number of applications by selecting a typical representation covering the required application cluster, then, they determine the security mechanisms for all subsets of trained applications.

In [5], Plossl, *et al.* have proposed a security architecture for VANET (SAV). The communication model of this architecture is based on the fact that there are two types of communication: communication messages passive such as beacons messages that are sent periodically and active communication messages that are sent when an event occurs and a warning is to be sent to neighboring vehicles. The security architecture they propose for VANET is divided into three layers: The lower layer that includes basic security features, The security layer to jump, and The multi-hop layer, it includes all the applications and services used in VANET.

In [6] Dhurandhar, *et al.* have presented Vehicular Security through Reputation and Plausibility checks (VSRP) approach to deploy security in VANET, their algorithms take into account three types of events: traffic jams, accidents, and braking applications. The algorithm uses a system based on the reputation of the sensors, not only to detect but also to isolate malicious nodes present in the network. This algorithm also allows managing the problems related to aggregation and deletion of data. This algorithm operates on an event-oriented approach. Three types of events are listed: a-jumping, multi-jump, and malicious intent. The protocol distinguishes three types of packages for messages: data packets, requests packets of neighbors (neighborreq packet), and response packets neighbors (neighborrep packet).

In [7] Golle, *et al.* proposed a general approach to assessing the validity of data in the VANET. In their approach, the node tries different possible explanations about the data it has collected; based on the assumption that a malicious node is afraid to attend. Their techniques to assess and classify the nodes depends on two assumptions: the nodes have an ability to exchange information with each other, plus a parsimony argument accurately reflects contradictory behavior in the VANET. This technique allows them to detect incorrect information about the identity of the node or nodes of the emitters of this incorrect information with high probability.

In [8], Tiffany Hyun-Jin, *et al.* proposed a model to distinguish spurious messages from legitimate messages. They explore six different sources of information to enable vehicles to filter malicious messages that are transmitted by a minority of disobedient vehicles. The six sources are as follows.

- The digital signature verification result.
- The geographical location of the source.
- Local sensors to the vehicle.
- The messages of other vehicles: Is there a contradiction between alerts?
- The validation infrastructure (RSU).
- The reputation of the issuer.

This model validation warning is based on two components: the level and the Certainty of Event (CoE). An alert is triggered when the certainty of the event exceeds a threshold.

III. Cryptography

Security is an unsurpassable prerequisite for the deployment of VANET. In fact, wireless networks are generally vulnerable to espionage and attacks, and the importance of information sent between vehicles, increase the probability of occurrence of these threats.

Cryptography is the technique used to make the confidential data by encrypting at the source node and deciphering at the destination node. It can be considered as a key solution to most of these threats.

We distinguish two types of encryption and decryption algorithms [9].

- Symmetric-key algorithms in which all nodes have the same encryption key.
- Asymmetric-key algorithms where we distinguish the use of two keys, one public known by all nodes and the other is private for each node.

To ensure that the information is only accessible by authorized entities, the most reliable solution is to use asymmetric algorithms. This infrastructure is known As Public Key Infrastructure (PKI).

In a PKI, the communication is encrypted with a digital certificate and obtain this certificate, the entity made a request to the Registration Authority. This generates a couple of keys (public key, private key) and sends the private key to the entity. Consequently, PKI communication takes place in several phases as shown in Figure 1.

- *Phase 1*: the entity B requests access to entity A.
- *Phase 2:* the entity A sends its certificate, which contains its public key.
- *Phase 3:* the entity B verifies the authenticity of the certificate of entity A. Specifically, it checks the signature of entity A. At this moment, the entity B is sure of the authenticity of the certificate of the entity A
- *Phase 4:* same as phase 2, the entity B sends its certificate.
- *Phase 5:* same as phase 3, the entity A verifies the certificate of entity B. At this time, the A entity is sure of the authenticity of the certificate of the entity B
- *Phase 6:* the entity A sends a message unencrypted randomly generated to entity B.
- *Phase 7:* the entity B encrypts the received message using its private key and sends it. The entity A decrypts the message using the public key of the entity B. At that moment the entity A is sure about the identity of entity B.
- *Phase 8:* same as phase 6, but in the other direction. The entity B sends a message unencrypted randomly generated to entity A.
- *Phase 9:* same as phase 7, but in the other direction. At this moment, the entity B is sure of the identity of the entity A
- *Phase 10:* exchange of information between the entity A and the entity B can be started in complete securely.

However, in VANET, the use of traditional PKI phases is a challenge because of the constraints the response time and the architecture of this network. However, the characteristics and requirements of applications and services require the definition of specific protocols.

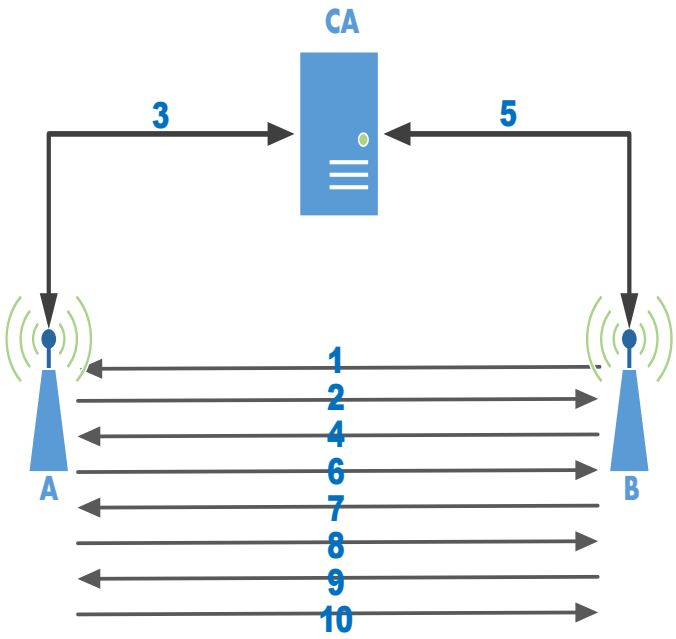


Fig. 1. Key exchange mechanism in the Public Key Infrastructure

IV. The Proposed Procedure

Due to high-speed vehicles, minimizing the end-to-end delay in VANET is highly importance to ensure the proper functioning of services and applications while satisfying the requirements of security in this type of network. However, using traditional security mechanisms cause negative effects on the quality of services and applications because these mechanisms are complex and require a lot of time which leads

to additional delays for the information to reach its destination even if the energy, memory, and computational capacity do not constitute any obstacle in VANET.

To remedy this problem, we propose a new procedure for the exchange of security keys while respecting the requirements of communication in VANET. So for this proposal, we assume that all of the vehicles are grouped into clusters as shown in Figure 2, and each cluster has only one manager node (Cluster Head). This Cluster Head will be responsible for vehicle integration and validation of the security keys.

In the rest of this paper, we use the following notation to describe the proposed procedure.

TABLE I. NOTATION AND SYMBOLS USED

Symbol	***Description***
CA	Certification Authorities
PK_i	The public key of a vehicle i.
SK_i	The private key of a vehicle i.
E(k,M)	The encrypted message M with the key k.
V_R	Random value.
H()	The hash function.

A. *Cluster Schema*

As previously mentioned, the network is divided into clusters as shown in Figure 2. The aim of the cluster is to maximize the lifetime of connections between vehicles, for this, the cluster creation is based primarily on two criteria: the direction and average speed of vehicles. Thus, the cluster has at least one member vehicle and at most one cluster Head. This Cluster Head will assume the role of Certification Authorities (CA).

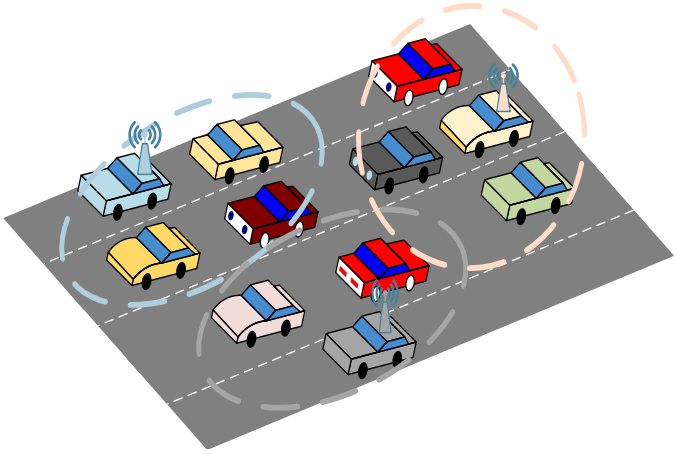


Fig. 2. The formation of clusters

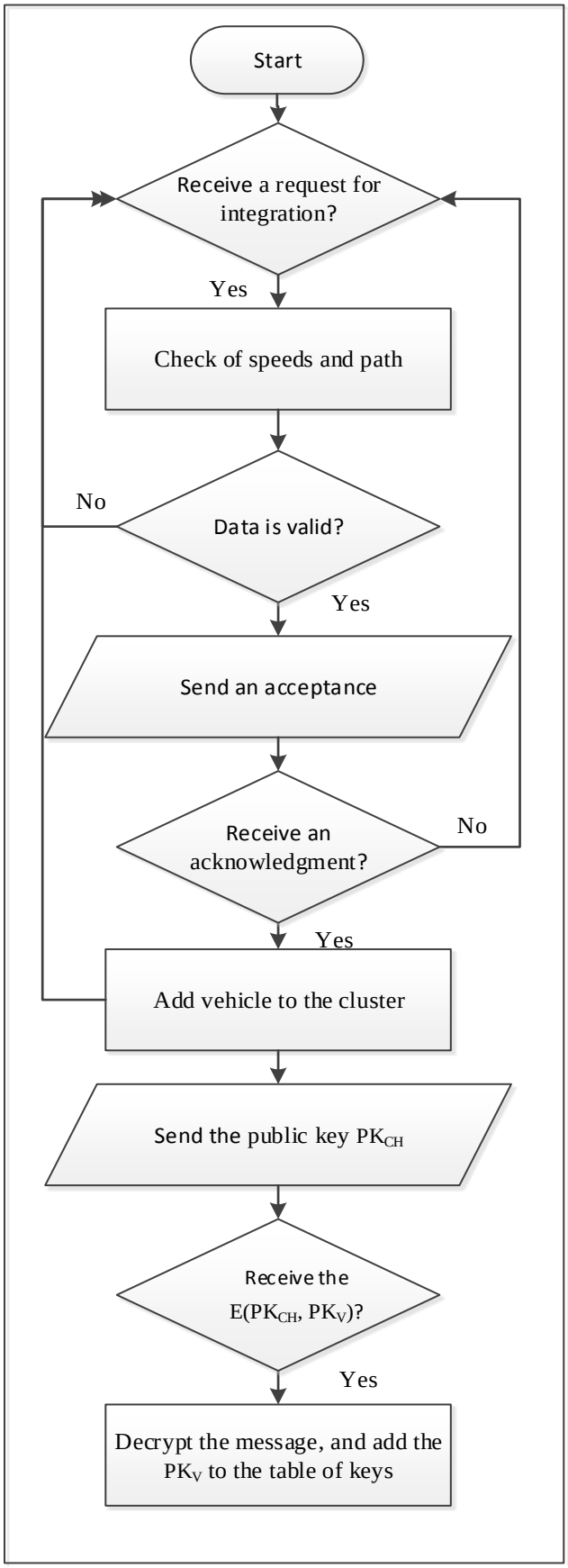


Fig. 3. The integration of a vehicle in the cluster

If the cluster is not yet formally established, and there are at least two vehicles, first, these vehicles will be a check of their speeds and path, the path is the common route segment between vehicles that should be sufficiently long to establish a connection and exchange information, the speed should be around of average speed of the other vehicles so that the vehicle remains in communication with the other vehicles of the same cluster, and the Cluster Head is elected according to its path that must be the longest path on the road compared to other cluster members.

On the other hand, if the cluster is already created, the vehicle broadcasts a request for integration with its speed and path and subsequently, the Cluster Head receives the request as shown in Figure 3. Then, they check the speed and around average speed of the cluster and the road segment in common between the vehicle and the Cluster Head is long enough, whether the Cluster Head sends an acceptance and waits for an acknowledgment. Once the vehicle is integrated into the cluster, the Cluster Head sends its certificate, which contains its public key PK_{CH}, and the vehicle sends $E(PK_{CH}, PK_V)$ to Cluster Head, which represents the public key PK_V encrypted using the public key of the Cluster Head PK_{CH}, the Cluster Head decrypts the message with his private key SK_{CH} and add the PK_V to the table of keys in its database.

B. The exchange of keys between cluster members

The exchange of public keys between the cluster members in the proposed procedure involves three entities.

- The initiator vehicle A
- The responder vehicle B
- The Cluster Head.

Both vehicles A and B are members of the same cluster, so the two vehicles are already identified at the Cluster Head, which owns their public keys.

The sharing of keys applies only to the vehicles A and B, and not the other cluster members, but once the share is finished the keys are stored in the vehicles A and B, which decreased considerably in the exchange of useless messages between vehicles, and thus improves network performance while ensuring the security of the communication.

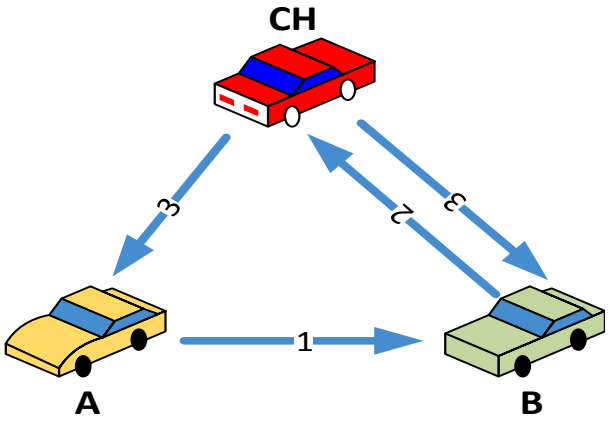


Fig. 4. Exchange public keys between cluster members

The proposed procedure will be conducted in three main phases as shown in Figure 4.

- ***Phase 1***: The vehicle initiator A sends a request to establish communication with the vehicle B. This request contains:
  - The identifier of the vehicle A.
  - A random value V_R generated by A, this value is randomly generated and different for each request of establishing communication.
  - $E(SK_A, V_R)$ the encryption of V_R with the private key SK_A of vehicle A.
  - $H(V_R|E(SK_A, V_R))$ the hash of the V_R and $E(SK_A, V_R)$.
- ***Phase 2***: The vehicle B build its own request to the Cluster Head, which contains:
  - The identifier of the vehicle A.
  - $E(SK_A, V_R)$ the encrypted message sent by the vehicle A.
  - $H(ID_A|E(SK_A, V_R))$ the hash of the V_R and the identifier of the vehicle A.
- ***Phase 3***: The Cluster Head build two messages the one for the vehicle A and the other for the vehicle B, the first message contains:
  - $E(SK_{CH}, PK_B)$ the encryption of the public key PK_B of vehicle B with the private key SK_{CH} of Cluster Head.
  - $E(SK_{CH}, V_R)$ the encryption of V_R with the private key SK_{CH} of Cluster Head.
  - $H(E(SK_{CH}, PK_B)|E(SK_{CH}, V_R))$ the hash of the encrypted PK_B and the encrypted V_R.

  And the second message contains:
  - $E(SK_{CH}, PK_A)$ the encryption of the public key PK_A of vehicle A with the private key SK_{CH} of Cluster Head.
  - $E(SK_{CH}, V_R)$ the encryption of V_R with the private key SK_{CH} of Cluster Head.
  - $H(E(SK_{CH}, PK_A)|E(SK_{CH}, V_R))$ the hash of the encrypted PK_A and the encrypted V_R.

V. Simulation and Analysis

A. Simulation

We choose SUMO (Simulation of Urban MObility) and NS2 (Network Simulator 2) as a simulation platform, in order to test the effectiveness of the proposed procedure. SUMO is designed to manage large real route maps, which can be downloaded from OpenStreetMap, which allow to simulate different scenarios in different parts of the world. SUMO has the ability to operate as a server and to report the simulation data in real time NS2.

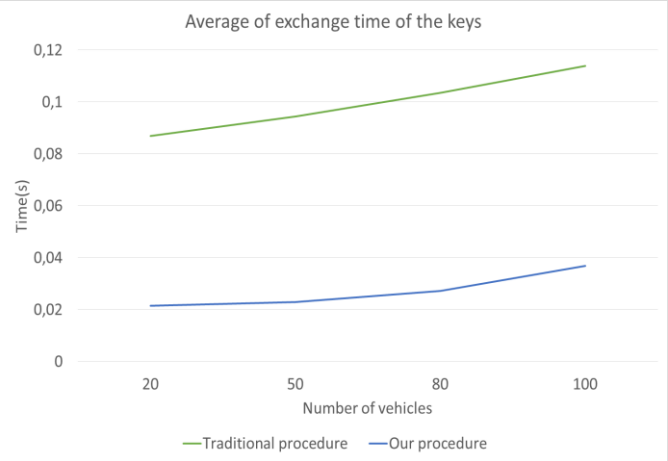


Fig. 5. Exchange public keys between cluster members

SUMO allows the changing of simulation scenarios in NS2 at runtime and thereby provide a dynamic simulation in NS2, and highlight the effectiveness of the proposed procedure.

The purpose of this simulation is to compare the performance of the proposed procedure with the traditional public key exchange procedure. For this, we consider multiple scenarios depending on the number of vehicles, and in each scenario, we randomly selected the path of each vehicle in different places on the map. We used same scenarios for both procedures; we use the following simulation parameters:

TABLE II. SIMULATION PARAMETERS

Map size	1500m x 1500m
Number of vehicles	20,50,80,100
Average speed	15m/s
Simulation time	900s
MAC protocols	IEEE 802.11p
Routing protocols	AODV
Hash function	SHA-1

For both the procedures, and for each scenario, we calculate the average of the exchange time of the keys between vehicles collected during execution.

In Figure 5 which illustrates the results of the simulation, we note that, if the network size increases, with more vehicles, the exchange time of the keys increases slightly in both the procedures. However, the exchange time in the proposed procedure is lower than the time spent in the traditional procedure.

Therefore, the proposed procedure shows a better performance in comparison with the traditional procedure, as the proposed procedure is achieved with only three phases, versus ten phases in the traditional procedure.

Therefore, the proposed procedure can significantly reduce the time to establish a secure communication between vehicles, thus improving the performance of VANET.

B. Security Analysis

The proposed procedure aims to secure the inter-vehicle communication, and thus ensures:

Authentication which consists of verifying the vehicle identity. In the proposed procedure, each vehicle stores an identifier and a pair of keys for secure communication. The signature of each message distributed by the private key provides the authentication of each number.

Integrity, which consists of verifying the integrity of the message when it's exchanged, and not subjected to voluntary or accidental tampering. In the proposed procedure, it is assured by a hash function, which is irreversible.

Confidentiality guarantees that only legitimate message recipient can read it. Therefore, encryption with a public key and decrypted with the private key in the case of receiving the message, and the case of sending encryption messages with the private key and the deciphering with the public key.

Non-repudiation is a much-desired property in VANET because the nature of VANET easily enables you to listen or disrupt the messages exchanged. The attacker can make a replay attack, it is a type of man in the middle attack that consists of intercepting the message and the retransmitted later. Non-repudiation depends on authentication, but the replay attack cannot be confronted by authentication and integrity only. That's why in each request a different random number is generated and included in the request to prevent this type of vulnerability.

VI. CONCLUSION

VANET is a promising technology for the intelligent transportation system. VANET is a promising technology for the intelligent transportation system. However, VANET has many constraints such as the fast moving of vehicles and collective communication medium without any administration by a centralized infrastructure, these constraints combine to make the difficult and complex VANET security to apprehend, and this makes VANET an ideal target for different attacks.

In this paper, a new solution for the exchange of security keys is presented in order to protect information exchanged between vehicles. The proposed procedure can reduce significantly, the delivery time and secure communication and improve VANET performance at the same time.

The proposed procedure is simulated and it has been compared with the traditional procedure keys exchange in several conditions, the experimental result shows that the proposed procedure is very effective. In the future work, we will try to improve the proposed procedure by adding more complexity in different attacks.

REFERENCES

[1] Liang, Wenshuang, Zhuorong Li, Hongyang Zhang, Yunchuan Sun, and Rongfang Bie. "Vehicular Ad Hoc Networks: Architectures, Research Issues, Challenges and Trends." In Wireless Algorithms, Systems, and Applications, pp. 102-113. Springer International Publishing, 2014.

[2] Sumra, I.A.; Bin Hasbullah, H.; Bin AbManan, J.-L., "Effects of attackers and attacks on availability requirement in vehicular network: A survey," in Computer and Information Sciences (ICCOINS), 2014 International Conference on , vol., no., pp.1-6, 3-5 June 2014.

[3] M. Raya and J.-P. Hubaux, "Securing vehicular ad hoc networks," Journal of Computer Security, vol. 15, no. 1, pp. 39–68, 2007.

[4] Kargl Frank, Zhendong Ma, and Elmar Schoch. "Security engineering for VANETs." Proc. 4th Wksp. Embedded Sec. in Cars, pp.15-22, 2006.

[5] Plossl, K.; Nowey, T.; Mletzko, C., "Towards a security architecture for vehicular ad hoc networks," in The First International Conference on Availability, Reliability and Security, 2006. ARES 2006., pp.8 pp.-, 20-22 April 2006

[6] Dhurandher, S.K.; Obaidat, M.S.; Jaiswal, A.; Tiwari, A.; Tyagi, A., "Securing vehicular networks: A reputation and plausibility checks-based approach," in GLOBECOM Workshops (GC Wkshps), 2010 IEEE , vol., no., pp.1550-1554, 6-10 Dec. 2010

[7] P. Golle, D. Greene and J. Staddon, Detecting and correcting malicious data in VANETs, in: Proceedings of VANET'04, 2004, pp. 29–37.

[8] Tiffany Hyun-Jin Kim, Ahren Studer, Rituik Dubey, Xin Zhang, Adrian Perrig, Fan Bai, Bhargav Bellur, and Aravind Iyer. "VANET alert endorsement using multi-source filters". In Proceedings of the seventh ACM international workshop on VehiculAr InterNETworking (VANET '10). ACM, New York, NY, USA, 51-60. 2010.

[9] A. J. Menezes, P. C. Van Oorschot, S. A. Vanstone. " Handbook of Applied Cryptography". CRC press series on Discrete mathematics and its Applications. CRC Press 1997.

A New Particle Swarm Optimization Based Stock Market Prediction Technique

Essam El. Seidy
Department of Mathematics, Faculty of Science, Ain Shams University
Cairo, Egypt

***Abstract*—Over the last years, the average person's interest in the stock market has grown dramatically. This demand has doubled with the advancement of technology that has opened in the International stock market, so that nowadays anybody can own stocks, and use many types of software to perform the aspired profit with minimum risk. Consequently, the analysis and prediction of future values and trends of the financial markets have got more attention, and due to large applications in different business transactions, stock market prediction has become a critical topic of research. In this paper, our earlier presented particle swarm optimization with center of mass technique (PSOCoM) is applied to the task of training an adaptive linear combiner to form a new stock market prediction model. This prediction model is used with some common indicators to maximize the return and minimize the risk for the stock market. The experimental results show that the proposed technique is superior than the other PSO based models according to the prediction accuracy.**

***Keywords*—Computational intelligence; Particle Swarm Optimization; Stock Market; Prediction**

I. INTRODUCTION

Stock market is, without a doubt, one of the greatest tools ever invented for building wealth. Stocks are main element, if not the cornerstone, of any investment portfolio. This demand coupled with advances in trading technology has opened up the markets so that nowadays nearly anybody can own stocks, and use many types of software to achieve the aspired profit with minimum risk. Consequently, a lot of attention has been devoted to the analysis and prediction of future values and trends of the financial stock markets, and due to large applications in different business transactions, stock market prediction has become a hot topic of research. Particle Swarm Optimization (PSO) has become popular choice for solving complex and intricate problems, which are otherwise difficult to solve by traditional methods. The usage of the Particle Swarm Optimization technique in coping with stock market prediction problems is the most important applications of PSO to predict the stocks that have maximum profit with minimum risk. In our earlier paper [1], a new Particle Swarm with Center of Mass (PSOCoM) Optimization algorithm is presented which gives a new efficient search technique. It gets benefit from the physical principle "Center of Mass" to move the particles to the new best predicted position. The new proposed technique improves the performance of the current PSO technique. In this paper, the presented particle swarm optimization with center of mass technique (PSOCoM) is applied to the task of training an adaptive linear combiner to form a new stock market prediction model. This prediction model is used with some common indicators to increase the profit and decrease the risk in stock market.

The survey of the relevant literature showed that there have been many studies for stock market prediction, Many research papers have appeared in the literature using evolutionary computing tools such as genetic algorithm (GA)[2], particle swarm optimization (PSO)[3], and bacterial foraging optimization (BFO)[4] in developing forecasting models. In [5], Hassan et al. described a novel time series forecasting tool, their fusion model combines a Hidden Markov Model (HMM), Artificial Neural Networks (ANN) and Genetic Algorithms (GA) to forecast financial market behavior.

In another work, Aboueldahab, *et al.* [6] introduced a new Enhanced Particle Swarm Optimization (EPSO) to train the Sigmoid Diagonal Recurrent Neural Networks (SDRNN) weights and applied this technique in the forecasting of both NASDAQ100 and S&P500 stock market indices. Majhi, *et al*. [4] used the standard particle swarm optimization (PSO) algorithm to develop an efficient forecasting model for prediction of S&P500 and DJIA stock indices. The connecting weights of the adaptive linear combiner based model are adjusted by the PSO so that its mean square error (MSE) is minimized. Also in [4], Majhi in 2008 developed two new forecasting models based on bacterial foraging optimization (BFO) and adaptive bacterial foraging optimization (ABFO) were presented to predict S&P500 and DJIA stock indices using technical indicators derived from the past stock indices. The structure of these models is basically an adaptive linear combiner, whose weights are trained using the ABFO and BFO algorithms.

A new adaptive-filter method for predicting assets on the stock markets is proposed in [7]. This method is applied through the prediction over the actual valuation of the PETR3 shares (Petrobras ON) traded in the Brazilian Stock Market. The authors evaluated the first-rate choices of the window length and the number of filter coefficient. Observing the correlation between the predictor signals did this and the actual course performed by the market in terms of both the window prevision length and filter coefficient values. It is shown that such adaptive predictors furnish, on the average, very substantial profit on the invested amount.

In [8], Jamous, *et al*. introduced many different forms of PSO which were used for stock market prediction such as Standard Particle Swarm Optimization [9], Linear Decreasing Weight Particle Swarm Optimization (LDWPSO) [10],

Exponential Particle Swarm Optimization (EPSO) [11], Center Particle Swarm Optimization[12], Mean Particle Swarm Optimization [13], and Fuzzy Particle Swarm Optimization (FPSO)[14].

The rest of the paper is organized as follows: in Section II the proposed technique is presented, Section III gives the evaluation of the proposed algorithm is presented, and Section IV concludes this paper with a summary of main points.

II. The Proposed Stock Market Prediction Technique

In this section, the proposed technique is described. The structure of the proposed stock market prediction technique is assumed to be an adaptive linear combiner with parallel inputs as shown in Figure1. It is an adaptive finite impulse response (FIR) filter with number of inputs equal to the number of features in the input patterns. These features are abstracted from the stock market series such as closing prices and technical indicator values. The connecting weights of the adaptive linear combiner are considered as the particles, and initial their values are set to random numbers in the range [-1, +1]. The swarm of particles is chosen to represent the initial solutions of the model. Each particle is adjusted during the training step by the way of minimizing the mean square error (MSE) as an objective function for PSOCoM technique. To give a clear sight about the methodology of proposed prediction model, let N represent the number of patterns (e.g. 100 days training set),and D is the size of an input pattern to an adaptive linear combiner (e.g. D = 8 means one day ahead obtained from the past stock prices plus seven technical indicator values related to this day), which equal the number of adjusted weights and also the dimension of the particles, so that every eight values (one day ahead price plus seven indicator values) are passed through an adaptive linear combiner, and multiplied with weights of an adaptive linear combiner and the partial sums are added together to give $y_i(k)$ as an output for the combiner. Then, this output is compared with the corresponding desired stock price d(k) to produce the error $e_i(k)$. A shift one day forward produces new error until reaching the end day in training set (100 days) is reached. After that, each produced error is squared and added to the others using the accumulator shown in Figure1. The summation is divided by number of patterns to give the mean square error for the i_{th} particle as shown in equation1 which is the objective function of the PSOCOM technique, so that the aim is to minimize this mean square error for best training.

$$MSE_i = \frac{\sum_{i=1}^{N} e_i^2(k)}{N} \quad (1)$$

It is important to refer that the previous scenario considers one day ahead closing price with its seven indicators values to train the prediction model, so when five days ahead closing price with their seven indicators values are used for training, the dimension of the particles will be D = 35. According, the number of connecting weights is equal to 35.

However, in the prediction step, the optimized weight values, obtained by PSOCoM technique, are used to give the predicted price for the same forecasting stock price through an adaptive linear combiner.

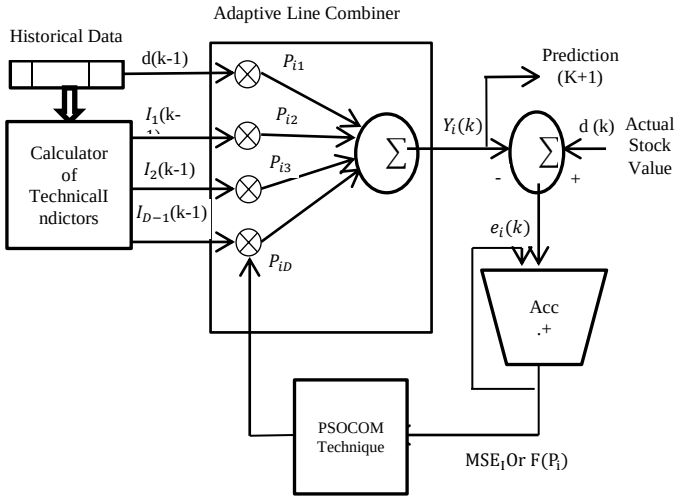


Fig. 1. The proposed stock market prediction model

If there is M number of test patterns, the mean absolute percentage error (MAPE) is used to measure the performance of prediction model during the forecasting stock prices as follows:

$$MAPE = \frac{1}{M}\sum_{k=1}^{M}\left|\frac{d(k)-y(k)}{y(k)}\right| \times 100 \quad (2)$$

Where: M is the number of test patterns.

Table I describes the indicators' equations, which are used in this paper. All previous indicators' equations were abstracted from "Encyclopedia of Technical Market Indicators" book published by Colby [15].

TABLE I. Technical Indicators Used with the Proposed Algorithm

Technical Indicators	Formula
Exponential Moving Average (EMA) (EMA10) (EMA20) (EMA30)	EMA = Previous EMA + (P–Previous EMA)*A where A = 2/(N+1) , A is smoothing factor, N is time period, P is current price. EMA10 = preEMA10 + (P- preEMA10)*(2/11) EMA20 = preEMA20 + (P- preEMA20)*(2/21) EMA30 = preEMA30 + (P- preEMA30)*(2/31)
Simple Moving Average (SMA) (SMA10)	$SMA10_j = \frac{\sum_{i=j-10}^{j}(CP_i)}{10}$, j the day to be calculated its SMA10 CpiClosing Price of day i
Relative StrengthIndex (RSI) (RSI9) (RSI14)	$RSI = 100 - \frac{100}{1+(U/D)}$, U = (total gain)/n D = (total losses)/n. n is number of RSI period.
Price Rate Of Change (PROC) (PROC27)	$\frac{(Today's\ Close - Close\ X\text{-}period\ ago)}{(Close\ X\text{-}period\ ago)} \times 100$

III. Evaluation of the Proposed Stock Market Prediction Technique

In this section, the performance of the proposed technique is evaluated. The historical data of used indices and the values of parameters settings are described. Finally, the results and their discussion are presented.

A. Historical Data

The historical data of three common indices, namely, National Association of Securities Dealers Automated Quotations 100 (NASDAQ-100), Dow Jones Industrial Average (DJIA) and Standard's & Poor's 500 (S&P 500), are used in this experiment for the evaluation of the proposed prediction model. These historical data consist of daily close prices and technical indicators derived from those indices. Total number of samples for the stock indices is 2500 trading days, from 2 January 2005 to 31 December 2014. Each sample consists of the opening price, highest price, lowest price, closing price and the total volume of the stocks traded for the day.

B. Parameter Settings

The same set of parameters is applied to the compared prediction models, namely, ALCPSO, ALCLDWPSO, ALCCenterPSO, ALCMeanPSO and the proposed model, inertia weight w is linearly decreased from 0.9 to 0.4, and isfixed at 0.9 in PSO and Mean PSO. The acceleration coefficients are set to $c_1 = c_2 = 2$.the maximum velocity is set to Vmax= 0.5 and Xmax= 1. The swarm size is set to 30. The maximum number of iterations was set to 100. Initialization is range of particle positions was $-1 \leq xi \leq 1$. All mean square errors (MSE) and mean absolute percentage errors (MAPE) are computed over 30 runs. The seven common technical indicators used for this evaluation are EMA10, EMA20, EMA30, SMA10, RSI9, RI14 and PROC27. In short term prediction experiment, the training period was set to 100, 200 and 500 days to predict test period of 100 days. In long term prediction experiment, the training period was set to 1000 and 1500 days to predict test period of 750 days.

C. Results and Discussion

There are two types of prediction to evaluate the proposed prediction model, short- and long-term prediction. Various experiments are carried out by varying the selection of technical indicators as a new feature with closing price to the inputs of the models. As a result, the best set of used indicators, which produced more accurate prediction are: EMA30, RSI14 and PROC27. These sets of indicators are applied to all calculations in this experiment. To clarify the learning characteristics of the compared models in short and long term, the mean square error (MSE) is considered as a measure during the training process. In short term prediction, Figure 2 to Figure 4 show the learning characteristics of the compared models obtained for one day advance with three technical indicators EMA30, RSI14 and PROC27, to predict DJIA, NASDAQ-100 and S&P500 stock indices, respectively. Figure 5 to Figure 7 show the learning characteristics of the compared models for long term. It is noted that the proposed PSOCoM converged faster than the other versions of PSO during the training process and reached the best minimum value of MSE indicating the convergence of the weights. This shows that the proposed PSOCoMis superior than the other PSO versions in learning characteristics, and in abstracting the important feature during training to perform more accurate prediction.

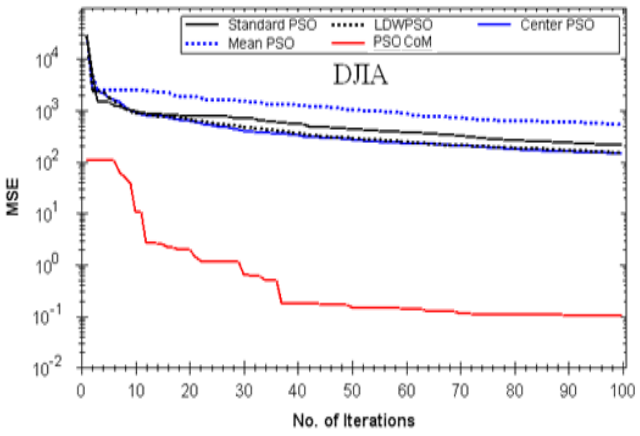


Fig. 2. Learning characteristics of compared models to predict DJIA for one day advance (short term prediction)

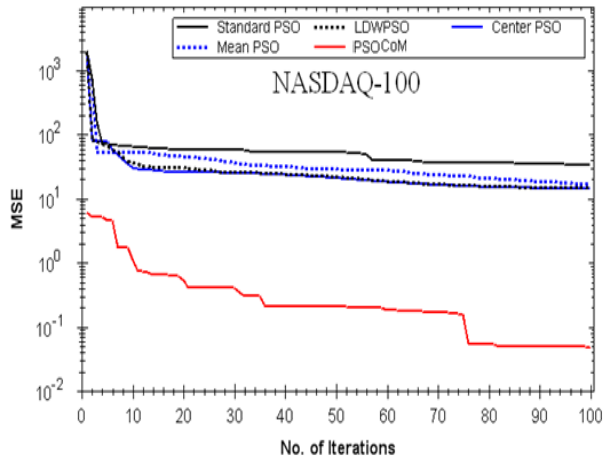


Fig. 3. Learning characteristics of compared models to predict NASDAQ-100 for one day advance (short term prediction)

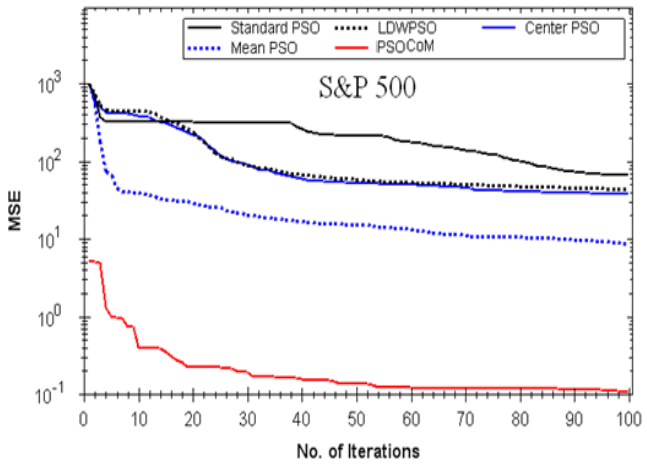


Fig. 4. Learning characteristics of compared models to predict S&P500 for one day advance (short term prediction)

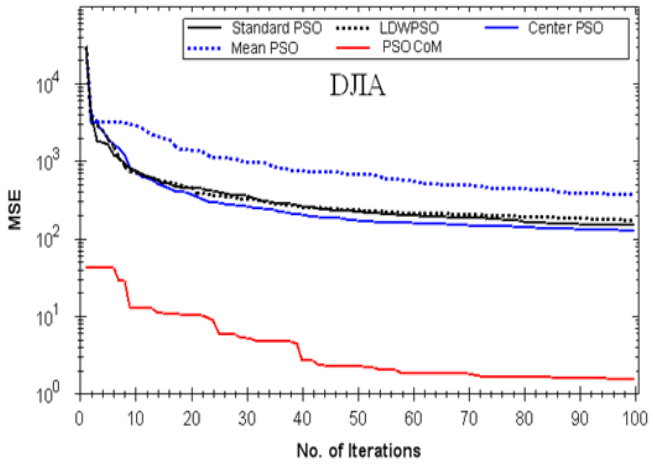


Fig. 5. Learning characteristics of compared models to predict DJIA for one day advance (long term prediction)

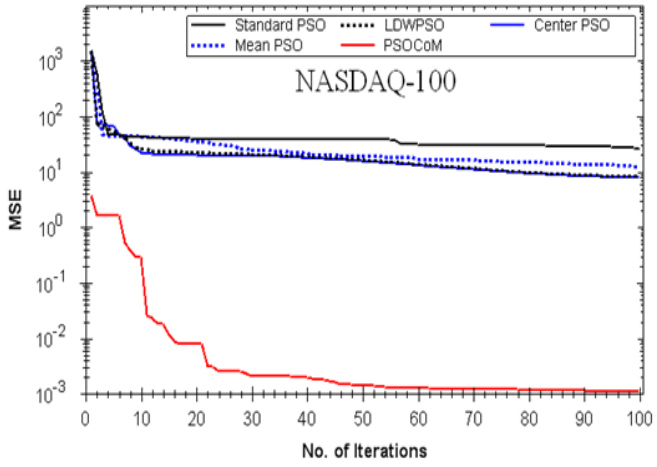


Fig. 6. Learning characteristics of compared models to predict NASDAQ-100 for one day advance (long term prediction)

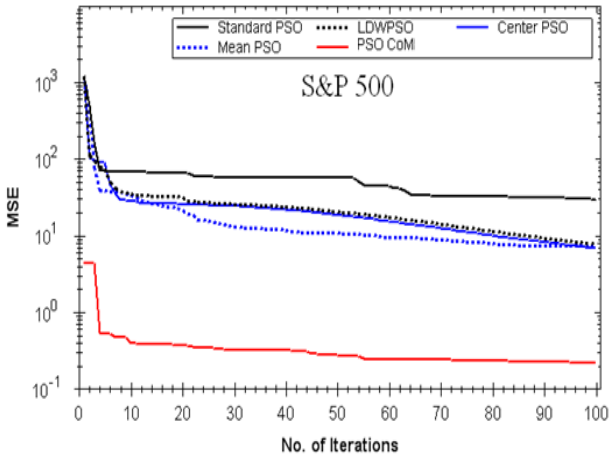


Fig. 7. Learning characteristics of compared models to predict S&P500 for one day advance (long term prediction)

Table II shows the best mean absolute percentage error (MAPE) obtained by the compared forecasting models for short and long term prediction. The comparative results of prediction the DJIA, NASDAQ-100 and S&P500 stock indices were obtained for one day and seven days ahead prediction over 30 runs. Observations of Table I indicate that the MAPE of the proposed predication model is the lowest between the compared PSO based models for all the stock market indices forecasting. This shows that the proposed forecasting model superior to the other PSO based models according to the prediction accuracy. It is noted that the MAPE obtained for short term prediction is less than that obtained for long term prediction, for example, the MAPE of short term prediction for DJIA index is equal to 0.0325%, while the MAPE of long term predication for DJIA index is equal to 0.8601%. Thus, the ratio between the short term and the long term MAPE predication equals to 0.1. This shows that all the PSO based forecasting models in addition to the proposed forecasting model give accurate prediction for short term prediction, while the prediction accuracy decreases for long term prediction. Back to the historical data for any index in stock markets, as known to all investors in financial stock markets, the last period of any stock can give important information about the next coming move of that stock, For short term investment the best selected period for abstract important information that guides to good prediction is three months to one year. Any information abstracted out of this period may disperse the investor and wrong prediction may be take place. On the other side, the best period for long-term investment is from one to three years. However, the previous notes verifies the obtained results as shown in Table II, where the MAPE increases as the training period increases for both of short- and long-term prediction. According to the complexity of the compared forecasting models, Table II shows that one day and seven days ahead used in all the calculations, where one day ahead indicates that there are four inputs to the adaptive linear combiner (one day ahead close price and three selected indicators). Consequently, the dimension of the particle is equal to four (the connecting weights of the adaptive linear combiner are equal to four), while for seven days ahead there are 28 inputs to the adaptive linear combiner (seven days closing prices and three selected

indicators for that seven days), so the dimension of the particle becomes 28 (the connecting weights of the adaptive linear combiner are equal to twenty eight).

As a result, when the particle dimension increases (more complexity) the MAPE of the compared forecasting PSO-based models increase (prediction accuracy decreases), while the proposed forecasting model gives almost the same MAPE for the two degree of complexity (one day and seven days ahead).This means that the optimal or near optimal solution (optimal values of connection weights in the adaptive linear combiner) is reached by the proposed PSOCoM technique. This shows that the prediction accuracy of the proposed forecasting model is almost the same while the complexity increases.

IV. Conclusion and Future Work

A novel stock market prediction technique has been proposed. Also, a new stock market prediction model based on the proposed PSOCoM technique has been provided. This prediction model uses PSOCoM technique to adjust the weights of an adaptive linear combiner. The results of the experiments showed that the proposed forecasting model is superior than the other PSO based models according to the prediction accuracy.

The PSO based forecasting models in addition to the proposed forecasting model give accurate prediction for short term prediction, while the prediction accuracy decreases for long term prediction. The MAPE obtained by prediction models increases as the training period increases for both of short term prediction and long term prediction. As a result, the proposed forecasting model is a new promising forecasting model for stock market prediction. In the future, based on the proposed prediction technique, it can design a new selection technique to select the best stocks with highest profit and minimum risk.

Furthermore, a new automated system can be developed based on the proposed work to become an intelligent agent that makes trades in stock markets to get maximum profit with minimum risk, gives the decision of buy and sell for the best selected stocks, and gives the final profit at the end of the required period.

References

[1] R. Jamous, E. El. Seidy, A. Tharwat, B. I. Bayoumi" Modifications of Particle Swarm optimization Techniques and Its Application on Stock Market: ASurvey", in (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 6, No. 3, 2015,PP:99-108.

[2] R. Hassan, Nath, B. and Kirley, M., "A fusion model of HMM, ANNand GA for stock market forecasting", Expert Systems with Applications,Vol. 33, 2007, pp. 171 -180.

[3] T. Cura., "Particle swarm optimization approach to portfolio optimization", Nonlinear Analysis Real World Applications, Vol. 10, Issue 4, 2009, pp. 2396-2406.

[4] R. Majhi, G. Panda, *, B. Majhi, G. Sahoo" Efficient prediction of stock market indices using adaptive bacterial foraging optimization (ABFO) and BFO based techniques "in Expert Systems with Applications, Vol .36 , 2009.pp. 10097–10104.

[5] R. Hassan, B. Nath, and M.Kirley, "A fusion model of HMM, ANN and GA for stock market forecasting", Expert Systems with Applications, Vol. 33, 2007, pp. 171-180.

[6] T. Aboueldahab, and M. Fakhreldin, , "Stock Market Indices Prediction via Hybrid Sigmoid Diagonal Recurrent Neural Networks and Enhanced Particle Swarm Optimization", International Congress for global Science and Technology, ICGST, Vol. 10, 2010, pp. 23-30.

[7] J. E. Wesen, V. Vermehren, H. M. de Oliveira" Adaptive Filter Design for Stock Market Prediction Using a Correlation-based Criterion", arXiv preprint arXiv: 1501.07504,2015.

[8] R. Jamous, E. El. Seidy, A. Tharwat, B. I. Bayoumi" A new Particle Swarm with Center of Mass Optimization", in International Journal of Engineering Research & Technology (IJERT), Vol. 4 Issue 05, 2015, PP:312-317.

[9] J. Kennedy, J. and Eberhart, C., "Particle Swarm Optimization".Proceedings of the 1995 IEEE International Conference on Neural Networks, Australia, 1995, pp. 1942-1948.

[10] D. Alrijadjis, K. Tanaka, and Mu S., "A Modified Particle Swarm Optimization with Nonlinear Decreasing Inertia Weight Based PID Controller for Ultrasonic Motor" , International Journal of Innovation, Management and Technology, Vol. 3, No. 3, June 2012.

[11] N. I.Ghali, N. El-Desouki, Mervat, and A. N. Bakrawi, "Exponential Paticle Swarm Optimization Approach for Improving Data Clustering", Proc. World Academy of Science, Engineering and Technology, Vol. 32, 2008, pp. 56-60.

[12] Y. Liu, Qin, Z., Shi, Z. and Lu, J., "Center particle swarmoptimization," Neurocomputing Vol. 70, 2007, pp. 672-679.[12] K. Deep, and Bansal, J.C. "Mean particle swarm optimizationfor function optimization," Int. J. Computational Intelligence Studies, Vol. 1, No. 1, 2009, pp.72-92

[13] K. Deep, and J.C. Bansal, "Mean particle swarm optimization for function optimization," Int. J. Computational Intelligence Studies, Vol. 1, No. 1, 2009, pp. 72-92.

[14] L. A. Zadeh, , "Fuzzy sets". Information and Control. Vol. 8, 1965, pp.338–353.

[15] R. W.Colby , "The Encyclopedia of Technical Market Indicators", McGraw-Hill Press, 2003.

TABLE II. COMPARATIVE RESULTS OF MEAN ABSOLUTE PERCENTAGE ERROR (MAPE) AVERAGED DURING PREDICTION PROCESS OVER 30 RUNS

Days ahead	***Predicti on Period***	***Training Period***	***Mean Absolute Percentage Error (MAPE)***					Stock Index
			ALC-PSO	***ALC-LDWPSO***	***ALC-CenterPSO***	***ALC-MeanPSO***	***ALC-PSOCoM***	
1	(Short term) 100	100	0.0281	0.0279	0.0278	0.0286	0.0271	DJIA
		200	0.0289	0.0287	0.0287	0.0295	0.0283	
		500	0.0335	0.0331	0.0332	0.0348	0.0325	
	(Long term) 750	1000	0.5833	0.5847	0.5820	0.5964	0.5761	
		1500	0.8693	0.8688	0.8661	0.8888	0.8601	
7	100	100	0.0298	0.0296	0.0288	0.0287	0.0271	
		200	0.0303	0.0298	0.0318	0.0296	0.0283	
		500	0.0345	0.0347	0.0344	0.0339	0.0325	
	750	1000	0.6263	0.6828	0.6510	0.5970	0.5762	
		1500	0.8949	0.8864	0.9560	0.8797	0.8603	
1	100	100	0.0371	0.0369	0.0368	0.0371	0.0343	NASDAQ100
		200	0.0384	0.0382	0.0381	0.0385	0.0358	
		500	0.0462	0.0449	0.0450	0.0454	0.0411	
	750	1000	0.6898	0.6777	0.6769	0.6835	0.6641	
		1500	1.0318	1.0157	1.0155	1.0238	0.9920	
7	100	100	0.0596	0.0518	0.0519	0.0393	0.0343	
		200	0.0581	0.0518	0.0543	0.0405	0.0357	
		500	0.0795	0.0661	0.0667	0.0469	0.0411	
	750	1000	0.9933	0.8964	0.9271	0.7195	0.6641	
		1500	1.4894	1.3306	1.3701	1.0595	0.9920	
1	100	100	0.0419	0.0390	0.0386	0.0335	0.0304	S&P500
		200	0.0411	0.0383	0.0382	0.0345	0.0318	
		500	0.0436	0.0396	0.0399	0.0402	0.0364	
	750	1000	0.7062	0.6520	0.6517	0.6511	0.6343	
		1500	1.0922	1.0621	1.0417	0.9636	0.9471	
7	100	100	0.0803	0.0662	0.0627	0.0362	0.0308	
		200	0.0743	0.0608	0.0593	0.0358	0.0319	
		500	0.1030	0.0811	0.0728	0.0422	0.0370	
	750	1000	1.2639	1.1024	1.0882	0.6750	0.6359	
		1500	1.8301	1.5534	1.4934	0.9910	0.9515	

Devising a Secure Architecture of Internet of Everything (IoE) to Avoid the Data Exploitation in Cross Culture Communications

Asim Majeed
School of Computing and Digital Technology
Birmingham City University (BCU)

Anwar Ul Haq
Department of Computer Science
QA Higher Education (ULST)

Rehan Bhana
School of Computing and Digital Technology
Birmingham City University (BCU)

Mike-Lloyd Williams
Department of Business Administration
QA Higher Education (ULST)

***Abstract*—The communication infrastructure among various interconnected devices has revolutionized the process of collecting and sharing information. This evolutionary paradigm of collecting, storing and analyzing data streams is called the Internet of Everything (IoE). The information exchange through IoE is fast and accurate but leaves security issues. The emergence of IoE has seen a drift from a single novel technology to several technological developments. Managing various technologies under one infrastructure is complex especially when a network is openly allowing nodes to access it. Access transition of infrastructures from closed networked environments to the public internets has raised security issues. The consistent growth in IoE technology is recognized as a bridge between physical, virtual and cross-cultural worlds. Modern enterprises are becoming reliant on interconnected wireless intelligent devices and this has put billions of user's data in risk. The interference and intrusion in any infrastructure have opened the door of public safety concerns because this interception could compromise the user's personal data as well as personal privacy. This research aims to adopt a holistic approach to devising a secure IoE architecture for cross-culture communication organizations, with attention paid to the various technological wearable devices, their security policies, communication protocols, data format and data encryption features to avoid the data exploitation. A systems methodology will be adopted with a view to developing a secure IoE model which provides for a generic implementation after analyzing the critical security features to minimize the risk of data exploitations. This would combine the ability of IoE to connect, communicate, and remotely manage an incalculable number of networked, automated devices with the security properties of authentication, availability, integrity and confidentiality on a configurable basis. This will help clarify issues currently present and narrow down security threats planning considerably.**

Keywords—privacy; privacy enhancing technology (PET); big data; information communication technology (ICT)

I. INTRODUCTION

The Internet of Everything (IoE) can be defined as the products and systems which are communicating and interacting with the environment, users and another system through the communication networks. The emergence of IoE has integrated various diverse type networks and wireless communication technologies under one platform [3]. The new open communication relationship among devices has complicated the trust relationship and raised security issues within communication systems and the heterogeneous entities. The IoE based organizations require a novel security architecture to be laid out after analysing the existing ICT infrastructure to solve these security issues [5]. The IoE among cross-cultural organisations is growing at an alarming pace and meeting the security demands is becoming hyper-complex since the advancement in capabilities of smart technologies. The cross-cultural communication creates vulnerabilities and cyber security challenges depending on the communication processes, products and security of data; consequently have a high impact on economic growth [19]. The integration of various devices on a multichannel enhances users experience but positions the organisation's interface where intruders could exploit the data. Organisations operating in various sectors of the world have potentially many business partners, advisers, customers and closer collaborations exchanging a significant amount of data with each other. This not only enriches the product development and recruiting experience but leaves information's flaws in complex data handling. Cross-cultural organisations using hybrid delivery models run processes and business services through the cloud; managed by external providers [10]. The hybrid models help organisations to look at the activities through IoE communication model and extend the security perimeter to detect and monitor cyber security attacks.

Cross-cultural awareness and understanding are becoming increasingly important in the modern era. The study conducted by Botha et al, [3] showed that young people are particularly comfortable in sharing their experiences and cultural signatures through mobile technology and SMS services. Smartphones were at the forefront of the technology from the late nineties until now. Increasingly smart devices and wearable technologies are driving a new technological revolution [18]. These devices are capable of using sensor technologies to monitor, alert, automate the processes and activities in our personal and work lives. The world is increasingly becoming

more global and the advent of new digital technologies is constantly diminishing the barriers of space and distance among communities. At a global scale, this phenomenon is presenting new challenges in terms of how to increase the awareness of the cultural sensitivities and safe-use in the new digital era of Internet of Everything (IoE) [12].

The evolution of computers from mainframes to PCs, the transformation into ubiquitous computing with the emergence of Wireless sensor networks lead to wide industry adoption of Internet of Everything [13]. Due to this rapid evolution process, Internet of Everything has become an integral part of our life in the form of smart homes, smart healthcare, and smart automobiles [17]. Similarly, this advancement in technology is becoming de facto standard for businesses to achieve their key performance indicators and remain on the cutting edge in this competitive market [13]. Although currently customer-centric approach is helping businesses to create positive customer experiences with the help of analytic techniques, which analyses Big Data and can add value to a company, a more intelligent approach is required to deal with real data involved in Internet of Everything [15]. It is expected that the number using the Internet of Everything, will grow up to 50billion by 2020. This is due to the fact that transitioning to Internet of Everything by adding intelligence to data, allowing continuous monitoring, updating and controlling it in a real time improves the operational decision-making process of business [8].

II. 'IOT-IZING' THE BUSINESS

The adoption of latest technologies is slow particularly in small businesses but IoE integration has envisaged all size businesses to add real value to their communications and day to day processes [2]. Modern businesses are required to be proactive to build a frame around of how they can stay IoT-ized especially in meeting the cross-cultural communication needs. As soon as an organization starts thinking about moving their internal and external communications and processes on IoE related technologies, they would need to think investments on resulting data, volume of data connectivity, infrastructure support, data intelligence and sensors [20]. Consequently, businesses would need to think about staff training of using the IoE technologies to take the full advantage of going IoT-ized [18]. The integration of IoE technology based infrastructure would also help cross culture staff training to stay up to date on the updates and changes taking places within the organizations. The journey of going IoT-ized would bring unexpected and unpredictable challenges in real time situations but cross culture conflicts and consortiums could be resolved to share best practices using IoE paradigm [12]. The management of cross-culture communications using IoE technologies to connect more and more devices would bring more opportunities for cyber criminals as well as hackers [6].

It is very important for all size businesses to consider the security threats to avoid risks of data exploitation. If businesses are using some devices for communication and recording, there is a huge risk of these devices can be hacked and information recorded in this device could be exploited [5]. These threats should be embraced as a challenge to the organization and design a framework which could authenticate and authorize the secure users only and if the infrastructure triggers any caution about a unauthenticated device, the access should not be allowed. Capgemini's, [4] survey shows the Internet of Everything (IoE) present a business opportunity for a trillion-dollar industry and growth of new industries to cater for this shift where technology infused the world is a norm. The 71% of executives (related to IoE industry) raise their concern on security threats and related consequences on the growth of this business and opportunity it presents [4]. Only 33% of executives in the survey believed that the current IoE based products and services are resilient to cyber security attacks [4]. One of the key factors for the increase in security threat is the fact that IoE based products and system increase the potential attack points in a system [5]. The users awareness and patterns of behavior could play a major role in safe-use of IoE based products and systems [18]. The understanding of cultural behavior and patterns of communities will also play a key part in the growth of IoE based industry, especially when a culture of one community may affect the culture and behavior pattern of another community in terms of educating and informing the safe use of IoE based products and systems [1].

Fig. 1. IoT-izing and Cross Culture Communication [4]

The adoption of Internet of Everything provides business data intuitiveness, which was never possible before [12]. Increased processing power of server machines, super-fast internet connection, and massive use of smart devices with their falling costs, seamless business to business communication and development of applications lead the businesses to adopt cloud-based solutions, to help achieve scalable, flexible and low-cost solutions to improve their customer experience [3]. Just establishing IT infrastructure and connecting to The Internet is not enough, the adoption of IoE and cloud services is also required for a business to improve its informed decisions by the stakeholders [2]. Cloud services allow storing and analyses of business data coming from different streams [4]. Internet of Everything will constantly generate new data, which can be used to enhance the business key performance indicators such as customer services [6]. In order to gain an advantage of the Internet of Everything, companies should proactively plan to which extent they can be 'IoT-ized ' [3]. This can be done by focusing on the installation of infrastructure and employee training, so they can handle both internal processes and customer's queries [9]. Internet of

Everything is all about the connection between devices and exchanging of data, which means there are increased security threats to data and devices [5]. As more and more new devices are connected to IoE, people must be made aware of how to implement security measures while connecting these devices; they also provide new opportunities to the hackers because the experts are also exposing more vulnerabilities [4].

A. Privacy in IoE

Since the IoE has become so widespread, the smart devices know more and more about how to collect our data, therefore, we should also be aware of how they are monitoring and collecting our data and spying on us without our consent [8]. Security and privacy are one of the critical concerns individuals have. The EU Commission's paper on Internet of Everything Governance also highlights the implementation of security controls to minimize cyber-attacks and individual surveillance [14]. This does not mean that Internet of Everything should be avoided but rather a cautious and planned approach should be taken [2]. The dawn of internet has raised the concerns over privacy preservation. When organizations are communicating cross-culturally through the IoE medium, many applications used by the devices will exacerbate the problem of leaving trails of communication, traceable signatures, locations and the individual's behaviors [6].

The privacy concerns of healthcare organizations are more relevant as they run many applications through IoE. The hospital management systems may require the tracking of medical equipment or the monitoring of patient's vital statistics within assisted living facilities or at home. In this situation, the new IoE devices which require association and decoupling with the owner should authenticate the security check so to identify the device. A mechanism of shadowing has been proposed to look at the data security [8]. The user objects use digital shadows which store the virtual identity of the device in terms of its attributes and information [19]. The association of diverse authentication methods for machines and humans would offer new opportunities to identify the device identity and increase security. The door of personal networks could be opened for an object combining it with bio-identification [20]. Different countries have different views on compliance and privacy especially since technology is consistently evolving on a daily basis and cross culture organizations need to be cognizant of how these matters and issue would apply to them [11].

III. Cross Culture Communication and IoE

In order to meet the current demands, businesses are advancing their technologies in both software as well as hardware. Various researchers and IT experts have warned that this model is going to be changed in the future especially in terms of IoE advances when looking at cross-cultural aspects. This model would lead to the concept of generating revenue not only from hardware but from its use of on a cross culture communication basis. The model of freemium subscriptions would be the preferred choice in the IoE era of cross culture organizations [9]. The assumed model would raise many security issues relating to user's data. The services designed around hardware would be more amenable to ecosystems and easily upgradable providing multiple opportunities to generate revenue [14]. The evolution of a service-centric model could result in cross culture businesses struggling to ensure that they prioritize processes in order that protecting user data is easy as well as secure and transparent. Organizations would benefit from this customer-centric communication in terms of keeping track of customer loyalty program information, payment methods and purchase history [7]. This information would help organizations to improve customer experience as well as creating a solid foundation for monetization and data security [16].

IV. Elements of a Secure Architecture: Cross Cultural Business Perspective

The basic principal and central approach of IT security should be to design a secure infrastructure instead having additional layers of the existing architecture [20]. In relation to design a secure IT infrastructure for cross-culture communication, following principals need to consider:

A. Alignment of Business Domains and Security Requirements

A traditional IT infrastructure is designed in alliance with business processes and domains. In particular, if we talk about the retail businesses their domain may be based on the entire value chain from store management to supply chain management [10]. On the contrary, the IT infrastructure design has to look at both the perspectives of risk exposures to existing assets and business processes in each domain. The security element should be embedded and made an integral part of the architecture rather than making it more complex after adding more security layers [7].

B. Grouping by Capability

The ICT infrastructure is made secure and manageable on the basis of similar privileges level for users [2]. The privileges are assigned to particular groups of security and business domains. The risk is assessed on processes and assets of the organisations through the capability level and if it requires, more consistent and adequate securities these are assigned to various groups [11]. The homogenous level of protection is obtained after adding capabilities to security domains

C. Modularity

The modularity part deals with adjusting the security level of domains without affecting the other domains [9]. The business encompasses various domains with different security levels and modular structure as this helps to adequately measure the risk and at the same time provides protection as well. The infrastructure security could be increased by deploying the pivotal points at various nodes to monitor the technology. Devising a secure interface only between a corporate network and public internet is insufficient [18]. The threats of hijacking the network after connecting and penetrating in the infrastructure would grow. These threats would not be protected by the outside network guards and require some inside topologies to be devised to keep it secure through triggers [13]. As soon as some users get connected with the IoE, an extra security layer should be activated which detects attacks. The system should be designed in an intelligent way, which consistently observes the inside activities, detects user behavior change and alerts the infrastructure. Once the

network is divided into security domains, it brings multiple benefits to detecting threats [4]

Information is a valuable source and most modern businesses rely on effective use of information for their processes, market reach, customer satisfaction and competitive advantage [9]. This demand for the valuable information puts a strain on privacy and data related to personal liking, disliking, and behaviour. Etc. The information system has brought huge success to businesses in achieving their goals. The information system gathers process, distribute, utilise and interact with information [6]. The success of information systems is dependent on channelling communications effectively between different components of such system including people. The information security is an established discipline and with well-defined procedures and measures to this effect.

V. Existing Architecture Limits

Time and the budget have always been a pressure on modern organizations even though they are willing to invest heavily to secure their IT infrastructure [8]. These constraints lead them not to integrate security triggers inside the infrastructure but layering a new security infrastructure on top of their existing IT architecture. This addition creates ring-fenced, haphazard and heterogeneous architectural landscape which requires vast system updates and manual intervention to maintain it [15]. The purpose is to develop a secure architectural infrastructure but instead, this approach creates unanticipated gaps as well as complexity in a cross culture communication environment. There will be challenges if organizations roll out automated and digitized services quickly [10]. The coding of planned pilots through the cloud should have been monitored before the launch and along the appropriate consideration of the existing landscape. The safe testing area should have been created otherwise organization would end up risking their IT infrastructure [9].

VI. A Journey Towards Devising a Secure IoE Architecture

The capabilities of secure enterprise architecture are identified through an initial security assessment and classified by threat level [1]. The most critical business assets such as underwriting data and trading algorithms are analyzed to identify security gaps. The compromise on the security gaps could lead to reputational harm as well as material losses. The processes and assets of high-risk and high-value nature are separated on the basis of threat based classifications but cross-culture communications still benefit through virtual environments and shared infrastructure [19]. Various applications and servers could be used to run the organizational website through a separate authorization engine to process the high-value financial transactions within the cross-culture communications. The activities to support process and data steps for online money transfer or other business transactions are classified under discrete capabilities [7]. The adequate level of risk and protection could be determined through the analysis of security zone architecture within the cross culture communication. The risk impact of breaching can be estimated through the regulatory, competitive, financial, reputational and operational processes of the organization [3]. The risk can also be estimated through the process downtime as this mishandling of customers personal information could lead to regulatory fines [4].

VII. Model of Secure Architecture for Cross Culture Communication: : Technical Perspective

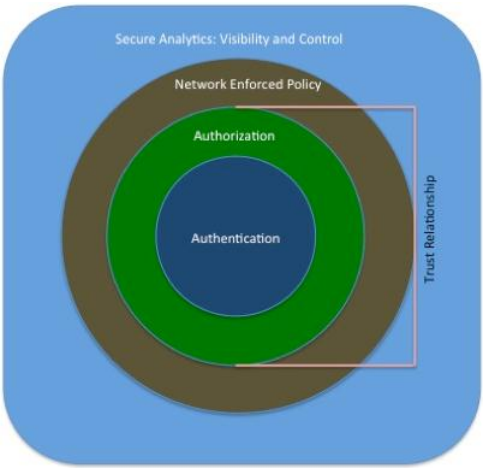


Fig. 2. Secure IoT Framework [20]

A. Authentication

The authentication layer is the central part of this framework which could be used to identify and verify the IoE entity information. As soon as the IoE devices start establishing the connection to each other, they require getting connected to IoE infrastructure [2]. The identity of the device should determine the trust relationship. Various IoE devices may have substantially different ways of storing, managing and presenting the information. It is noted that eligible users in organizations access the network for both local and cross culture communication through human credentials of password and username [4]. In terms of IoE the endpoints should be setup through fingerprint means, so not to require human interaction. The embedded sensors within the IoE devices should set artificially intelligent to scan and then recognize the user identity based on the particular device storage mechanism [12]. The X.509 certificates could also be used to establish a strong authentication system to establish this identity. The X.509 certificates are cryptographic and require enough memory to be executed consequently it may not be possible for various IoE devices to validate these certificates. The authentication protocol 802.1X defined by IEEE could also be used to authenticate the footprints leveraging the capacity to store strong human credentials and managing CPU load. The new modalities and form factors bring out the challenges of coining smaller footprint credential types based on less intensive constructs of cryptographic as authentication layer for cross-cultural communication [17].

B. Authorization

Authorization is the second layer of this framework controls all device access throughout the infrastructure environment [9]. The core authentication layer is also embedded in this by integrating the entities identity information. The exchange of appropriate information starts as soon as a trust relationship is established between authorization and authentication components [11]. The same car vendor can

develop a trust alliance between his all cars, so one car can share certain safety capabilities related information with another car. This established trusted alliance relationship between cars and their dealers may allow transmitting and exchanging additional information such as their last maintenance records or odometer reading [17]. The mechanism of user's access and management to enterprise networks is well validated in the current policy structure of IoE devices. Building an architecture handling communication of billions of IoE devices with varying trust relationships would be a big challenge for cross-culture communications [18]. These challenges would also extend to the point of end-to-end communication with appropriate controls and traffic policies to segment and synchronize the data traffic. The major factor to be looked after in this architecture would be the minimization of data exploitation.

C. Network Enforced Policy

The network enforced policy layer involves the traffic of all things that will route and transport on the infrastructure securely including controlling and management of the data exchange over IoE devices. Various mechanisms and protocols are already established regarding network enforced policy to secure the infrastructure of a network when IoE devices communicate cross-culturally [20].

D. Secure Analytics: Visibility and Control

The process of controlling the IoE ecosystem with the purpose of gaining visibility, a service is defined by the secure analytics layer through which data centers, network infrastructures, and all endpoints participate in providing telemetry [15]. A massive parallel database (MPD) platform can be deployed as it would process large volumes of data efficiently [20]. The anomalies of the secured data can be picked out and real time statistical analysis could be performed when integrating analytics with this technology [4]. This is a telemetry provision of all those elements that correlate and aggregate the information required for threat detection. This model envisages that, if the data is accessed by unauthenticated and unauthorized IoE devices, threat mitigation should automatically shut down the attacker and raise those triggers. The IoE devices generate data and that is only valuable if the correct security process and analytical algorithms are applied to identify and resolve the threats [6]. The security algorithms are applied on various layers of this model and data collected from those sources could produce a better analytical outcome of dealing with security threats. Every day new technology is evolving and network fabrics are becoming more complex in nature. The infrastructures topologies are moving to private and public clouds and this move require defense capabilities along with threat intelligence detection and resolution at the same time on clouds. The derivation of accurate intelligence requires control, context, and visibility [13].

VIII. Conclusion

The IoE constructs have vast security implications so deconstructing an existing security framework could be a foundation of security for future cross-culture communications environments. The proposed framework by the authors could be used in operational environments where policy enforcement is a key feature as well as protocol lead product development frameworks. There is a huge potential for zero-day attacks since the IoE industry is consistently emerging from multi-culture communications to cross-culture communications. This offers the devised architecture to apply security at the appropriate layer. The last layer of this architecture is the end point highly constrained devices and this integration minimized the malware growth on this stage. There is a tremendous increase on IP-based sensors and this leads to attack the data. These evolvements in technology highlight the need for new identification techniques and coining new security protocols. The revised structure should be applied to endpoint IoE devices within the cross culture communication in accordance with their enhanced capabilities. It is clear that IoE always leverages new challenges to security architects and networks. There is a need to evolve smart security systems which include predictive analysis, anomaly detection and threat detection for cross-culture communications.

References

[1] Bekkering, Ernst and J.P. Shim. (2016) "i2i Trust in Videoconferencing." Communications of the ACM 49.7 103-107. 8.

[2] Boh, W. F., & Yellin, D. (2006). Using Enterprise Architecture Standards in Managing Information Technology. Journal of Management Information Systems, 13(3), 163-207.

[3] Botha, A., Vosloo, S., Kuner, J., & Berg, M. v. (2009). Improving Cross-Cultural Awareness and Communication through Mobile Technologies. International Journal of Mobile and Blended Learning, 39 - 53.

[4] Capgemini. (2016). Securing the Internet of Everything opportunity: putting cybersecurity at the heart of the IoE. Retrieved from capgemini.com: https://www.uk.capgemini.com/resources/securing-the-internet-of-things-opportunity-putting-cybersecurity-at-the-heart-of-the-IoE

[5] D. Miorandi, S. Sicari, P. De, and I. Chlamtac. (2012) Internet of things: Vision, applications and research challenges. Ad Hoc Networks, 10(7), 1497-1516.

[6] Fruchter, R, Chen, M, Ando, C. (2003) Geographically Distributed Teamwork Mediated by Virtual Auditorium, Proc. of SID2003 2nd Social Intelligence Design Symposium, ed. D. Rosenberg, T. Nishida, R. Fruchter, London, UK.

[7] Gubbi, J., Buyya, R., Marusic, S., & Palaniswami, M. (2013). Internet of Everything (IoE): A vision, architectural elements, and future directions. Future Generation Computer Systems, 1645 - 1660.

[8] H. Ning and S. Hu. (2011) Technology classification, industry, and education for future Internet of Things. International Journal of Communication Systems, 25(9), 1230-1241.

[9] Järvenpää, S.L., Tractinsky, N. (1999): Consumer Trust in an Internet Store: A Cross-Cultural Validation. Journal of Computer-Mediated Communication, Vol.5 (2), December 1999, available online at http://www.ascusc.org/jcmc/.

[10] K, Karimi. (2014) What the Internet of Things (IoT) needs to become a reality. Available http://www.freescale.com/files/32bit/doc/white_paper/INTOTHNGSWP.pdf, [Accessed 24 January, 2016]

[11] L. Tan and N. Wang. (2010) Future Internet: The Internet of Things. Advanced Computer Theory and Engineering (ICACTE), 5(1), 376-380.

[12] Lam, W. (2005) Investigating success factors in enterprise application integration: A case-driven analysis. European Journal of Information systems, 14(2), 175-187.

[13] Martin, N. L., Pearson, M., & Furumo, K. (2007). IS Project Management: Size, Practices and The Project Management Office1,2. The Journal of computer Information Systems, 47(4), 52-60.

[14] Metastorm (2008). Metastorm releases enhanced ProVision enterprise modeling suite. http://www.metastorm.com/news/2008/040208.asp

[15] R. Weber. (2010) Internet of Things – New security and privacy challenges. Computer Law & Security Review, 26(1), 23–30.

[16] Ross, J. W. (2003). Creating a strategic IT architecture competency: learning in stages. MIS Quarterly Executive, 2(1), 31-43.

[17] Sherwood, J. (2005). Enterprise security architecture: a business-driven approach. San Francisco: CMP Books.

[18] Samovar, Larry A., Richard E. Porter, and Edwin R. McDaniel.(2005) Intercultural Communication: A Reader. Thomson Wadsworth.

[19] S. Gaglio and R. Lo(2012). Advances onto the Internet of Things: How ontologies make the Internet of Things meaningful. Cham: Springer.

[20] X. Su, J. Riekki, J. Nurminen, J. Nieminen, and M. Koskimies. (2014) Adding semantics to Internet of Things.

The Group Decision Support System to Evaluate the ICT Project Performance Using the Hybrid Method of AHP, TOPSIS and Copeland Score

Herri Setiawan
Department of Computer Science and Electronics, Faculty of Mathematics and Natural Sciences
Gadjah Mada University, Yogyakarta, Indonesia

Retantyo Wardoyo
Department of Computer Science and Electronics Faculty of Mathematics and Natural Sciences
Gadjah Mada University, Yogyakarta, Indonesia

Jazi Eko Istiyanto
Department of Computer Science and Electronics Faculty of Mathematics and Natural Sciences
Gadjah Mada University, Yogyakarta, Indonesia

Purwo Santoso
Departement Politics and Government Faculty of Social and Political Sciences
Gadjah Mada University, Yogyakarta, Indonesia

***Abstract*—This paper proposed a concept of the Group Decision Support System (GDSS) to evaluate the performance of Information and Communications Technology (ICT) Projects in Indonesian regional government agencies to overcome any possible inconsistencies which may occur in a decision-making process. By considering the aspect of the applicable legislation, decision makers involved to provide an assessment and evaluation of the ICT project implementation in regional government agencies consisted of Executing Parties of Government Institutions, ICT Management Work Units, Business Process Owner Units, and Society, represented by DPRD (Regional People's Representative Assembly). The contributions of those decision makers in the said model were in the form of preferences to evaluate the ICT project-related alternatives based on the predetermined criteria for the method of Multiple Criteria Decision Making (MCDM). This research presented a GDSS framework integrating the Methods of Analytic Hierarchy Process (AHP), Technique for Order Preference by Similarity to Ideal Solution (TOPSIS) and Copeland Score. The AHP method was used to generate values for the criteria used as input in the calculation process of the TOPSIS method. Results of the TOPSIS calculation generated a project rank indicated by each decision maker, and to combine the different preferences of these decision makers, the Copeland Score method was used as one of the voting methods to determine the best project rank of all the ranks indicated by the desicion makers.**

Keyword*—*GDSS; ICT; MCDM; AHP; TOPSIS; Copeland Score; Decision Maker

I. INTRODUCTION

The main advantage which this Multiple Criteria Decision Making (MCDM) offers is its ability to provide decision-making processes through the analysis of complex problems, aggregation of the criteria used in evaluation processes, the possibility of making the right decision, and the scope for decision makers to participate actively in the decision-making processes[1].

Several research in ICT project performance evaluation-related decision making employed this MCDM method [1][2][3][4]. Selection of effective and efficient projectsis crucial for every organizationas the decision-making processes to assess the feasibility of a certain project are extremely complex. The research was conducted by employing the methods of AHP and Moora as the research approaches [1].

To cope with uncertainties and obscurity found in humans' subjective perceptionsin decision making processes, a *Fuzzy Multi-criteria Decision-Making* (FMCDM) based evaluation method was applied to measure the performance of the software development projects [2]. What constitutes a problem in the MCDM is that it is the decision maker (DM) who have to choose which one is the best alternative that meets the criteria. Generally, it is not easy to an alternative that meets all the criteria simultaneously and thus a compromise solution was preferred. The problem's complexity may increase ifa number of DMs do not have the same perception relating to the existing alternatives. The VIKOR-based ranking method was proposed to identify such a compromise solution. This method used the suitable value for the alternative assessment with unquantifiable criteria, especially if the evaluation was undertaken based on the aspect of linguistics.

Kazemi et al [3] offers a project supervision method in order that such projects are consistent with the strategic objectives. The initial step to diminish the risk of project failure is to choose an optimum project with the MCDM approach using AHP and TOPSIS methods. In another model, *Linear Programming* (LP) and MCDM for decision making were applied in the priority project selection evaluation based on a number of predetermined criteria [4]. The analysis results indicated that MCDM can be used for evaluating project performance.

In Indonesian government agencies, especially in regional governments, there is a type of report called LAKIP, which is the Performance Accountability Report of Government Agencies, which serves as an instrument for measuring

performance of related agencies with regard to the extent of the successful implementation of their programs/activities. Unfortunately, this type of measurement is undertaken on a general basis with a variety of variables used, not specific to ICT. In another research, Ishak [5] examines the effectiveness of performance assessments in each SKPD (the Local Apparatus Work Unit). By using the analysis method based on a variety of data sources, it was concluded that the accountability of Indonesian governments remained focusing merely on financial management, while in the daily reality such financial information failed to answer public curiosity about government accountability and thus an appropriate measuring tool to measure performance of SKPD is necessary. Consequently, e-Government projects need to be evaluated to determine causes of the resulting changes, deficiencies, and irregularities [6].

This paper described a GDSS for ICT project performance evaluation in regional government agencies. This GDSS was used as a tool for decision makers to expand their capabilities, but not as a substitute for their judgment. Broadly speaking, this paper consists of several sections. The first one presents a brief overview of AHP, TOPSIS and Copeland Score. Then, the methodology, i.e. the measures to apply the hybrid method is described by also providing examples on the ways itwas implemented. In the final section, findings of the research that had been conducted are concluded.

Unlike the previous research, in addition to GDSS implementation using the hybrid method, the assessment criteria used were the ones that can be used for the assessment in any categories of ICT projects, not just limited to software and hardware related ICT projects. Moreover, to determine the assessment criteriato be used it is necessary to take into account the technical and managerial aspects in order to accommodate all the DMs.

II. The Overview of Multi-Criteria Decision Making

Based on the number of criteria used, decision related issues can be divided into two categories, namely single-criterion decisions and multi-criteria decisions. The *Multi Criteria Decision Making* (MCDM) is defined as a decision-making method to determine the best alternative of various alternatives based on certain criteria [7]. This MCDM is divided into *Multi Objective Decision Making* (MODM) dan *Multi Attribute Decision Making* (MADM) [8].

There are several methods to use to solve MADM related problemssuch as: 1) the Simple Additive Weighting (SAW) method; 2) Weighted Product (WP); 3) ELimination Et Coix Traduisant la realitE (ELECTRE); 4) Technique for Order Preference by Similarity to Ideal Solution (TOPSIS); and 5) Analytic Hierarchy Process (AHP).

A. The Analytical Hierarchy Process (AHP)

AHP is a decision support model developed by Thomas L. Saaty. This decision support model will describe complicated multi-factor or multi-criteria problems in a hierarchy [9][10]. A hierarchy is defined as a representation of a complex problem in a multi-level structure where the first level is a goal, followed by the levels of factors, criteria, sub-criteria, and so on downwards with alternatives as the lowest level. A hierarchy helps to untangle a complex problem into groups which later are organized into a hierarchical form so that the problem itself will appear more structured and systematic. AHP has its own advantages as it has the ability to perform analyses in a simultaneous and integrated manner of the criteria, both qualitative and quantitative ones. Basically, the steps in the AHP method consist of:

a) Defining the hierarchical structure of a problem

The problem is decomposed into a hierarchical tree illustrating the relationship between the problem, the criteria and the alternative solutions.

b) Undertaking a weighting process of the criteria at each level of the hierarchy

At this stage, all the criteria in each level of the hierarchy are measured in terms of their relative importance compared with the other criteria. It can be done using Saaty's weighting standards with a scale ranging from 1 to 9 and its opposite. The scale used can be changed using other values in accordance with the condition of cases to resolve. Information about the scale used by Saaty can be seen in Table 1.

TABLE I. Saaty Rating Scale-Based Assessment of the Relative Importance of the Criteria

Scale a_{ij}	Description
1	Both criteria are equally important
3	Criterion i is slightly more important than Criterion j
5	Criterion i is more important than Criterion j
7	Criterion i is strongly more important than Criterion j
9	Criterion i is absolutely more important than Criterion j
2, 4, 6, 8	The median of Criteriaiand j is between two adjacent decision values
opposite ($a_{i,j} = 1/a_{i,j}$)	Criterion I has a higher importance value than Criterion j, thus Criterion j has an opposite value

Based on the values of those criteria, the pairwise comparison matrix A can be formulated as follows:

$$A = \begin{bmatrix} a_{1,1} & a_{1,2} & a_{1,3} & \cdots & a_{1,j} \\ a_{2,1} & a_{2,2} & a_{2,3} & \cdots & a_{2,j} \\ a_{3,1} & a_{3,2} & a_{3,3} & \cdots & a_{3,j} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{i,1} & a_{i,2} & a_{i,3} & \cdots & a_{i,j} \end{bmatrix} \quad (1)$$

$A_{i,j}$ refers tothe element of Matrix A in the i^{th} row and the j^{th} column .

c) Calculatingthe criteria weightingand weighting consistency

At this stage, the weighting priority is calculated by looking for the eigenvector value of matrix A through the following processes:

- SquareMatrix A. The value of the element of Matrix A^2 is determined using the following formula:

$$a_{i,j}{}^2 = \sum_{k=1}^{n} a_{i,k} \cdot a_{k,j} \quad (2)$$

ai,k refers tothe element of Matrix A in the i^{th} row and the k^{th} columnand ak,j refers tothe element of Matrix A in the k^{th} row and the j^{th} column.

– Add up the whole elements of each rom inMatrixA^2 until Matrix B is generated using the following formula:

$$b_i = \sum_{j=1}^{n} a_{i,j} = a_{i,i} + a_{i,2} + a_{i,3} + \cdots + a_{i,j} \quad (3)$$

$b_i$refers tothe element of Matrix B in the i^{th} row. Matrix B is arranged by Element b_i in the following pattern:

$$B = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ \cdots \\ \cdots \\ \cdots \\ b_i \end{bmatrix} \quad (4)$$

Add up the whole elements of Matrix B using the following formula:

$$\sum_{i=1}^{n} b_i = b_1 + b_2 + b_3 + \cdots + b_i \quad (5)$$

– After Matrix B is obtained in Stepbabove, normalization is undertaken to Matrix B to obtain its eigenvector value. This eigenvector value of Matrix B is described in the form of Matrix E as follows:

$$E = \begin{bmatrix} e_1 & = & b_1 & / & \sum_{i=1}^{n} b_i \\ e_2 & = & b_2 & / & \sum_{i=1}^{n} b_i \\ & & \vdots & & \\ e_i & = & b_i & / & \sum_{i=1}^{n} b_i \end{bmatrix} \quad (6)$$

e_i refers tothe element of Matrix E in the i^{th} row.

– Those three processes above are performed repeatedly and at the end of each iteration, the differential of the eigenvector values of Matrix E is calculated using the previous eigenvector values of Matrix E until an amount whose value is close to zero is generated. Matrix E obtained in the last step indicates the criteria priority indicated by the eigenvector value coefficient.

d) Calculating the alternative weighting

In this stage, alternative weighting is performed for each criterion in the pairwise comparison matrix. The process to undertake such alternative weighting is similar to that performed to calculate criteria weighting.

e) Showing the order of alternatives under consideration and selecting the alternatives

In this stage, the eigenvector values obtained in the alternative weighting for each criterionand the eigenvector values generated from the criteria weighting are calculated. This is done to determine the alternative chosen from all the available alternatives.

f) Repeating Steps c, d and e for the whole levels of the hierarchy

g) Calculating the eigenvector value of each pairwise comparison matrix

Eigenvector values are the score for each element. This step aims to synthesize the element priority from the lowest hierarchy level to the goal attainment.

h) Examining the consistency of the hierarchy. If the value is greater than 10 percent, it means that the judgment data assessment should be revised

B. Technique for Order Preference by Similarity to Ideal Solution (TOPSIS)

The Technique for Order Preference by Similarity to Ideal Solution (TOPSIS) is developed based on the concept that the best selected alternative should not only has the shortest distance from the positive ideal solution, but it also has the longest distance from the negative ideal solution [11]. Generally, TOPSIS procedures are given in the following steps:

a) Calculating normalization values

TOPSIS requires performance rating of eachalternative of Ai in each normalized criterion of Cj, namely:

$$r_{ij} = \frac{x_{ij}}{\sqrt{\sum_{i=1}^{m} x_{ij}^2}} \quad (7)$$

Description of the symbols:

rij refers to the normalization value of each alternative (i) compared with criterion (j) where i=1,2,...,m; and j = 1,2,...,n.

xij refers to a value of an alternative (i) compared with criterion(j) where i=1,2,...,m; and j = 1,2,...,n.

b) Calculating weighted normalization values

After calculating the normalization values, the next step is to calculate weighted normalization values by multiplying the value of each alternative in the normalization matrix by thescore given by decision makers. The following equation used is:

$$y_{ij} = w_i r_{ij} \quad (8)$$

Description of the symbols:

- yijrefers toweighted normalization values.
- wi refers to the score for each criterion.
- rij refers tonormalization valuesof each alternative where i=1,2,...,m; and j = 1,2,...,n.

Identifying positive ideal solutions and negative ideal solutions

Positive ideal solutions and negative ideal solutionscan be calculated based on theweighted normalization values as:

$$A^+ = \left(y_1^+, y_2^+, \cdots, y_n^+\right); \quad (9)$$

where

$$A^- = \left(y_1^-, y_2^-, \cdots, y_n^-\right); \quad (10)$$

$$y_j^+ = \begin{cases} \max\limits_i y_{ij}; & \text{If j is the benefit attribute} \\ \min\limits_i y_{ij}; & \text{if j is the cost attribute} \end{cases} \quad (11)$$

j = 1,2,..,n.

$$y_j^- = \begin{cases} \min\limits_i y_{ij}; & \text{if j is the benefit attribute} \\ \max\limits_i y_{ij}; & \text{If j is the cost attribute} \end{cases} \quad (12)$$

Description of the symbols:

- The positive ideal solution (A+) is obtained by searching the maximum value of the weighted normalization value (yij) if the attributeis the benefit attribute and the minimum value of the weighted normalization value (yij) if the attribute is the cost attribute.
- The negative ideal solution (A-) is obtained by searching the minimum value of the weighted normalization value (yij) if the attribute is the benefit attribute and themaximum value of the weighted normalization value (yij) if the attribute is the cost attribute.

c) Calculating the distance between each alternativeand either the positive ideal solution or the negative ideal solution

The distance between Alternative Ai andthe positive ideal solution is formulated as:

$$D_i^+ = \sqrt{\sum_{j=1}^{n} \left(y_i^+ - y_{ij}\right)^2}; \qquad \text{i=1,2,…,m} \qquad (13)$$

The distance between Alternative Ai andthe negative ideal solution is formulated as:

$$D_i^- = \sqrt{\sum_{j=1}^{n} \left(y_{ij} - y_i^-\right)^2}; \ \text{i=1,2,…,m} \qquad (14)$$

Description of the symbols:

- The distance between Alternative Ai and the positive ideal solution (yj+) represented by the symbol Di+ is derived from the square root of the total values of each alternative obtained and the weighted normalization value for each alternative (yij) minus the positive ideal solution (yi+) and then squared.
- The distance between Alternative Ai and the negativeideal solution (yj-) represented by the symbol Di- is derived from the square root of the total values of each alternative obtained and the weighted normalization value for each alternative (yij) minus the negative ideal solution (yi-) and then squared.

d) Determiningthe proximity value of each alternative towards the ideal solutions (preference)

The preference valuefor each alternative (Vi) is given as follows:

$$V_i = \frac{D_i^-}{D_i^- + D_i^+}; \qquad (15)$$

Description of the symbols:

- Vi (the preference value for each alternative) is obtained from the value of the distance between Alternative Ai and the negative ideal solution (Di-) divided by the total value of the distance between Alternative Ai and the negative ideal solution (Di-) plus the sum of the value of the distance between Alternative Ai and the negative ideal solution (Di+).
- The value of Vi which is greater indicates that Alternative Ai is preferred.

C. The Copeland Score

One of the common problems in the GDSS is the way to aggregate decision makers'opinions in order to make an appropriate decision. The methods of group decision-making (especially those related to MCDM) will usually experience problems if each decision maker gives their individual preference[12]. In general, the GDSS consists of two stages in, namely stimulating decision makers' preferences separately and performing group aggregation towards any preferences given.

Among the tools used in the aggregation of group-based decision making is voting. Voting is defined as an act to select the most frequently appearing value among the selected alternatives [13].

Copeland score is one of the voting methods whose technique is based on a subtraction of the frequency of winning minus the defeat frequency of a pairwise comparison[13]. Another research [14] describes the way the voting method of Copeland Score accommodates decision makers'score based on their respective level of expertise. The example of how to determine pairwise comparisons in the Copeland Score method is presented in Figure 1.

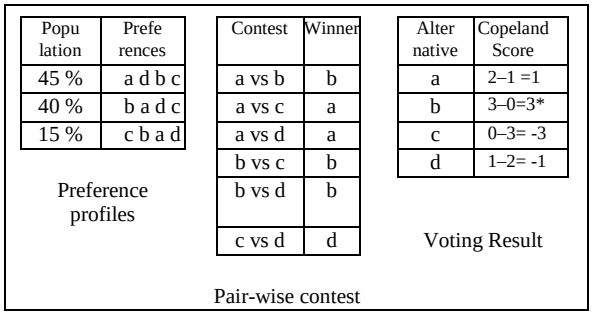


Fig. 1. Determination of pairwise comparisons in the Copeland Score method

Figure 1 above presents three tables, namely the Table of Preference Profiles, the Pairwise Contest Table, and the Table of Voting Results. The Table of Preference Profiles suggests that there are four options, namely A, B, C, and D. 45% of the population prefers A to D, B, and C (see the first row of the Table of Preference Profiles). The Pairwise Contest Table indicates that an option (for exampleA) is compared with all the available options (B, C, D). This pairwise comparison is completedone by one and applies for the overall items of choices.

In the first rowrelating to the pairwise comparison between A and B (see the Table of Preference Profiles), 45% of the population prefers A to B; in the second row, 40% of the population prefers B to A; in the third row, 15% of the population prefers B toA. This implies that there is 45% of the population that prefers A to B, while the remaining 55% (the total number of the population that prefers b) of the population prefers B to A. Thus, B is chosen as the winner of a pairwise comparison between A to B. Comparisons are also made to other candidatesas described above.

The Pairwise Contest Table shows that as the winner, Option Aappears 2x (twice). Option B appears 3x (three times). Option D appears 1x (once), and Option C does not appear.

According to the Pairwise Contest Table, it is revealed that Option A has two chances of winning over C and D, and a chance of losing to B. To determine whether Option A is the best choice or not, a subtraction of the frequency of winning minus the defeat frequency is performed. The Table of the Voting Results shows that Option B has the highest frequency. Based on the frequency, it is decided that Alternative B wins.

III. Research Method

A. Classification of the Types of ICT Projects

The projects in the regional government agencies belonging to ICT projects are [15]:

a) Software Establishment/Development,

b) Hardware Provision/Maintenance,

c) Network Building/Maintenance,

d) Bandwith Purchase/Lease, and

e) Educational Programs/Training for ICT staff

B. The Implemented GDSS Method

Evaluation of ICT projects designedwas a model of *Multiple-Criteria Decision Making* (MCDM) using the methods in *Multi-Attribute Decision Making* (MADM). Determination of the best alternative among a number of alternatives was done based on several predetermined criteria. The scoring criteria to evaluate ICT projects were a compilation of project management concepts in general [16], ISO/IEC 15939 concerning how to measure software [17] and benchmarks that can be used to measure computer performance [18].

Table 2 describes decision makers along with the parameters and criteria used in the evaluation of ICT projects.

TABLE II. Criteria of ICT Project Evaluation

Output Parameter		
Criteria		Decision Maker
1	Project Schedule (C1)	Business Process Owner Units (DM1)
2	Project Costs (C2)	
3	Project Scope (C3)	
4	Project Risks (C4)	ICT Management Work Units (DM2)
5	Project Performance (C5)	
Outcome Parameter		
Criteria		Decision Maker
6	Project Effectiveness (C6)	Executing Parties of Government Institutions (DM3) and Society represented by DPRD (DM4)
7	Project User Satisfaction (C7)	

Evaluation of ICT projects in government agencies requires assessments undertaken by the Executing Parties of Government Institutions, ICT Management Work Units, Business Process Owner Units, and Society. Stakeholders of the ICT management as a group of decision makers have specified the assessment criteria based on performance indicators according to the duties and functions. Such performance assessments employed both qualitative and quantitative criteria, where the qualitative criteria used linguistic variables. These linguistic variables referred to variables whose values are indicated in the forms of words or sentences in natural or artificial language [19].

Then, to draw the conclusion relating to the ICT project results attained, the performance assessment scale based on the existing criteria was used. The measurement scale was developed based on the consideration of each decision maker. Table 3 presents the assessment scoring scale used in this research.

TABLE III. Assessment Scoring for Criteria Performance

Score	Assessment			Percentage
5	Very Good	Very Large	Ignored	90 s/d 100
4	Fairly Good	Fairly Large	Minor	80 s/d 89,99
3	Good	Large	Moderate	60 s/d 79,99
2	Less Good	Less Large	Serious	40 s/d 59,99
1	Not Good	Not Large	Critical	< 39,99

C. Scoring for each criteria is elucidated as follows:

- Project Schedule

Based on the criteria of the project schedule timeliness, the percentage between the planned project scheduleand the actual project schedule [20].

Formula:

[1 – ABS (ALT – PLT) / PLT] x 100% (16)

ALT=Actual Finish Date – Actual Start Date

PLT= Planned Finish Date-Planned Start Date

- Project Costs

The ability to provide the agreed scope of duties concerning costs, hours of work, laboratories and travel expenses. Based on the percentage between the committed (baseline) and expected costs (actual + forecast) [20].

Formula:

[1 – (ECost – CCost) / CCost] x 100% (17)

Expecteed Cost (Ecost) = actual + forecast

Commited Cost (Ccost)

- Project Scope

In this criteria category, the scoring used several linguistic variables, namely: Very Large, Fairly Large, Large, Less Large, and Not Large.

- Project Risks

These refer to the arising impacts of the risks, which are defined as follows [20]:

– *Critical*: If this risk occurs,a project will fail. The minimum requirements of the project cannot be met.

- *Serious*: If this risk occurs, a project will encounter increases in terms of the costs/ schedule. The minimum requirements of the project that are acceptable can be met while the secondary requirements cannot.
- *Moderate*: If this risk occurs, a project will encounter increases in terms of the costs/ schedule. The minimum requirements of the project that are acceptable and a few of the secondary requirements can be met.
- *Minor*: If this risk occurs, a project will encounter slight increases in terms of the costs/ schedule. The minimum requirements of the project that are acceptable and some of the secondary requirements can be met.
- *Ignored*: If this risk occurs, it will not affect a project. All the requirements can be met.

- Project Performance, Project Effectiveness and Project User Satisfaction

In this criteria category, the scoring used several linguistic variables, namely: Very Good, Fairly Good, Good, Less Good, and Not Good.

Each has its own performance assessment criteria indicated in a measurement scale.

D. The Hybrid Methodof AHP, TOPSIS and Copeland Score

a. Performing criteria scoring(AHP)
b. Calculating normalization values (TOPSIS)
c. Calculating weighted normalization values (AHP-TOPSIS)

$$y_{ij} = w_i r_{ij} \quad (18)$$

Description of the symbols:
- y_{ij} refers to weighted normalization values.
- w_i refersto the score of each criteria (generated from AHP scoring)
- r_{ij} refers to the normalization value of each alternativewherei=1,2,...,m; and j = 1,2,...,n

d. Identifying positive and negative ideal solutions (TOPSIS)
e. Calculating the distance between each alternative and the positiveand negative ideal solutions (TOPSIS)
f. Determining the proximity value of each alternative towards the ideal solution (preference) (TOPSIS)
g. Repeating **steps a** to **f** for each Decision Maker
h. Ranking all the DMs (TOPSIS-Copeland Score)

IV. Result and Analysis

This section provides examples of the ICT project evaluation model implementation. The sample data used were retrieved from ten regional government ICT projects that have been completed. In this GDSS model, there were four decision makers (namely DM1, DM2, DM3, and DM4), seven criteria (namely C1, C2, C3, C4, C5, C6, and C7) to assess and ten ICT project alternatives (namely P1, P2, P3, P4, P5, P6, P7, P8, P9, and P10) to evaluate.

DM1 evaluated each alternative based on three criteria C = {C1, C2, C3), DM2 evaluated each alternative based on two criteria C = {C4, C5}, and lastly DM3 and DM4 evaluated each alternative based on two criteria C = {C6, C7}.

a) Performing criteria scoring(AHP)

The first step was to create a pairwise comparison matrix of criteria for DM1, followed by scoring the criteria. Then,the total value of a_{ij} for each pairwise comparison matrix column was calculated as shown in Table 4.

TABLE IV. The Pairwise Matrix for the Criteria of DM1

	C1	C2	C3
C1	1	0.5	0.3
C2	2	1	0.5
C3	3	2	1
	6	3.5	1.8

After normalization had been completed,the results are presented inTable 5.

TABLE V. Scores for Normalized Criteria

	C1	C2	C3	Rata-rata	
C1	0.1667	0.1429	0.1818	0.1638	W1
C2	0.3333	0.2857	0.2727	0.2973	W2
C3	0.5000	0.5714	0.5455	0.5390	W3
	1.0000	1.0000	1.0000	1.0000	

b) Calculating normalization values (TOPSIS)

Based on the dataon the evaluation results given by DM1 on the criteria for each ICT project alternative, the following data on assessment results presented in Table 6 are obtained.

TABLE VI. Scoring for DM1

	C1	C2	C3
P1	4	4	5
P2	3	3	4
P3	5	4	2
P4	4	4	5
P5	3	3	4
P6	5	4	2
P7	4	4	5
P8	3	3	4
P9	5	4	2
P10	4	4	5
	12.8841	11.7898	12.6491

TABLE VII. Normalized Scoring for DM1 (Matrix R)

R	0.3105	0.3393	0.3953
	0.2328	0.2545	0.3162
	0.3881	0.3393	0.1581
	0.3105	0.3393	0.3953
	0.2328	0.2545	0.3162
	0.3881	0.3393	0.1581
	0.3105	0.3393	0.3953
	0.2328	0.2545	0.3162
	0.3881	0.3393	0.1581
	0.3881	0.3393	0.1581

c) Calculating normalization values (AHP-TOPSIS)

The scoring of normalized values for DM1/ Matrix Y presented in Table 8 was obtained from the multiplication of the normalized value of each criterion in Table 7 by the normalized scoring for DM1/ Matrix R in Table 6.

TABLE VIII. MATRIX (Y) OF DM1

Y	0.0508	0.1009	0.2130
	0.0381	0.0756	0.1704
	0.0636	0.1009	0.0852
	0.0508	0.1009	0.2130
	0.0381	0.0756	0.1704
	0.0636	0.1009	0.0852
	0.0508	0.1009	0.2130
	0.0381	0.0756	0.1704
	0.0636	0.1009	0.0852
	0.0636	0.1009	0.0852

d) Identifying positive and negative ideal solutions (TOPSIS)

A+	0.0636	0.1009	0.2130
A-	0.0381	0.0756	0.0852

e) Calculating the distance between each alternative and the positive and negative ideal solutions (TOPSIS)

D+1	0.012712	D-1	0.130907
D+2	0.110520	D-2	0.085217
D+3	0.043126	D-3	0.035806
D+4	0.110088	D-4	0.120157
D+5	0.154066	D-5	0.132746
D+6	0.035516	D-6	0.000000
D+7	0.110088	D-7	0.120157
D+8	0.103272	D-8	0.116746
D+9	0.101414	D-9	0.134992
D+10	0.092180	D-10	0.114053

f) Determining the proximity value of each alternative towards the ideal solution (preference) (TOPSIS)

P1	0.911489
P2	0.435366
P3	0.453630
P4	0.521865
P5	0.462834
P6	0.000000
P7	0.521865
P8	0.530622
P9	0.571017
P10	0.553032

P1	0.911489	Winner
P9	0.571017	
P10	0.553032	
P8	0.530622	
P4	0.521865	
P7	0.521865	
P5	0.462834	
P3	0.453630	
P2	0.435366	
P6	0.000000	

*g) Repeating **steps a** to **f** for each Decision Maker*

After the scoring process had been completed by each DM (DM1, DM2, DM3 and DM4), the following results of project ranking presented in Table 9 were obtained.

TABLE IX. RANKING OF ALL THE DMS

R	DM1		DM2		DM3		DM4	
1	P1	0.911489	P3	0.614761	P1	0.789999	P1	0.777171
2	P9	0.571017	P10	0.614761	P3	0.707060	P3	0.532960
3	P10	0.553032	P4	0.597624	P4	0.707060	P5	0.532960
4	P8	0.530622	P7	0.597624	P10	0.707060	P10	0.532960
5	P4	0.521865	P2	0.573689	P6	0.675952	P6	0.524957
6	P7	0.521865	P6	0.569533	P9	0.675952	P7	0.524957
7	P5	0.462834	P9	0.569533	P2	0.594851	P9	0.524957
8	P3	0.453630	P8	0.553453	P7	0.594851	P2	0.522668
9	P2	0.435366	P5	0.541074	P8	0.585358	P4	0.522668
10	P6	0.000000	P1	0.453388	P5	0.572409	P8	0.522668

h) Ranking the project evaluation results of all the DMs (TOPSIS-Copeland Score)

The following are the stages ofvoting results for the best ICT projects:

- Preference Profile

The preference profile presented in Table 10 shows that there are a total of ten ICT project alternatives (namely P1, P2, P3, P4, P5, P6, P7, P8, P9, and P10). Each decision maker in the process of decision making had their own score which had been determined according to their respective expertise and competence. The score for DM1 was equal to 0.1, the score for DM2 was equal to 0.4, and the scoresfor DM3 and DM4 were equal to 0.2.

- Performing pairwise contests

A pairwise contest is defined as a paired-comparison process comparing one candidate (alternative) to the other candidates, which was performed by:

- Displaying alternative contests in pairs. For example, P1 is compared with P2, P1 is compared with P3, and so on. Similarly, P2 is compared with P3, P2 is compared with P3, and so on. This pairwise comparisons are taken one at a time and done to all the options until P9 is compared with p10.
- Searching for the winner of the comparisons (contests) of each paired alternative.
- For example, in the contest of the pairwise comparison between P1 and P2 (see Table 11), the winner is P1 as in DM1, P1 ranks 1st while P2 ranks 9th so that the winner in DM1 is P1. In DM2, P1 ranks 10th while P2 ranks 6th, and as a resultP2 wins in DM2. In DM3, P1 ranks 2nd while P2 ranks 7th and consequently the winner in DM3 is P1. Lastly, in DM4,P1 ranks 1st while P2 ranks 8th and thus the winner in DM4 is P1. These imply that the rank of P1 is three times higher than the rank of P2, after calculating the scores of each DM it is revealed that the total score for P1 is equal to 0.1 + 0.3 + 0.2 = 0.6 while for P2 is equal to 0.4. Thus, P1 wins.

TABLE X. PREFERENCE PROFILE

Weight	Preferences (Rangking)									
	1	2	3	4	5	6	7	8	9	10
DM1 (0.1)	P1	P9	P10	P8	P4	P7	P5	P3	P2	P6
DM2 (0.4)	P3	P10	P4	P7	P2	P6	P9	P8	P5	P1
DM3 (0.3)	P1	P3	P4	P10	P6	P9	P2	P7	P8	P5
DM4 (0.2)	P1	P3	P5	P10	P6	P7	P9	P2	P4	P8

TABLE XI. PAIRWISE CONTESTS

Contest			Winner
P1 (01+0.3+0.2)	VS	P2 (0.4)	P1
P1 (01+0.3+0.2)	VS	P3 (0.4)	P1
P1 (01+0.3+0.2)	VS	P4 (0.4)	P1
. . .	. . .	. . .	. . .

P1 (01+0.3+0.2)	VS	P10 (0.4)	P1

- Calculating the voting results

The voting results present the results of voting (score) for each candidate afterpairwise contests, based on the following stages:

- Calculating the frequency of winningof the candidates (alternatives) which had been compared using the pairwise contest.
- Calculating the defeat frequency of the candidates (alternatives) which had been compared using the pairwise contest.
- Calculating the differential betweenthe frequency of winningand the defeat frequency of the candidates (alternatives) which had been compared
- Presenting the frequency differential obtained as a score for each candidate.

The voting results are presented in order according to the ranking of the frequency-of-winning scores from the highest to the lowest one for the four DMs, which can be seen in Table 12.

TABLE XII. VOTING RESULTS

Alternatif	Win	Loss	W-L
Proyek 1	9	0	9
Proyek 3	8	1	7
Proyek 10	7	2	5
Proyek 4	6	3	3
Proyek 7	5	4	1
Proyek 6	4	5	-1
Proyek 9	3	6	-3
Proyek 2	2	7	-5
Proyek 8	1	8	-7
Proyek 5	0	9	-9

V. CONCLUSION

This paper offered a hybrid method in MCDM to evaluate ICT projects in Indonesian regional government agencies based on the concept of Group Decision Support Systems (GDSS).

The GDSS concept can overcome any possible inconsistencies which may occur in decision making as it makes decisions based on the mathematical calculation model. Contributionsof the decision makers in the model were in the fom of preferences for choosing ICT Project alternatives based on predetermined criteria using the hybrid method of AHP, TOPSIS and Copeland Score. Based on the implementation examples, projects with the best rank were produced from the assessment undertaken by all DMs, namely Projects 1, 3 and 10 which had the same performance while Project 5 had the worst performance.

Our next research will focus on the development of a web-based prototype to implement the proposed model. The prototype developed attempts to provide an answer to the problems relating to ICT project performance evaluation in regional government agencies.

ACKNOWLEDGMENT

The first author is an employee of Indo Global Mandiri Foundation (*Yayasan Indo Global Mandiri*, IGM) as a lecturer at Faculty of Computer Science, Indo Global Mandiri University. Now he is pursuing a doctoral program on Computer Science, Department of Computer Science and Electronics, Faculty of Mathematics and Natural Sciences, Gadjah Mada University. This work is supported and founded by IGM.

REFERENCES

[1] T. Bakshi, A. Sinharay, and B. Sarkar, "Exploratory Analysis of Project Selection through MCDM," in ICOQM-10, 2011, pp. 128–133.

[2] G. Büyüközkan and D. Ruan, "Evaluation of software development projects using a fuzzy multi-criteria decision approach," Math. Comput. Simul., vol. 77, no. 5–6, pp. 464–475, May 2008.

[3] S. M. Kazemi, S. M. M. Kazemi, and M. Bahri, "Six Sigma project selections by using a Multi Criteria Decision making approach: a Case study in Poly Acryl Corp.," in Proceedings of the 41st International Conference on Computers & Industrial Engineering, 2011, pp. 502–507.

[4] H. Ismaili, "Multi-Criteria Decision Support for Strategic Program Prioritization at Defence Research and Development Canada," University of Ottawa, 2013.

[5] M. Ishak, "Kebijkaan Pengukuran Kinerja Pemerintah Daerah," INOVASI, vol. 6th, pp. 143–151, 2009.

[6] G. J. Victor, A. Panikar, and V. K. Kanhere, "E-government Projects – Importance of Post Completion Audits," in International Conference of e-government (ICEG), 2007, pp. 189–199.

[7] S. Kusumadewi, S. Hartati, A. Hardjoko, and R. Wardoyo, Fuzzy Multi Attribute Decision Making (Fuzzy MADM), First. Jogyakarta: Graha Ilmu, 2006.

[8] H.-J. Zimmermann, Fuzzy Set Theory and Its Applications, Second Edi. Boston, MA: Kluwer Academic Publishers, 1991.

[9] T. L. Saaty, The Analytic Hierarchy Process: Planning, Priority Setting, Resource Allocation. New York, NY: McGraw-Hill, 1980.

[10] T. L. Saaty, Fundamentals of Decision Making and Priority Theory With the Analytic Hierarchy Process. Pittsburgh: RWS Publications, 2000.

[11] C.-L. Hwang and K. Yoon, Multiple Attribute Decision Making : Methods and applications. New York: Springer Berlin Heidelberg, 1981.

[12] S. K. Cheng, "Development of a Fuzzy Multi-Criteria Decision Support System for Municipal Solid Waste Management .' A Thesis," University of Regina, 2000.

[13] B. Gavish and J. H. Gerdes, "Voting mechanisms and their implications in a GDSS environment," Ann. Oper. Res., vol. 71, pp. 41 – 74, 1997.

[14] Ermatita, S. Hartati, R. Wardoyo, and A. Harjoko, "Development of Copeland Score Methods for Determine Group Decisions," Int. J. Adv. Comput. Sci. Appl., vol. 4, no. 6, pp. 240–242, 2013.

[15] H. Setiawan, J. E. Istiyanto, R. Wardoyo, and P. Santoso, "The Use of KPI In Group Decision Support Model of ICT Projects Performance Evaluation," in International Conference on Electrical Engineering, Computer Science and Informatics (EECSI 2015), 2015, no. August, pp. 19–20.

[16] PMI, A Guide to the Project Management Body of Knowledge - PMBOK Guide. 2013.

[17] ISO/IEC, Information technology — Software engineering — Software measurement process, no. September. 2001.

[18] "Standard Performance Evaluation Corporation." [Online]. Available: https://www.spec.org/. [Accessed: 02-Jan-2016].

[19] L. A. Zadeh, "The Concept of Linguistic Variable and ITs Application to Aprroximate Reasoning-II," in Information Sciences, vol. 357, 1975, pp. 301–357.

[20] P. A. Engert and Z. F. Lansdowne, "Risk Matrix User ' s Guide," Bedford, Massachusetts, 1999.

A Novel Efficient Forecasting of Stock Market Using Particle Swarm Optimization with Center of Mass Based Technique

Razan A. Jamous
Department of Mathematics, Faculty of Science, Ain Shams University
Cairo, Egypt

Essam El.Seidy
Department of Mathematics, Faculty of Science, Ain Shams University
Cairo, Egypt

Bayoumi Ibrahim Bayoum
Department of Mathematics, Faculty of Science, Ain Shams University
Cairo, Egypt

***Abstract*—This paper develops an efficient forecasting model for various stock price indices based on the previously introduced particle swarm optimization with center mass (PSOCOM) technique. The structure used in the proposed prediction models is a simple linear combiner using (PSOCOM) by minimizing its mean square error (MSE) to evaluate the proposed model. The comparison with other models such as standard PSO, Genetic algorithm, Bacterial foraging optimization, and adaptive bacterial foraging optimization had been done. The experimental results show that PSOCOM algorithms are the best among other algorithms in terms of MSE and the accuracy of prediction for some stock price indices. Whereas, the proposed forecasting model gives accurate prediction for short- and long-term prediction. As a result, the proposed stock market prediction model is more efficient from the other compared models.**

Keywords*—Stock market forecasting; particle swarm optimization; Bacterial foraging optimization; Adaptive bacterial foraging optimization; Genetic algorithm

I. INTRODUCTION

Stock market is an organized and regulated financial market where securities such as bonds, notes, and shares are bought and sold at prices governed by the forces of demand and supply [1]. In addition, stock market is, without a doubt, one of the greatest tools ever invented for building wealth. Stocks are a part, if not the cornerstone, of any investment portfolio. This demand coupled with advances in trading technology has opened up the markets so that nowadays nearly anybody can own stocks, and use many types of software to perform the aspired profit with minimum risk [2]. Consequently, a lot of attention had been devoted to the analysis and prediction of future values and trends of the financial stock markets, and due to applications in different business transactions, stock market prediction has become a hot topic of research [3]. There is no doubt that the majority of the people related to stock markets are trying to achieve profit. Profit is achieved by investing in stocks that have a good future (short or long term future).

In this paper, our earlier presented particle swarm optimization with center of mass technique (PSOCoM) is applied to the task of training an adaptive linear combiner to form a new stock market prediction model. This prediction model is used with some common indicators to maximize the return and minimize the risk for the stock market.

The rest of the paper is organized as follows: The survey of the relevant literature is summarized in Section 2. The description of the proposed technique is given in Section 3. Simulation results are shown in Section 4, followed by conclusions in Section 5.

II. RELATED WORK

Many research papers have appeared in the literature using evolutionary computing tools such as genetic algorithm (GA)[4], particle swarm optimization (PSO)[5], and bacterial foraging optimization (BFO)[6], and Adaptive bacterial foraging optimization (ABFO) in developing forecasting models.

A new evolutionary computation technique called Bacterial foraging optimization (BFO) had been proposed in [3]. It is inspired by the pattern exhibited by bacterial foraging behavior. Bacteria have the tendency to gather to the nutrient-rich areas by an activity called chemotaxis. It is known that bacteria swim by rotating whip like flagella driven by a reversible motor embedded in the cell wall. E. coli has 8–10 flagella placed randomly on a cell body. When all flagella rotate counterclockwise, they form a compact, helically propelling the cell along a trajectory, which is called run. When the flagella rotate clockwise, they pull on the bacterium in different directions and cause the bacteria to tumble. The bacterial foraging system primarily consists of four sequential mechanisms namely chemotaxis, swarming, reproduction, and elimination-dispersal [7].

Bacterial Foraging Optimization (BFO) is a recently developed nature-inspired optimization algorithm, which is based on the foraging behavior of *E. coli* bacteria. Up to now, BFO has been applied successfully to some engineering problems due to its simplicity and ease of implementation. However, BFO possesses a poor convergence behavior over complex optimization problems as compared to other nature-inspired optimization techniques. This paper first analyses how the run-length unit parameter of BFO controls the exploration of the whole search space and the exploitation of the promising areas. Then it had been presented a variation on the original BFO, called the adaptive bacterial foraging optimization

(ABFO) [8], employing the adaptive foraging strategies to improve the performance of the original BFO. This improvement is achieved by enabling the bacterial foraging algorithm to adjust the run-length unit parameter dynamically during algorithm execution in order to balance the exploration exploitation tradeoff [9]. Majhi in [10] developed two new forecasting models based on bacterial foraging optimization (BFO) and adaptive bacterial foraging optimization (ABFO) to predict S&P500 and DJIA stock indices using technical indicators derived from the past stock indices. The structure of these models is basically an adaptive liner combiner, the weights of which are trained using the ABFO and BFO algorithms.

Kyoung-jae Kim and Won Boo Lee [11] developed a feature transformation method using genetic algorithms. This approach reduces the dimensionality of the feature space and removes irrelevant factors involved in stock price prediction.

Another research done on genetic algorithms (GAs) by Kyoung-jae Kim [12] to predict stock market by using GA not only to improve the learning algorithm, but also to reduce the complexity of the feature space. Thus, this approach reduces dimensionality of the feature space and enhances the generalizability of the classifier.

The authors in [13][14], proposed data mining approach using genetic algorithms (GA) to solve the knowledge acquisition problems that are inherent in constructing and maintaining rule-based applications for stock market. Although there are an infinite number of possible rules by which it could trade, only a few of them would have made us a profit if it had been following them. The authors intend to find good sets of rules which would have made the most money over a certain historical period.

Kennedy and Eberhart in [15] introduced particle Swarm Optimization (PSO) in 1995. Individuals in a particle swarm follow a very simple behavior: to emulate the success of neighboring individuals and their own successes. The collective behavior that emerges from this simple behavior is that of discovering optimal regions of a high dimension al search space [16]. PSO algorithm maintains a swarm of particles, where each particle represents a potential solution. In analogy with evolutionary computation paradigms, a swarm is similar to a population, while a particle is similar to an individual. In simple terms, the particles are "flown" through a multidimensional search space, where the position of each particle is adjusted according to its own experience and that of its neighbors.

Let $x_{id}^{(t)} = (x_{i1}, x_{i2}, \ldots, x_{id})$ denote the position of particle i in the search space at time step t, $V_{id}^{(t)} = (v_{i1}, v_{i2}, \ldots, v_{id})$ denote the velocity particle i in the search space at time step t, $P_i = (p_{i1}, p_{i2}, \ldots, p_{id})$ denote the best solution achieved so far by the particle itself, $P_{gd} = (p_{g1}, p_{g2}, \ldots, p_{gd})$,denote the best solution achieved so far by the whole swarm. The new position of the particle is changed by adding a velocity to the current position, as follows:

$$x_{id}^{(t+1)} = x_{id}^{(t)} + v_{id}^{(t+1)} \quad (1)$$

$$v_{id}^{(t+1)} = w.V_{id}^{(t)} + c_1 r_1\left(P_{id} - X_{id}^{(t)}\right) + c_2 r_2\left(P_{gd} - X_{id}^{(t)}\right) \quad (2)$$

Where c_1 and c_2 are two positive constants, r_1 and r_2 are two random numbers in the range [0, 1]; w is the inertia weight. The velocity vector drives the optimization process, and reflects both the experiential knowledge of the particle and socially exchanged information from the particle's neighborhood. The experiential knowledge of a particle is generally referred to as the cognitive component, which is proportional to the distance of the particle from its own best position (referred to as pbesti). The socially exchanged information is referred to as the social component of the velocity equation (2), which is proportional to the distance of the particle from the best position found by the swarm (referred to as gbest). For the global best PSO, or gbest PSO, the neighborhood for each particle is the entire swarm. The social component of the particle velocity update reflects information obtained from all the particles in the swarm. In this case, the social information is the best position found by the swarm. For the local best PSO, or pbesti PSO, the neighborhood for each particle is small number of particles in the swarm. Thus, the social component reflects information exchanged within the neighborhood of the particle, reflecting local knowledge of the environment. In this case, the social information is the best position found by the experiential knowledge of the particle. The velocity calculation as given in equation (2) consists of three terms: the previous velocity, $V_{id}^{(t)}$, the cognitive component, $c_1 r_1\left(P_{id} - X_{id}^{(t)}\right)$, and the social component, $c_2 r_2\left(P_{gd} - X_{id}^{(t)}\right)$.

PSO has become popular choice for solving complex and intricate problems which are otherwise difficult to solve by traditional methods [17]. The usage of the PSO technique in coping with stock market prediction problems is the most important applications of PSO to predict the stocks that have maximum profit with minimum risk. In our earlier paper[18], we introduce many different forms of PSO which used for stock market prediction such as Standard Particle Swarm Optimization, In our another earlier paper [19], we present a new PSOCoM Optimization algorithm. Also, in our [20], we apply the presented PSOCoM technique to the task of training an adaptive linear combiner to form a new stock market prediction model. This prediction model is used with some common indicators such as S&P500, DJIA and NASDAQ-100 that give advice of buy and sell to increase the profit and decries the risk in stock market.

III. The Proposed Stock Market Prediction Technique

In this section, the description of the proposed prediction technique is provided. The new efficient search technique, that is, PSOCoM Optimization algorithm, is used to design the proposed efficient forecasting of stock market. PSOCoM benefits from the physical principle "Center of Mass" to move the particles to the new best predicted position. A virtual particle called center of mass is inserted to the formula of velocity to help the cognitive behavior component to avoid local optima, and to help maintaining the diversity of the swarm during the searching process. This increases the opportunity of fast convergence to global (or near global optima), where the center of mass particle will attract particles to the region of best found solutions, and this gives particles

the best chance to occupy the position of global best found solution during the search process. The PSOCoM technique is applied to the task of training an adaptive linear combiner to form a new stock market prediction model. This prediction model is used with some common indicators to increase the profit and decries the risk in stock market.

The structure of the proposed stock market prediction technique is assumed to be an adaptive linear combiner with parallel inputs as shown in Figure 1. The numbers of the inputs equal to the number of features in the input patterns, these features are abstracted from the stock market series such as closing prices and technical indicator values. The connecting weights of an adaptive linear combiner are considered as the particles and initially their values are set to random numbers in the range [-1, +1]. The swarm of particles is chosen to represent the initial solutions of the model. Each particle is adjusted during the training step by the way of minimizing the mean square error (MSE) as an objective function for PSOCoM technique. The formula of mean square error for the i_{th} particle is given in Equation 3.

$$MSE_i = \frac{\sum_{i=1}^{N} e_i^2(k)}{N} \quad (3)$$

Where: error $e_i(k)$.

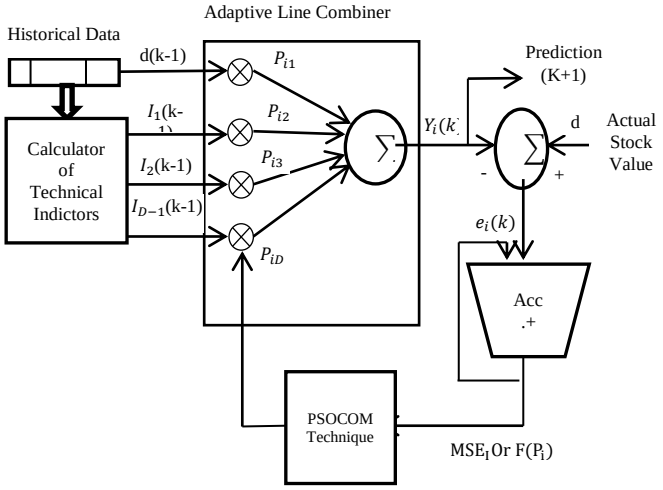


Fig. 1. The proposed stock market prediction model

The steps of the proposed prediction model are described below:

1. Start
2. For all history data of specific stock
3. {
4. //Abstract features (closing price, technical indictor values).
5. Calculate one- day –ahead price //d(k-1)
6. Calculate the technical indicators // $I_{(1)}(k-1), \dots I_{D-1}(k-1)$.
7. }
8. Using adaptive linear combiner(ALC)
9. Multiply the input of (ALCi) (d(k-1) and $I_{(1)}(k-1), \dots I_{D-1}(k-1)$) with weight of ALCi
10. For all training set(100 days as example)
11. {
12. Calculate the error $e_i(k) = d_i(k) - y_i(k)$ // $d_i(k)$ corresponding desired stock price (i.e. close price).
13. Calculate Mean Square Error for the i^{th} particle //$MSE_i = \frac{\sum_{i=1}^{N} e_i^2(k)}{N}$.
14. Use (MSE_i) as an objective function for PSOCoM technique to minimize MSE.
15. Use the output of improved PSOCOM to improve $y_i(k)$.
16. End.

IV. Evaluation of the Proposed Stock Market Prediction Technique

Results and Analysis

In this section, the performance of the proposed technique is evaluated. For that, the experimental data of used indices and the values of parameters settings are described. Finally the results and discussion of these results are presented.

A. Experimental Data

The data for the stock market prediction experiments have been collected for Standard's and Poor's 500 (S&P 500), National Association of Securities Dealers Automated Quotations 100 (NASDAQ-100), and Dow Jones Industrial Average (DJIA). These common known indices in USA stock market are used for evaluation of the proposed prediction model. These experimental data consist of daily close price and technical indicators derived from those indices. Total number of samples for the stock indices is 2500 trading days, from 2 January 2005 to 31 December 2014. Each sample consists of the opening price, highest price, lowest price, closing price and the total volume of the stocks traded for the day.

B. Parameter Settings

In this section, the setting of the parameters which were used in the experiments are presented. The inertia weight w was linearly decreased from 0.9 to 0.4; acceleration coefficients were set to $\mathbf{c_1} = \mathbf{c_2} = 2$; the maximum velocity was set to $V_{max} = 0.5$ and $X_{max} = 1$. The swarm size was set to 30. The maximum number of iterations was set to 100. Initialization range of particle positions was $-1 \leq x_i \leq 1$. All mean square errors (MSE) were computed over 30 runs. In short term prediction experiment, the training period was set to 100, 200, and 500 days to predict test period of 100 days. In long term prediction experiment, the training period was set to 1000 and 1500 days to predict test period of 750 days.

C. Results and Discussion

The evaluation of the proposed prediction model was performed using two types of prediction, short-term prediction and long-term prediction. The convergence characteristics of PSOCOM, ABFO, BFO, GA and PSO models for 1 day ahead prediction of DJIA , NASDAQ-100 and S&P 500 stock indices

for short term and long term prediction are shown in figures from 1 to 6, respectively.

To clarify the learning characteristics of the compared models in short and long term, the mean square error (MSE) was considered as a measure during training process.

As we see, Figures from 1 to 6, show the variation in the MSE vs. the number of iterations. It's clear that the MSE decreases when the number of iterations increases.

The comparison between of learning characteristics models to predict DJIA for one day advance in short term prediction is shown in Figure 2, and for long term in figure 3.

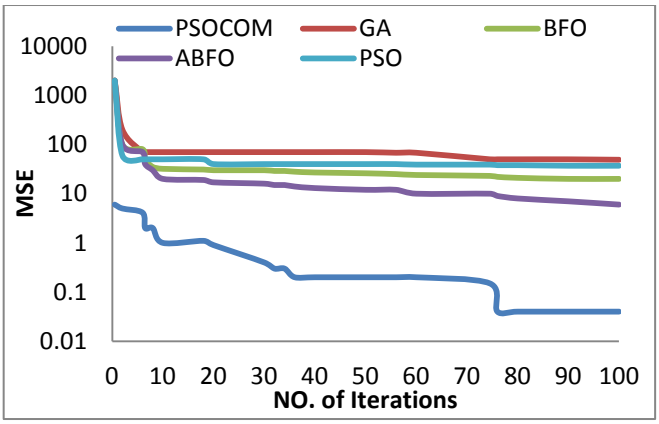


Fig. 2. Comparison of learning characteristics models to predict DJIA for one day advance (short term prediction)

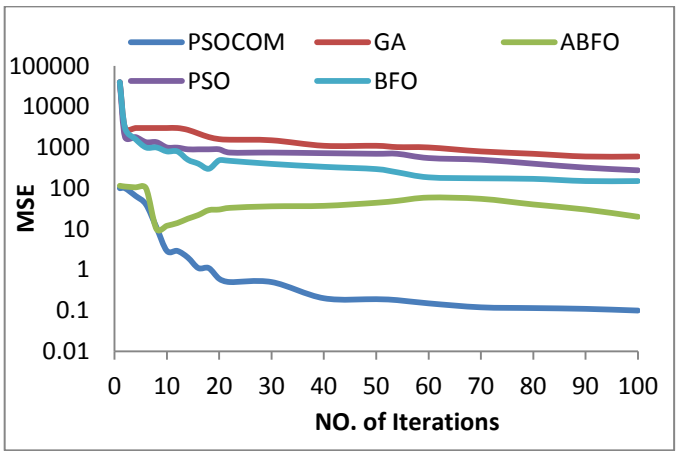


Fig. 3. Comparison of learning characteristics models to predict DJIA for one day advance (long term prediction)

The comparison of learning characteristics models to predict NASDAQ-100 for one day advance in short term prediction is shown in Figure 4, and for long term in Figure5.

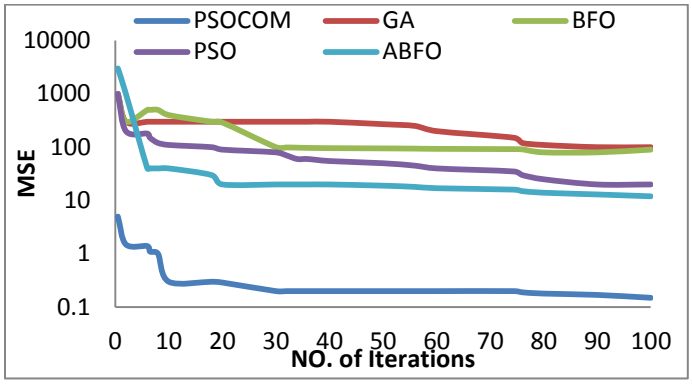


Fig. 4. Comparison of learning characteristics models to predict for one day advance (short term prediction)

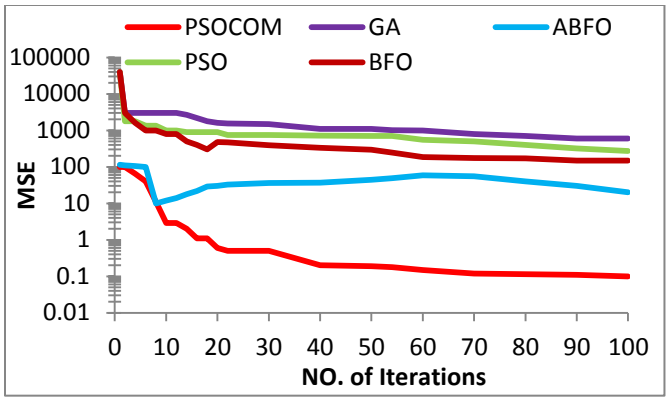


Fig. 5. Comparison of learning characteristics models to predict NASDAQ-100 for one day advance (long term prediction)

The comparison of learning characteristics models to predict S&P500 for one day advance in short term prediction is shown in Figure 6, and for long term in Figure 7.

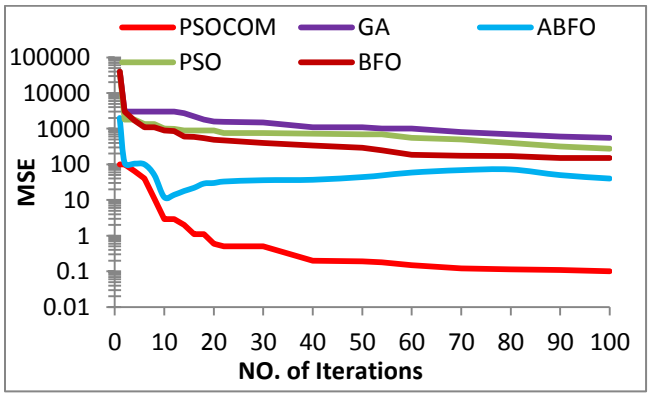


Fig. 6. Comparison of learning characteristics models to predict S&P500 for one day advance (short term prediction)

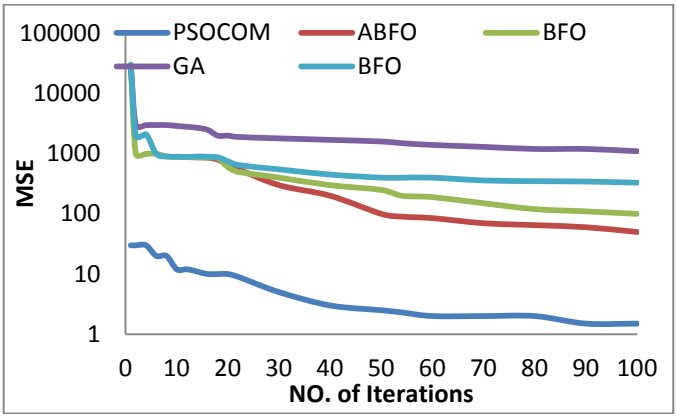


Fig. 7. Comparison of learning characteristics models to predict S&P500 for one day advance (long term prediction)

The figures proposed PSOCoM converged faster than the other methods such as ABFO, BFO, GA, and standard PSO during training process and reached the best minimum value of MSE indicating to the convergence of the weights. This emphasizes that the proposed PSOCoM overcomes the other compared methods, in learning characteristics, then the proposed prediction model superiors the other models in abstracting the important feature during training to perform more accurate prediction. According to MSE, the previous results clearly indicated that the proposed PSOCoM based model offers faster convergence during training followed by ABFO, PSO, BFO and then GA, based models.

The following curves show the comparison between the actual price and predicted price produced by the proposed PSOCOM model for DJIA, NASDAQ100, and S&P500 respectively.

Figures from 8 to 10 show the actual vs. predicted price for DJIA, NASDAQ100, and S&P500 indices for seven days ahead using the proposed PSOCOM model when test data are used as input. Comparison reveals very good agreement between the actual and predicted prices for DJIA, NASDAQ100, and S&P500 indices. It is in general observed that the proposed models predict DJIA, NASDAQ100, and S&P500 stock indices with less than 1% error for seven days ahead, because that the proposed PSOCOM converged faster than the other versions of other techniques during training process and reached the best minimum value of MSE indicating to the convergence of the weights. This leads to the fact that the proposed PSOCOM superiors the other techniques in learning characteristics, so the proposed prediction model superiors the other models in abstracting the important feature during training to perform more accurate prediction.

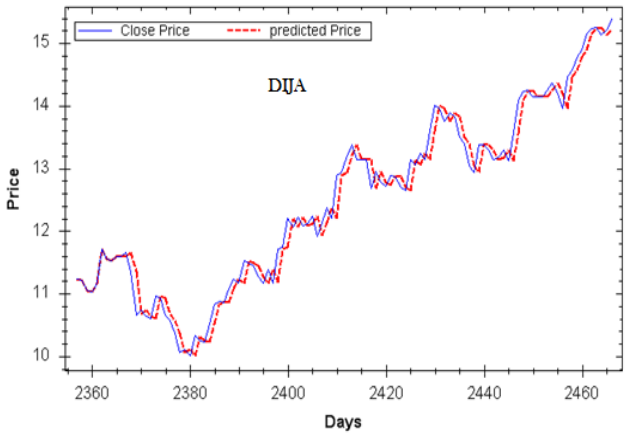


Fig. 8. DJIA index

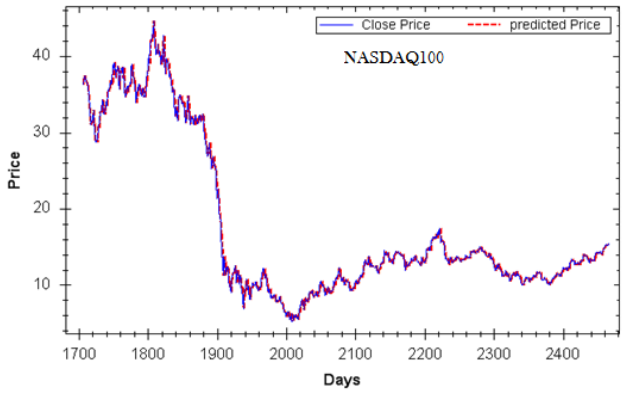


Fig. 9. NASDAQ100 index

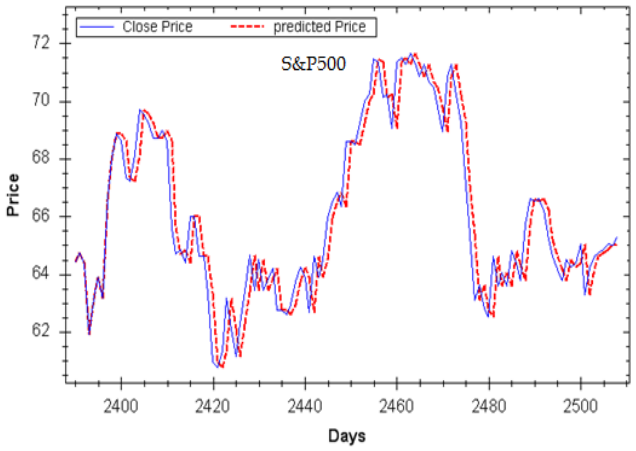


Fig. 10. S&P500 index

V. Conclusion and Future Work

A new stock market prediction model, based on the PSOCoM technique was proposed. The PSOCoM technique is used in the Suggested prediction model to adjust the weights of the adaptive linear combiner. The results of the experiments showed that the proposed forecasting technique is better than the other methods in terms of the accuracy of the prediction. The proposed forecasting model gives accurate prediction for short term and long term prediction. As a result, the proposed stock market prediction model is more efficient from the other compared models. So, the suggested prediction model is a new promising forecasting model for stock market prediction. In the future, more experiments and more comparison with other prediction models can be done. In addition, based on the proposed prediction technique, a new selection model can be designed in order to select the best stocks with highest profit and minimum risk. Also, it can develop a new automated system based on the presented technique to become an intelligent agent that makes trades in stock markets to get maximum return with minimum loss and gives the decision to buy or sell for the best selected stocks, and gives the final return at the end of the determined period.

References

[1] C.Hargreaves and Y.Hao,"Prediction of Stock Performance Using Analytical Techniques " Journal of Emerging Technologies in Web Intelligence, Vol 5, No 2 (2013), 136-142, May 2013.

[2] S.Arun Joe Babulo, B. Janaki, C. Jeeva,"Stock Market Indices Prediction with Various Neural Network Models" International Journal of Computer Science and Mobile Applications, Vol.2 Issue. 3, March-2014.

[3] T.Helstrom, and K.Holmstrom, "Predicting the stock market". Published as Opuscula ISRN HEV-BIB-OP-26-SE. 1998.

[4] D. Contras , O. Matei" Translation of the Mutation Operator from Genetic Algorithms to Evolutionary Ontologies' in International Journal of Advanced Computer Science and Applications(IJACSA), Volume 7 Issue 1, 2016 .

[5] A.Jordehi and J.Jasni, "Parameter selection in particle swarm optimisation: a survey", Journal of Experimental & Theoretical Artificial Intelligence, 25.4: 527-542, 2013.

[6] N. K. Jhankal ; D. Adhyaru "Bacterial foraging optimization algorithm: A derivative free technique" Engineering (NUiCONE), ,Nirma University International Conference on 2011.

[7] R. Vijay 'Intelligent Bacterial Foraging Optimization Technique to Economic Load Dispatch Problem' in International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue-2, May 2012.

[8] H. Shen, Y. Zhu" Adaptive Bacterial Foraging Optimization Algorithm Based on Social Foraging Strategy" Journal of Networks, Vol 9, No 3 (2014), 799-806, Mar 2014.

[9] J. Li " Analysis and Improvement of the Bacterial Foraging Optimization Algorithm" Journal of Computing Science and Engineering, Vol. 8, No. 1, March, pp. 1-10, 2014.

[10] R.Majhi ,G.Panda,B.Majhi, andG.Sahoo, , "Efficient prediction of stock market indices using adaptive bacterial foraging optimization (ABFO) and BFO based techniques", Expert Systems with Applications. Vol. 36(6), pp. 10097-10104, 2009.

[11] K. Kyoung-jae, L. Won Boo. "Stock market prediction using artificial neural networks with optimal feature transformation". Neural Computing and Applications (2004),Volume: 13, Issue: 3, Publisher: Citeseer, Pages: 255-260

[12] K. Kim, I. Han. "Genetic algorithms approach to feature discretization in artificial neural networks for the prediction of stock price index". Expert Systems with Applications, Volume 19, Issue 2, August 2000, Pages 125–132.

[13] J. Kennedy, andC. Eberhart, , "Particle Swarm Optimization". Proceedings of the 1995 IEEE International Conference on Neural Networks, Australia, 1995, pp. 1942-1948.

[14] J .Štěpánek, ,J. Šťovíček, R. Cimler " Application of Genetic Algorithms in Stock Market Simulation" Cyprus International Conference on Educational Research (CY-ICER-2012)North Cyprus, US08-10 February, 2012.

[15] L. Demidova, E. Nikulchev , Y. Sokolova "The SVM Classifier Based on the Modified Particle Swarm Optimization"in International Journal of Advanced Computer Science and Applications(IJACSA), Volume 7 Issue 2, 2016.

[16] Y.Zhang, S. Wang, and G.Ji "A Comprehensive Survey on Particle Swarm Optimization Algorithm and Its Applications" journal of Mathematical Problems in Engineering, Volume , Article ID 931256, 38 pages,2015.

[17] H. Subramanian, S. Ramamoorthy, P.Stone, B. J. Kuipers "Designing safe, profitable automated stock trading agents using evolutionary algorithms" In Proceedings of the Genetic and Evolutionary Computation Conference, July 2006.

[18] E.El. Seidy,; "A New Particle Swarm Optimization based Stock Market Prediction Technique "accepted to publicationin International Journal of Advanced Computer Science and Applications,Volume 7 No. 2, February 2016.

[19] R. Jamous, E. El. Seidy, A. Tharwat, B. I. Bayoumi" A new Particle Swarm with Center of Mass Optimization", in International Journal of Engineering Research & Technology (IJERT), Vol. 4 Issue 05, 2015, PP:312-317

[20] R. Jamous, E. El. Seidy, A. Tharwat, B. I. Bayoumi"Modifications of Particle Swarm Optimization Techniques and Its Application on Stock Market: A Survey "in International Journal of Advanced Computer Science and Applications(IJACSA), Volume 6 Issue 3, 2015.

Throughput Measurement Method Using Command Packets for Mobile Robot Teleoperation Via a Wireless Sensor Network

Kei SAWAI
Department of Information and Communication Engineering, Tokyo Denki University
Tokyo, Japan

Ju Peng
Department of Information and Communication Engineering, Tokyo Denki University
Tokyo, Japan

Tsuyoshi Suzuki
Department of Information and Communication Engineering, Tokyo Denki University
Tokyo, Japan

***Abstract*—We are working to develop an information gathering system comprising a mobile robot and a wireless sensor network (WSN) for use in post-disaster underground environments. In the proposed system, a mobile robot carries wireless sensor nodes and deploys them to construct a WSN in the environment, thus providing a wireless communication infrastructure for mobile robot teleoperation. An operator then controls the mobile robot remotely while monitoring end-to-end communication quality with the mobile robot. Measurement of communication quality on wireless LANs has been widely studied. However, a throughput measurement method has not been developed for assessing the usability of wireless mobile robot teleoperation. In particular, a measurement method is needed that can handle mobile robots as they move around an unknown environment. Accordingly, in this paper, we propose a method for measuring throughput as a measure of communication quality in a WSN for wireless teleoperation of mobile robots. The feasibility of the proposed method was evaluated and verified in in practical field test where an operator remotely controlled mobile robots using a WSN.**

***Keywords*—Wireless Sensor Networks; Rescue Robot Teleoperation; Communication Quality Measurement**

I. INTRODUCTION

A wireless sensor network (WSN) would be useful for teleoperation of a mobile robot, but methods for measuring throughput between the operator and robot have not been argued enough. A common approach for throughput measurement is to calculate the maximum transfer amount per unit time, which provides the communication speed for delivery of payloads over the connection between sensor nodes. This method enables high-precision evaluation of throughput in networks where wireless communication quality is stable. However, throughput cannot be accurately measured in unstable networks where various types of noise occur. Furthermore, measurement by this method requires a few minutes, because many communication packets are sent. A rescue robot moves around and explores a disaster area, but such an environment contains debris and many obstacles leading to a risk of noise due to multipath fading of radio waves. Therefore, to operate the rescue robot promptly and smoothly, a rapid measurement method is necessary for evaluating the usability of rescue robot teleoperation.

Gathering information with a rescue robot in a disaster area is very important for assessing the situation, avoiding secondary disasters, and managing disaster mitigation [1]-[6]. In general, gathering information from a bird's eye view with an unmanned air vehicle is a useful method for ascertaining the situation in a disaster area. However, in an urban area with many underground spaces where information cannot be gathered from the air, the extent of the damage can be difficult to assess. Gathering information in underground spaces is important for avoiding secondary disasters [7]-[8] and planning rescue operations. When the communication infrastructure is damaged, cooperation among rescue workers is hindered by communication disconnection between aboveground and underground spaces, meaning that there is a high risk of rescue workers being involved in a secondary disaster, particularly in situations that are continuously or unexpectedly changing. Research on a disaster area information gathering system including WSN technology and rescue robots has recently been conducted considering past incidents in enclosed areas [7]-[8]. This system enables reduction of the risk of secondary disaster by employing a rescue robot rather than a human, and the WSN provides a wireless communication infrastructure for wireless rescue robot teleoperation. The WSN consists of many sensor nodes deployed by the robot and distributed spatially for cooperatively monitoring environmental conditions, such as temperature, sound, vibration, air pressure, and motion. The topology of the WSN changes automatically to optimize the communication path according to the communication quality between sensor nodes (Fig. 1). The rescue robot is defined as one of the sensor nodes constructing the WSN. Then, the WSN provides a wireless communication infrastructure where none existed before. Therefore, WSNs have been discussed as a method for gathering information and constructing communication infrastructure in disaster areas, and have been studied widely. Also, we have been investigating a robot wireless sensor network (RWSN) (Fig. 2) [9]-[13].

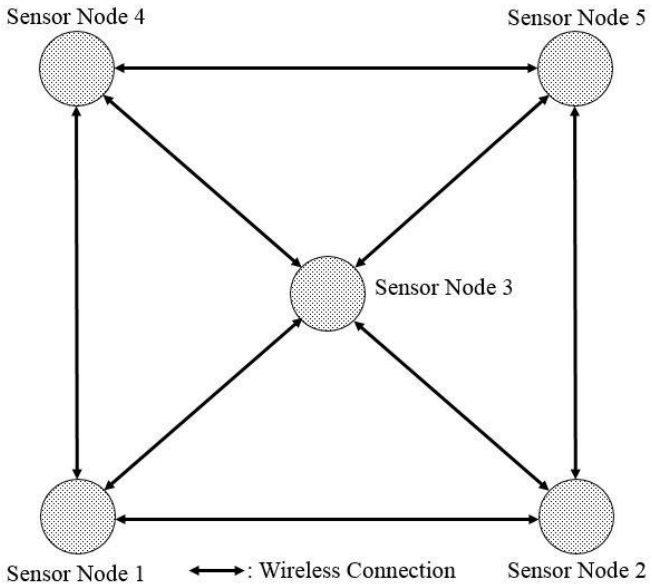


Fig. 1. Actively changed network topology of wireless sensor networks

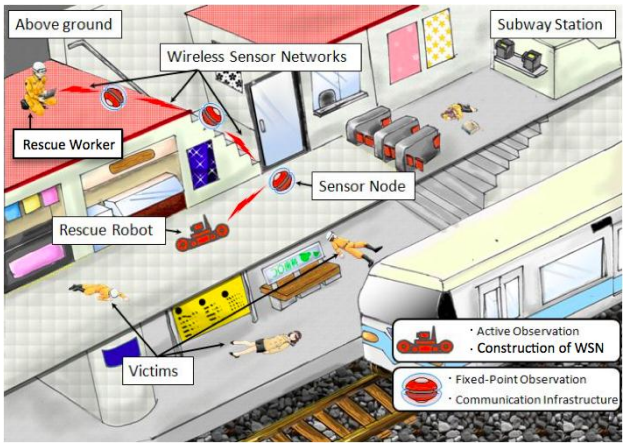


Fig. 2. Example deployment of a gathering disaster area information system using a WSN and rescue robot

However, in rescue robot teleoperation using a WSN, it is difficult to maintain communication connectivity between the operator and the rescue robot. Doing so requires measuring the throughput between the rescue robot and operator. Yet, the robot is a continually changing communication point because it moves around to gather information in the disaster area. The propagation of radio waves in a disaster area, such as an underground space, is difficult to estimate because of complicated building construction and damage caused by the disaster. Furthermore, the communication distance between the rescue robot and the adjacent sensor node in an ad hoc network changes as the robot moves around, and so the electric field signal strength between the rescue robot and adjacent sensor node also fluctuates. For these reasons, existing methods cannot perform accurate measurement of throughput between the rescue robot and operator in an unstable communication environment such as a post-disaster underground space.

Most existing methods for measuring throughput assume a normal living space as the environment for the WSN, but this assumption does not hold for a post-disaster underground space. Furthermore, the existing methods are not intended for networks with a continually moving sensor node such as a rescue robot. In an RWSN, accurate measurement of throughput cannot be performed because the moving sensor node continually changes the communication point. For existing method to measure throughput accurately, many packets must be sent from the measurer to the measurement object, and then the throughput is calculated by using the size and number of packets received by the measurement object. Sending and receiving packets requires a few minutes, which reduces the usability of rescue robot teleoperation. When such a method is used, the operator cannot control the rescue robot during throughput measurement because many communication packets are sent instead of control packets. If the robot cannot communicate over the network, it could become uncontrollable and cause a secondary disaster. To prevent network disconnection while deploying sensor nodes, during each interval the operator must stop the rescue robot by utilizing command packets and then measure throughput as indicator of the strength of wireless communication connectivity between the rescue robot and operator. Thus, existing methods have the problem of reducing the practical utility of the disaster area information gathering system. Furthermore, in research on mobile robots and WSN systems, the communication connectivity of the system has not been defined clearly [14]-[17]. Most studies adopt received signal strength indicator (RSSI) as a measure of wireless communication connectivity to maintain the network configuration of the system, and RSSI is measured by an RF module over a very short time [18]-[26]. However, RSSI indicates communication quality in the physical layer and cannot evaluate communication quality in the transport layer for sending and receiving communication packets. For example, even if the RSSI level is good, a communication environment with strong radio wave interference or absorption will have reduced throughput. Thus, RSSI cannot correctly indicate the strength of communication connectivity. Quick evaluation of communication connectivity with high precision is necessary for monitoring packet traffic in the transport layer.

Accordingly, here we considered an approach to reduce the time required for measuring throughput between the rescue robot and the operator by using the command packets for robot control, rather than communication packets. The proposed system uses a TCP/IP-compliant communications protocol and an IEEE 802.11 wireless LAN protocol that are compatible with teleoperation of a rescue robot. The proposed method was actually implemented using a WSN and rescue robot in a field test to examine the feasibility of the measurement time reduction. The next section describes the RWSN of our proposed disaster area information gathering system comprising a rescue robot and WSN. Section 3 presents details of the proposed throughput measurement method, and then field test results for performance evaluation are presented in Sections 4 and 5. Conclusions are given in Section 6.

II. Robot Wireless Sensor Networks

An RWSN is constructed from a WSN by using a rescue robot that deploys sensor nodes along its movement path. In our sensor node deployment method for constructing the WSN, the rescue robot delivers sensor nodes that were connected in

advance and configured with a defined routing path. Thus, the rescue robot delivers all the sensor nodes with a defined routing path into an underground space along the robot's movement path. Deployed sensor nodes communicate with adjacent sensor nodes wirelessly, and as sensor nodes are deployed, the wireless communication distance between the operator and the rescue robot is extended via the WSN in the underground space. The operator controls the rescue robot through the communication infrastructure provided by the constructed WSN. In the network topology of the RWSN, sensor nodes are linearly connected to prevent communication errors due to auto-routing control (Fig. 3, Fig. 4).

Generally, the WSN can automatically decide the routing path of data transfer by various means, including the RSSI between adjacent sensor nodes, the time difference of arrival of communication packets, the end-to-end throughput, or the packet loss rate. The routing paths in the WSN are reconstructed according to changes in communication quality. However, repeated reconstruction of the routing path causes communication disconnection and reconnection between sensor nodes, a situation that causes serious problems in the teleoperation of mobile robots. The rescue robot cannot establish a wireless communication infrastructure by an existing method because the risk of disconnection between rescue robot and adjacent sensor node is high. Teleoperation with irregular interruptions of wireless communication degrades the usability of the rescue robot, and so information gathering performance degrades as well. Such a situation must be avoided to prevent the robot from becoming uncontrollable and creating a risk of secondary disaster. Furthermore, communication quality strongly varies according to secondary disaster damages in the disaster area. Therefore, the routing path is not repeatedly reconstructed in the WSN.

In the proposed RWSN system, it is necessary to maintain communication connectivity in order to predict disconnection between a sensor node to be deployed and the adjacent sensor node. To predict disconnection when extending the WSN, the change in throughput should be monitored between the sensor node to be deployed and the adjacent sensor node. End-to-end throughput between the operator and the rescue robot is known by connecting all sensor nodes in advance. Maintaining this throughput stabilizes end-to-end communication connectivity, and then it is necessary to maintain the throughput value between adjacent sensor nodes. An IEEE 802.11 wireless LAN module can control throughput speed automatically by monitoring RSSI, allowing the throughput to be decreased when RSSI drops below a defined threshold. Thus, in order to maintain throughput between a sensor node to be deployed and the adjacent sensor node, it is necessary to monitor the RSSI value at the same time as throughput measurement. Maintaining throughput with RSSI monitoring stabilizes end-to-end communication connectivity in the RWSN system. Regarding elemental technologies for an RWSN information gathering system in post-disaster underground spaces, we have previously reported details of the RWSN, development of impact-resistant sensor node, a rescue robot with a sensor node deployment mechanism, and an RWSN construction method that employs a rescue robot with WSN technology. In particular, we have demonstrated the importance of communication connectivity between deployed sensor node and adjacent sensor nodes.

In this paper, the proposed method to reduce throughput measurement time is implemented and evaluated on the assumption that the RWSN is already constructed in the target environment. Details of this method are described in the next section.

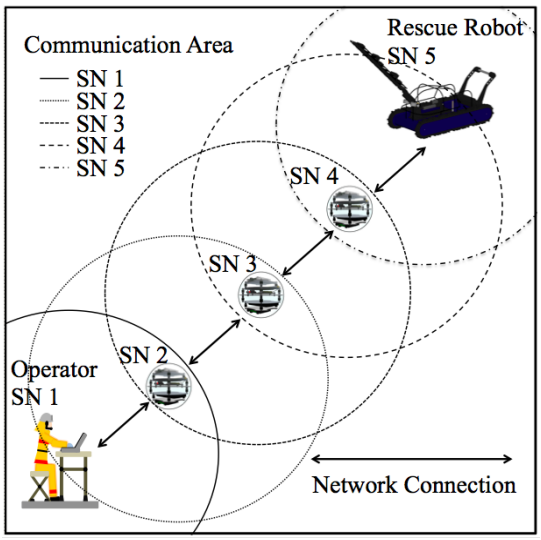


Fig. 3. Teleoperation of a rescue robot in an RWSN (SN: sensor node)

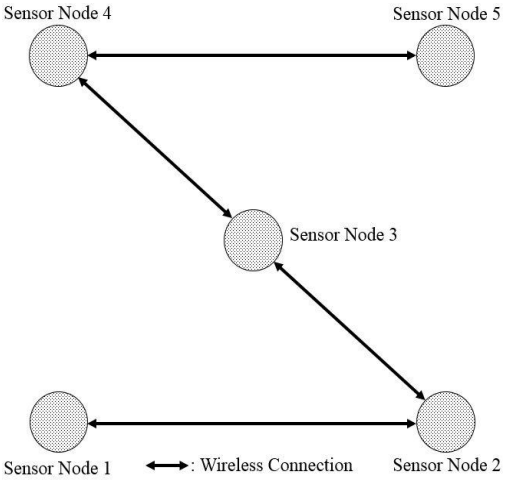


Fig. 4. Network topology of the RWSN

III. Throughput Measurement Method Using Command Packet for Teleoperation of Rescue Robot

A. Assumed Environment for Constructing the WSN

We assume that in a post-disaster underground space, it is necessary to construct the disaster area information gathering system on the first basement floor between the exit and entrance. In particular, this environment is configured by two EXIT stairs, and 30 m path between them. In the Japanese Building Standards Law, 30 m is specified as the maximum permissible distance between EXIT stairs for escaping to the ground from underground structures. Moreover, most paths in underground spaces are designed to be straight in consideration

of emergency evacuation. Therefore, we assume that RWSN is constructed in this environment, and the design of the proposed method is based on these environmental conditions (Fig. 5).

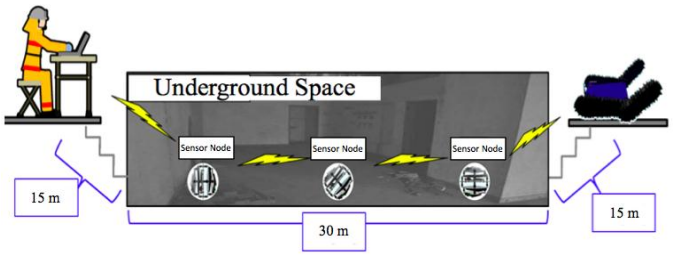


Fig. 5. Assumed environment for constructing the RWSN

B. Throughput Measurement Method without Interruption of Communication for Rescue Robot Teleoperation

We propose a throughput measurement method to reduce measurement time by measuring small command packets that are sent and received for mobile robot teleoperation (Fig. 6). The operator sends command packets configured from action parameters for moving and sensing, and then the mobile robot sends back status information based on captured image data, temperature, movement speed, acceleration, and so forth. These communications between the operator and the rescue robot should be continued without interruption to maintain the operability of teleoperation. Even when the operator does issue a command to the rescue robot, packets are continuously sent and received by the operator and the rescue robot in the background to predict communication disconnection and an uncontrollable status of the rescue robot. Generally, teleoperation systems use small packets for stable communication quality over a network.

Therefore, we propose a throughput measurement method that uses these command packets. Throughput is the amount of transferred data per unit time, and is a measure of communication speed. Throughput T [bps] is calculated by using equation (1).

$$T = \frac{8B}{t} \quad (1)$$

Here, B [byte] is the packet size and t [s] is the transfer time.

The size of command packets for teleoperation of the rescue robot is defined as described above. Then, ID numbers are consecutively assigned to the packets as they are sent to the operator or the rescue robot. Furthermore, the transmission interval between packets is constant. Thus, in the proposed measurement method, throughput is calculated by equation (1). By using command packets in this way, it is not necessary to interrupt teleoperation of the rescue robot teleoperation or to send many communication packets for throughput measurement as in existing method. Moreover, the operator can obtain the throughput within a few seconds by the proposed method, whereas the existing method takes a few minutes for throughput measurement. In a network system implementing various network applications, existing methods measure the maximum value of throughput between terminals by sending many communication packets, which preload communication band to the utmost remits in the network. However, a network system that uses only a defined network application such as disaster area information gathering system does not have to maximize throughput. To maintain communication connectivity, it is important to monitor the change in throughput between terminals. Although the maximum throughput value cannot be measured, the proposed method can evaluate the throughput stability more quickly than the existing method. Therefore, we consider our approach more suitable than existing methods for systems such as the disaster area information gathering system in post-disaster underground space.

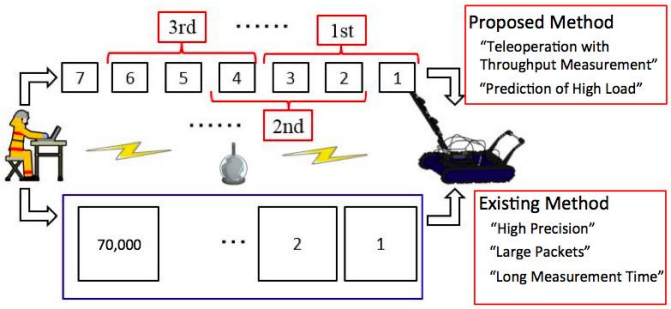


Fig. 6. Comparison of existing and proposed throughput measurement methods

C. Definition of Packet Size

To measure an aspect of communication quality such as throughput, it is important to define packet size. If the packet size is too small, the packet transfer time is adversely affected from noise due to network traffic. If the packet size is too big, a transferred packet is automatically divided into multiple packets by TCP/IP and UDP. This mechanism is defined as IP fragmentation, in which transferred packets are divided in order to pass a link where the maximum transmission unit is smaller than the original datagram size. However, this mechanism changes the transmission time of packets between terminals. Therefore, in our proposed method, we adopt a packet size of 1,400 bytes to prevent IP fragmentation.

D. Packet Transfer Method in the Proposed Throughput Measurement

The proposed packet transfer method transfers multiple packets to an extent that does not adversely affect the communication path. Throughput measurement by command packet transfer can be affected by various types of noise along the network path, and so the throughput is inaccurate due to transfer of packets, changes in the surrounding environment, and electrical noise. In particular, the building structures in a post-disaster environment continually change as a result of secondary disaster damage, and the features of the radio wave propagation environment also continuously changes. Thus, in the throughput measurement method, multiple packets are transferred to predict the influence of noise along the communication path, and the overall average of all received packets at a determined time is taken as the measurement result. Throughput measured in this way takes into consideration environmental noise, and can be used for continuous monitoring of communication connectivity between terminals. Generally, a teleoperation system of a rescue robot continuously communicates in order to improve the usability of

teleoperation between the operator and rescue robot by sending and receiving command packet. Therefore, the throughput measurement method using continuously transfer of command packets on the communication path can be widely implemented in teleoperation systems of rescue robots. As described in the next section, we conducted a field test to evaluate the proposed throughput measurement method using an actual rescue robot and WSN.

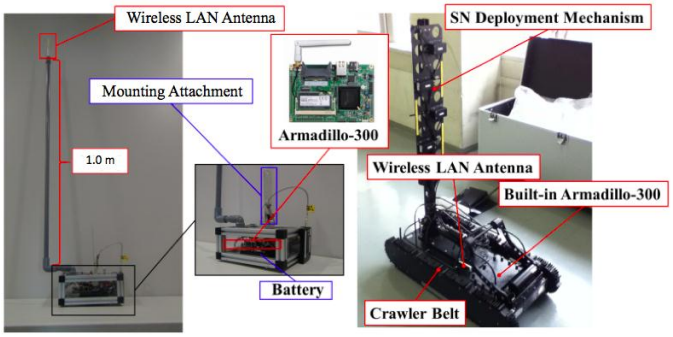
IV. Performance Evaluation of Throughput Measurement Method Using Command Packets Controlling Rescue Robot

A. Performance Evaluation

This section presents a performance evaluation of our proposed throughput measurement method. We measured throughput by using an existing method and the proposed method, and then experimental results are verified in a comparative evaluation. In the experiment, an RWSN is adopted as a disaster area information gathering system in a post-disaster underground space.

B. Experimental Condition

The RWSN used in this experiment was constructed by a rescue robot deploying sensor nodes. The rescue robot was equipped with three sensor nodes and deployed them at 15 m intervals. Thus, the RWSN was intended to span 60 m, and the WSN provided a wireless communication infrastructure with a communication path of this length. The operator used the RWSN to control the rescue robot, and the RWSN construction was based on a sensor node deployment method that we proposed in the past. The sensor node deployment method was provided to predict network disconnection when constructing the RWSN. In the experiment, the proposed method defined the number of transferred packets to be used for a single measurement as 40 packets in consideration of packet traffic noise. Among existing methods comparable to the proposed method, "utest" (NTTPC Communications Ltd.) was adopted, which measures an accurate value of throughput by sending 1,000,000 packets each of 1500 bytes for a single measurement. In each of the existing method and the proposed method, measurements were performed 10 times, and then the overall mean was calculated.



(a) Developed sensor node (b) Rescue robot with sensor node deploying mechanism

Fig. 7. Sensor node and rescue robot using in constructing RWSN

To construct the WSN, we adopted the sensor node device developedin our previous studies (Fig. 7 (a)). The sensor node is equipped with a CPU board, memory device, CompactFlash disc drive, IEEE 802.11b/g wireless LAN module, a digital camera, an A/D converter, and a battery. The sensor nodes are controlled by software implemented in Linux (Debian). In this way, a WSN was constructed by utilizing the AODV-uu protocol for ad hoc networks. Table 1 shows the specification of our developed sensor node. As shown in Fig. 7 (b), a crawler-type mobile robot (S-90LWX, Topy Industries, Ltd. was used as the rescue robot in this experiment. A sensor node deployment mechanism was developed for WSN construction and installed on the rescue robot; this mechanism can hold five sensor nodes using five solenoid-operated locks. Figure 8 shows the framework of this mobile robot with the sensor node deployment mechanism. Then, the operator can control the crawler robot and the deployment mechanism remotely by utilizing TCP/IP and UDP. The experiment is performed in the passageway with a length of 300 [m] or more in Tokyo Denki University. Figure 9 (a), (b) shows experimental overview in this performance evaluation.

TABLE I. Specifications of Sensor Nodes for Constructing the RWSN

Sensor Node	
Operating system	Linux Kernel 2.6 (Debian)
CPU board	Armadillo-300 (ARM 200 MHz)
Weight	1.5 kg
Height × Width × Length	225 mm × 180 mm × 380 mm
Battery No. 1	Output: 5 V, 1.8 mAh
Battery No. 2	Output: 5 V, 2.1 mAh
Operating time	10 [h]

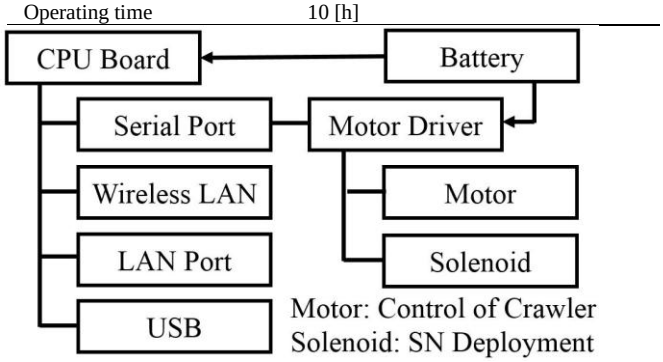
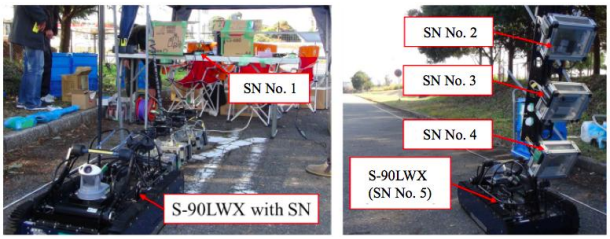


Fig. 8. Configuration of the rescue robot with sensor node deployment mechanism



(a) Experimental Condition (b) Rescue Robot mounting SNs

Fig. 9. Scenes from performance evaluation field test

C. Experimental Result

Figure 10 shows the experimental results. Figure 11 shows a comparative evaluation of error range. Throughput values in both the existing method (utest) and the proposed method showed similar trends in the performance evaluation test. The

proposed method gave slightly higher throughput values than utest, showing that the proposed method did not slow down packet traffic.

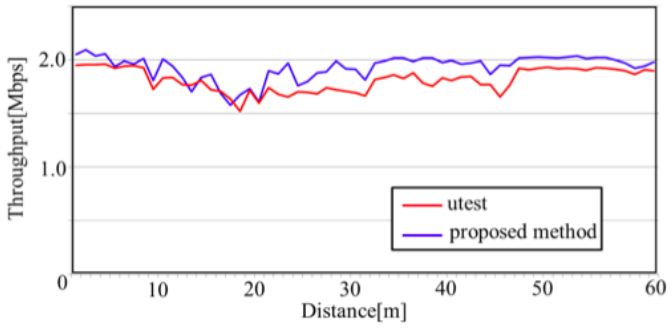


Fig. 10. Experimental Results Using "utest" and Proposed Method

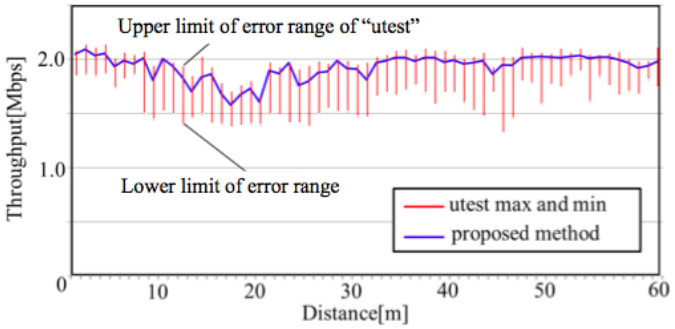


Fig. 11. Comparative Evaluation of Error Range

V. Discussion

The correlation coefficient between the throughputs measured by the utest and the proposed method was 0.8655, showing that there was a high correlation in this performance evaluation. Overall mean values of throughput measured by the proposed method were slightly higher than the mean values measured by utest, because the proposed method did not produce excessive packet traffic on the communication path. However, all throughputs measured by the proposed method were within the upper limit of the error range of measurements by utest. Thus, we confirmed that the throughput values obtained by the proposed method were correct. Furthermore, this proposed method enabled to control of the rescue robot by a remote operator. These results show the applicability of the proposed method to RWSN.

VI. Conclusion

In this paper, we proposed a method for reducing the time required for measuring the throughput between a rescue robot and an operator in a constructed WSN with the aim of improving the usability of a disaster area information gathering system in post-disaster underground spaces. This method was designed to avoid excessive sending of communication packets as in existing methods and to suit an assumed underground disaster environment. The proposed method was designed to utilize the command packets for controlling the rescue robot and used a wireless LAN IEEE 802.11 protocol compatible with wireless teleoperation of a rescue robot. The developed method was implemented in a rescue robot and WSN, and then feasibility of the proposed method was confirmed on the basis of measurement error and measurement time from the results of a performance evaluation field test for comparison with utest. Furthermore, throughput measurement using the proposed method allowed the throughput to be measured without stopping the robot.

In future work, we plan to improve the graphical user interface to display throughput to the operator while controlling the rescue robot. Also, we will consider a method for measuring packet jitter between the rescue robot and operator for implementing a video streaming function that uses the rescue robot's camera to gather disaster area information in a post-disaster underground space.

Acknowledgment

This work was partially supported by MEXT KAKENHI Grant Number 15631330.

We are grateful to Mr. Shigeaki Tanabe and Mr. Ryuta Kunimoto who were graduate students in Tokyo Denki University in 2011 and 2012 respectively for their support in the RWSN system development.

References

[1] CHI Hao-yuan, LIU Xu, XU Xiao-dong, "A Framework for Earthquake Disaster Mitigation System," Proceedings of 2011 China located International Conference on Information Systems for Crisis Response and Management (ISCRAM), pp.490-495, 2011.

[2] Huang AN, "China's Emergency Management Mechanisms for Disaster Prevention and Mitigation," Proceedings of International Conference on E-Business and E-Government (ICEBEG), pp.2403-2407, 2010.

[3] Yoshiaki KANAEDA, Kazushige MAGATANI, "Development of the device to detect SPO2 in the Field,"31st Annual International Conference of the IEEE EMBS, pp.412-415, September 2009.

[4] Y. Kawata, "Disaster Mitigation due to next Nankai earthquake tsunami occurring in around 2035," Proceedings of International Tsunami Symposium 2001, session 1, pp. 315-329, 2001.

[5] Y. Kawata, "The great Hanshin-Awaji earthquake disaster: damage, social response, and recovery," Journal of Natural Disaster Science, Vol. 17, No. 2, pp.1-12, 1995.

[6] H. Kawakata, Y. Kawata, H. Hayashi, T. Tanaka, K. C. Topping, K. Yamori, P. Yoshitomi, G. Urakawa and T. Kugai, "Building an integrated database management system of information on disaster hazard, risk, and recovery process," Annuals of Disas. Prev. Res. Inst., Kyoto Univ., No.47 C, 2004.

[7] Abishek T K, Chithra K R and Maneesha V. Ramesh, "ADEN:Adaptive Energy Efficient Network of Flying Robots Monitoring over Disaster Hit Area," Proceedings of 8th IEEE International Conference on Distributed Computing in Sensor Systems (IEEE DCOSS), pp.306-310, 2012.

[8] Abishek T K, Chithra K R, Maneesha V Ramesh, "AER: Adaptive Energy Efficient Routing Protocol For Network of Flying Robots Monitoring over Disaster Hit Area," Proceedings of 21st Annual Wireless and Optical Communi-cations Conference (WOCC), pp.166-169, 2012.

[9] K. Sawai, T. Suzuki, H. Kono, Y. Hada and K. Kawabata, "Development of a SN with impact-resistance capability for gathering disaster area information," 2008 Internatio-nal Symposium on Nonlinear Theory and its Applications (NOLTA2008), pp.17-20, 2008.

[10] Tsuyoshi Suzuki, Kei Sawai, Hitoshi Kono and Shigeaki Tanabe, "Sensor Network Deployment by Dropping and Throwing Sensor Node to Gather Information Underground Spaces in a Post-Disaster Environment," Descrete Event Robot, iConcept PRESS, in Press. 2012.

[11] K. Sawai, H. Kono, S. Tanabe, K. Kawabata, T. Suzuki, "Design and Development of Impact Resistance Sensor Node for Launch Deployment into Closed Area," In international journal of sensing for

industry(Sensor Review), Emerald Group Publishing Ltd., Vol. 32, pp.318 – 326, 2012.

[12] S. Tanabe, K. Sawai and T. Suzuki, "Sensor Node Deployment Strategy for Maintaining Wireless Sensor Network Communication Connectivity," International Journal of Advanced Computer Science and Applications (IJACSA), The Science and Information organization, Vol.2, No. 12, pp.140 – 146, 2011.

[13] H. Sato, K. Kawabata and T. Suzuki, "Information Gathering by wireless camera node with Passive Pendulum Mechanism," International Conference on Control, Automation and Systems 2008 (ICCAS2008), pp.137-140, 2008.

[14] T. Yoshida, K. Nagatani, E. Koyanagi, Y. Hada, K. Ohno, S. Maeyama, H. Akiyama, K. Yoshida and Satoshi Tadokoro, "Field Experiment on Multiple Mobile Robots Conducted in an Underground Mall," Field and Service Robotics Springer Tracts in Advanced Robotics, vol. 62, pp365-375, 2010.

[15] H. Jiang, J. Qian, and W. Peng, "Energy Efficient Sensor Placement for Tunnel Wireless Sensor Network in Underground Mine," Proceedings of 2nd International Conference on Power Electronics and Intelligent Transportation System(PEITS 2009), pp. 219-222, 2009.

[16] J. Xu, S. Duan and M. Li, "The Research of New Type Emergency Rescue Communication System in Mine Based on Wi-Fi Technology," Proceedings of IEEE 3rd International Conference on Communication Software and Networks (ICCSN), pp. 8-11, 2011.

[17] K. Nagatani, S. Kiribayashi, Y. Okada, K. Otake, K. Yoshida, S. Tadokoro, T. Nishimura, T. Yoshida, E. Koyanagi, M. Fukushima and S. Kawatsuma, "Emergency Response to the Nuclear Accident at the Fukushima Daiichi Nuclear Power Plants using Mobile Rescue Robots," Journal of Field Robotics, vol. 30, no. 1, pp. 44-63, 2013.

[18] Parker, E., L., Kannan, B., Xiaoquan, F., Yifan, T. (2003). Heterogeneous Mobile Sensor Net Deployment Using Robot Herding and Line of Sight Fromations, Proceedings of 2003 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS2003), Volume 3. pp.2488-2493, 2003.

[19] T. Umeki, H. Okada, K. Mase, "Evaluation of Wireless Channel Quality for an Ad Hoc Network in the Sky SKYMESH," Proceedings of Sixth International Symposium on Wireless Communication Systems 2009 (ISWCS'09). pp.585-589, 2009.

[20] Helge-Bjorn Kuntze, Christian W. Frey, Igor Tchouchenkov, Barbara Staehle, Erich Rome, Kai Pfeiffer, Andreas Wenzel and Jurgen Wollenstein, "SENEKA - Sensor Network with Mobile Robots for Disaster Management," Homeland Security (HST), pp.406-410, 2012.

[21] E. Budianto, M.S. Alvissalim, A. Hafidh, A. Wibowo, W. Jatmiko, B. Hardian, P. Mursanto and A. Muis, "Telecommunication Networks Coverage Area Expansion in Disaster Area using Autonomous Mobile Robots : Hardware and Software Implementation," Proceedings of International Conference on Advanced Computer Science and Information Systems (ICACSIS), pp.113-118, 2011.

[22] Andrew Chiou, and Carol Wynn, "Urban Search and Rescue Robots in Test Arenas: Scaled Modeling of Disasters to Test Intelligent Robot Prototyping," Proceedings of International Conference on Autonomic and Trusted Computing (ATC), pp.200-205, 2009.

[23] A. Howard, J. Matric, and J. Sukhatme, "Mobile Sensor Network Deployment using Potential Fields," A Distributed Scalable Solution to the Area Coverage Problem, Distributed Autonomous Robotics Systems 5, Springer-Verlag. pp.299-308, 2002.

[24] Wing-Yue Geoffrey Louie, and Goldie Nejat, "A victim identification methodology for rescue robots operating in cluttered USAR environments," Advanced Robotics, vol. 27, issue. 5, pp. 373-384, 2013.

[25] Andrew Markham and Niki Trigoni, "Magneto-Inductive NEtworked Rescue System (MINERS):Taking Sensor Networks Underground," Proceedings of the 11th international conference on Information Processing in Sensor Networks (IPSN '12), pp. 317-328, 2012.

[26] Josh D. Freeman, Vinu Omanan, and Maneesha V. Ramesh, "Wireless Integrated Robots for Effective Search and Guidance of Rescue Teams," Proceedings of 8th International Conference on Wireless and Optical Communications Networks (WOCN 2011), pp. 1-5, 2011.

User Interface Menu Design Performance and User Preferences: A Review and Ways Forward

Dr Pietro Murano
Department of Computer Science, The Universal Design of ICT Research Group
Oslo and Akershus University College of Applied Sciences, Oslo, Norway

Margrete Sander
Department of Computer Science, The Universal Design of ICT Research Group
Oslo and Akershus University College of Applied Sciences, Oslo, Norway

***Abstract*—This review paper is about menus on web pages and applications and their positioning on the user screen. The paper aims to provide the reader with a succinct summary of the major research in this area along with an easy to read tabulation of the most important studies. Furthermore, the paper concludes with some suggestions for future research regarding how menus and their positioning on the screen could be improved. The two principal suggestions concern trying to use more qualitative methods for investigating the issues and to develop in the future more universally designed menus.**

***Keywords*—Menus; navigation of interfaces; universal design; research methods**

I. Introduction

The research and practical problem of how to design navigation methods or menus for web sites and applications has been around for some time. Many have opinions regarding which type of menu is best for users and this is reflected in the web sites we can see on the internet. However, most web sites seem to employ either a left or right vertically positioned menu or a horizontal top positioned menu, with some web sites sometimes opting for a combination of two or more styles of menu (e.g. see [1]). Many web sites also use the bottom horizontal part of a screen to position a form of menu (e.g. see [2]). Some are also more unique, by placing a horizontal menu in the middle of the screen (e.g. see [3]).

Despite the opinions and numerous studies around this subject, there are still unanswered questions regarding which menu position or design might be optimal in terms of performance and user preference.

Therefore, this paper is a review paper of the most relevant research conducted around evaluating menu types and their positioning on the screen. It is hoped that other researchers can benefit from this work, because it helps to bring together in one place a number of sources from diverse authors and publications that can be difficult to find in general searches. In addition, this review will focus on looking into the main issues of this field and to find any unexplored aspects in the field. The authors will also recommend some potential ways forward. This review paper should be of benefit to researchers working in this area, students of computer science and any professional designer or developer involved in menu design and implementation.

Furthermore, the authors acknowledge that there are in existence certain types of menus that are less conventional and less used in every day applications and web sites, e.g. radial and flower menus. Please see Rubio and Janecek [4], Samp and Decker [5], Bailly, Lecolinet and Nigay [6] and Murano and Khan [7] for some examples. Discussions of these will be deliberately not included in this review paper, as the authors wish to focus on menus which are more commonly used and hopefully can contribute some ways forward to the existing patterns of menu use. Clearly, if radial and flower menus should become more main stream, then a future review should deal with this category of menu too.

In section II, this paper will proceed by presenting the authors' salient selection of relevant papers. This will be followed by a summary table of the most important papers covered so as to allow a more 'at a glance' option for quick reference. In section III, some ways forward are proposed for further investigation into this subject. Lastly, in section IV, the overall conclusions are presented.

II. Review of the Most Relevant Works

There have been studies investigating all manner of aspects to do with navigation or presentation of information.

Pittsley and Memmott [8] investigated issues of users not noticing certain navigation cues on web pages designed for information retrieval in a US university library web site for research purposes.

They tested some changes to the user interface. The redesign used in the evaluation involved a tabbed horizontal menu and left vertical context menu with larger font (both with the same labels) and a larger tabbed horizontal menu only. They also had a 'Comparison group' which was not fully described in the paper. They collected monthly usage on each information guide, page and secondary pages as a guide to show if the redesigns had been successful. Their basic findings were that both prototype 'menus' showed an increase in secondary page hits compared to the 'comparison group'.

In another study by Melguizo, Vidya and van Oostendorp [9], menu types, the complexity of a navigation path and the users' spatial skills in relation to finding web information were investigated. They also looked at 'web disorientation'.

There were two types of menu studied, left vertical expandable and left vertical sequential. They measured task

accuracy, task response time and lostness. The primary results of their work were that there was no significant difference for menu type and task accuracy, task response time and lostness. However, they did see some differences in relation to users with high or low spatial ability. Their results suggest that users with high spatial abilities perform well in all their conditions. Users with low spatial abilities seem to generally perform better with expandable menus.

A different study by Patsula, Detenber and Theng [10] involved web menus, with the emphasis on investigating the 'structure processing mechanism' in human working memory.

In the first study they compared a structured rule-based menu, a structured semantic-based menu, an unstructured mismapped menu and an unstructured random design. They recorded retention, time, errors and subjective opinions. The actual context of the menus involved abstract content consisting of arbitrary words and characters. Based on a statistical analysis and interviews their results showed retention, time and errors were better with the structured menus. Further, almost half of all participants felt the unstructured menus needed more mental effort.

In the second study they compared a more realistic context using structured menus, based on the PhotoImpression 3.0 application and unstructured menus, based on the Norton Anti Virus 4.0 application. They recorded retention, time, errors and subjective opinions. Based on a statistical analysis, a questionnaire and interviews their results showed retention, time and errors were better with the structured menus. Also, participants seemed to engage in behaviour involving structure processing.

Another study to be considered is by Leuthold, Schmutz, Bargas-Avila, Tuch, and Opwis [11]. In this investigation they tested three types of vertical menu placed on the left side of a page. These were a simple menu, extended menu and a dynamic menu. Their context was an ecommerce web site. The principal data was recorded in terms of number of eye fixations, the time to do the first click, the correctness of the first click, the navigation strategy and subjective opinions.

The central results they achieved were that first clicks were more successful for complex and simple tasks with the extended menu. For simple tasks, participants used fewer eye fixations with the extended menu. For complex tasks participants used fewer eye fixations with the extended menu. For simple tasks participants were faster with the simple menu. For complex tasks participants were faster with the extended menu. Lastly, the extended menu was perceived by participants to be easier and more helpful than the other two menus.

One of the authors of this paper has been investigating menu design for a few years and in Murano and Oenga [12] the details of an experiment were described. A left vertical menu and a fisheye menu placed horizontally at the top of a page were studied in the context of a simulated supermarket web site and compared with a real supermarket web site which had a horizontal menu at the top of the screen. No major differences in the two designs were observed. One of the reasons for not observing any differences in the data collected could have been due to the tasks being too easy for participants. However it could also have been some confounding variable.

This resulted in another experiment conducted by Murano and Lomas [13] where the tasks were designed to be more demanding than in the study by Murano and Oenga [12]. An experiment with four conditions was executed, which had the context of a simulated web shop. The four conditions examined four different menu positions (left, right, top and bottom of the screen). Overall, the top and left placed menus elicited fewer errors and less mouse clicks in users. Also users' preferences generally were in the same direction with the results for errors and mouse clicks.

In another study by Bernard, Hamblin and Chaparro [14] different menu layouts were compared. The first layout they used, was an 'index menu'. This had menu options as links in the centre of the screen. The second was a 'horizontal menu', which was at the top of the screen. The third was a 'vertical menu', which appeared at the left side of the screen. They observed that the 'index menu' performed best and was also preferred by the users. They also observed that the 'horizontal menu' was the worst in terms of performance and preference.

In a study by Burrell and Sodan [15], the position of menus on the screen was studied. They studied six different menu positions. In relation to a screen, these were: top tabbed, top, left, top and bottom, top and left and top and right of the screen. Some web site context was used in their study and their data showed them that the tabbed menu was liked more by the users.

Moreover, McCarthy, Sasse and Riegelsberger [16] looked at menu positioning in the context of a web site. They had a complex and simple version of the web site. They then had three menu positions. These were left, top and right of the screen. The time taken for a task was significantly longer with the complex web site. However the menu positions described above were not statistically significant in difference when averaged across the tasks and the simple and complex web sites. The left sided menu performed better concerning interaction with the first web page. No differences for the second page were recorded.

Also Fang and Holsapple [17] did an interesting study looking at navigation structures for web sites. They specifically looked at three types of hierarchy. These were: 'subject-oriented, usage-oriented' and a combination of the first two. Further, they had some simple and complex tasks. Fang and Holsapple observed that the 'usage-oriented' and the combined form of hierarchy indicated increased usability when compared with the 'subject-oriented' hierarchy.

Furthermore, Yu and Roh [18] studied different menus. They studied 'a simple selection menu, a global and local navigation menu and a pull-down menu'. These were tested via an ecommerce web site. Participants were then used to perform some tasks in the context of information finding about products. The principal results indicated that the pull-down menu was faster for searching. However, the performance in browsing task speed was stronger in the global and local navigation menus. The users preferences were approximately the same in the three experimental conditions.

In a more industrial and realistic context, Kalbach and Bosenick [19] did an evaluation for the Audi Cars web site. In this context, they evaluated linear menus on the left and right sides of the screen, using the Audi web site. They observed no significant differences for task times and the two menu styles they tested.

In a study by Faulkner and Hayton [20], they evaluated left and right placed menus. This was in the context of an ecommerce web site, selling Christmas products. For tasks involving purchasing something, there were no significant differences in the times taken.

The review above has shown that there are many different types of results, which have been achieved with many different forms of menu or navigation. In order to help visualise and see what has been done more at a glance, the table below tabulates the above endeavours. Only what the authors have considered to be pertinent to this paper have been included in the table and it is suggested that if a particular work seems interesting, then the full paper should be accessed. Furthermore, in the table, use of the word 'statistical' is made. The authors here are using the word loosely, because different authors of the research cited have used different 'statistical' techniques. In some cases, they tend to be more about presenting high level 'statistics', while on certain other occasions more in-depth significance testing has been done. The same applies to the word 'experimental' used in the table. This is used loosely, because across all the studies shown in the table, the fundamental characteristics of what was done point to an 'experimental' type method. However, the degree of rigour and robustness of how the studies were conducted varies within this set of papers. Some experiments were very rigorous and some less so.

TABLE I. TABULATED SUMMARY OF THE MOST IMPORTANT RESEARCH ON MENU POSITIONING AND MENU TYPES

Authors	What Was Evaluated			Method	Type of Analysis	Significant Results
	Menu Position/Type	Dependent Measures	Context			
Faulkner, Hayton, 2011, [20]	Left, Right.	Time.	Ecommerce web site.	Experimental	Statistical	None.
Kalbach, Bosenick, 2003, [19]	Left, Right.	Time, Subjective Opinion.	Ecommerce web. site	Experimental	Statistical, Interviews.	None.
Yu, Roh, 2002, [18]	Top, Combination Top/Left.	Time (searching and browsing), Subjective Opinion.	Ecommerce web site.	Experimental	Statistical	Pull down top menu faster for searching. For browsing simple selection menu slowest. Similar to left vertical, but each sub-menu overlays the previous menu and not full left justified. Not significant for subjective opinion.
Fang, Holsapple, 2007, [17]	Hypertext links for navigation varied in structure: subject oriented, usage oriented and combination of above two.	Navigation time, Correct answers to a test, Subjective opinions.	Information web site.	Experimental	Statistical	Usage oriented structure or combined structure more useable for simple and complex tasks.
McCarthy, Sasse, Riegelsberger, 2003, [16]	Left, Top, Right.	Eye movements, Finding the target on navigation menu, Site complexity, Time.	Internet service provider web site.	Experimental	Statistical	Complex site took longer for searching. Left menu significantly performed better for first page visit. User glances were mostly towards the central part of the screen. Menu position not significant.
Burrell, Sodan, 2006, [15]	Top tabbed, Top, Left, Top and bottom, Top and left, Top and right.	Subjective opinions.	Information web site.	Experimental	Statistical	Tabbed top menu preferred.
Bernard, Hamblin, Chaparro, 2003, [14]	Top, Left, Index (links in the centre of page).	Time, Subjective opinions.	Ecommerce web site.	Experimental	Statistical	Index faster. Some evidence for index menu being preferred by users.
Murano, Lomas, 2015, [13]	Top, Left, Right, Bottom.	Time (task), Errors, Mouse clicks, Subjective opinions.	Ecommerce web site.	Experimental	Statistical	Least errors - in order: Top, left. No difference between top and right menus. Least mouse clicks – in order: Top, left.
Murano, Oenga, 2012, [12]	Left vertical, Top fisheye, Real supermarket top horizontal.	Time, Errors, Overall success, Subjective opinions.	Ecommerce web site.	Experimental	Statistical	Errors more with left vertical and top fisheye menu for 1 task out of 5. Some significance to show real supermarket top horizontal menu was preferred.

Leuthold, Schmutz, Bargas-Avila, Tuch, Opwis, 2011, [11]	Left vertical (three types: Simple menu, Extended menu, Dynamic menu).	Number of eye fixations, Time to do first click, Correctness of the first click, Navigations strategy, Subjective opinions .	Ecommerce web site.	Experimental	Statistical	First clicks were more successful for complex and simple tasks with the Extended menu. For simple tasks participants used fewer eye fixations with the extended menu. For complex tasks participants used fewer eye fixations with the extended menu. For simple tasks participants were faster with the simple menu. For complex tasks participants were faster with the extended menu. The extended menu was perceived by participants to be easier and more helpful than the other two menus.
Patsula, Detenber, Theng, 2010, [10]	Study 1: Structured rule-based, Structured semantic-based, Unstructured mismapped, Unstructured random design. Study 2: Structured, based on PhotoImpression 3.0 and Unstructured, based on Norton Anti Virus 4.0.	Study 1: Retention, time, errors, Subjective opinions. Study 2: Retention, time, errors, Subjective opinions.	Study 1: Web menus with abstract content consisting of arbitrary words and characters. Study 2: Menus based on PhotoImpression 3.0 and Norton Anti Virus 4.0.	Experimental	Study 1: Statistical, Interviews. Study 2: Statistical, Questionnaire, Interviews.	Study 1: Retention, time and errors were better with the structured menus. Almost half of all participants felt the unstructured menus needed more mental effort. Study 2: Retention, time and errors were better with the structured menus. Participants seemed to engage in behaviour involving structure processing.
Melguizo, Vidya, van Oostendorp, 2012, [9]	Left vertical expandable, Left vertical sequential.	Task accuracy, Task response time, Lostness.	Information web site.	Experimental	Statistical	No significance for menu type and task accuracy, task response time and lostness.
Pittsley, Memmot, 2012, [8]	Tabbed horizontal menu and left vertical context menu with a larger font (both with the same labels). Larger tabbed horizontal menu only. 'Comparison group' (not fully described in the paper).	Monthly usage on each information guide, page and secondary pages.	Information web site.	Experimental	Statistical, Usage data	Both prototype 'menus' showed an increase in secondary page hits compared to the 'comparison group'.

III. Ways Forward for Investigating Menu Design

An examination of these studies shows that there is no totally clear picture to suggest which menu type or position on the screen may be optimal in terms of performance, e.g. fewer errors and less clicks etc. There is also no totally clear picture regarding which type of menu or position may be preferred by the majority of users. These statements regarding a totally clear picture are important because designers and developers tend to like to have clear and unambiguous guidelines when developing elements of a user interface.

Some aspects that loosely can be linked across the studies are that the strongest suggestion of better performance comes from menus which are as straightforward as possible and well structured. Also in the study by Murano and Lomas [13] the menu placed at the bottom of the screen horizontally performed worst with users (finding not included in the table above).

In some cases, the evidence suggests a 'leaning' towards a top horizontal menu being the best performer.

Most studies used some sort of web site in their evaluations and most studies used exclusively some type of statistical analysis of the data. A few studies used interviews and other more qualitative methods. Interviews were the most frequently occurring qualitative method used. One study also used usage data in their analysis. Further, most studies investigated some sort of performance in terms of task times and errors and most recorded some kind of subjective opinions.

Therefore it is clear that further studies are needed and the studies presented above document some indications. However in the authors' opinions indications are often arguable.

One reason for this rather unclear picture could be to do with the fact that most users these days have become so used to seeing and interacting with different types of menus and in different positions on the screen, that it reduces significant effects of performance and even preferences. One way to test this statement would be to have complete novices to menus and their screen positioning taking part in extensive evaluations. However in practice obtaining such a sample of users is challenging as these days even children are being exposed to computer technology from birth onwards.

A further observation regarding these studies is that they all applied an 'experimental' type method to data collection. While this is an accepted and very good method and the authors do not seek to criticise it, it is the authors' idea that since the results are quite varied and do not show any overall patterns that could help designers and developers, perhaps some studies using more qualitative techniques could shed further light on the matter. Some possible approaches could be to use more ethnographic, interview or case study approaches.

A further way forward that could be looked at concerning all menu types and their positioning on the screen would be to try and apply more universal design [21] principles. The first aspect we will not dwell on here but needs mentioning is for developers to ensure that their implementation follows appropriate guidelines to help achieve aspects of universal design and accessibility (e.g. WCAG [22] etc.). The second aspect is to meet the challenge of navigation in a way that makes things as easy as possible (e.g. Apple Inc [23] has guidelines for menu design to aid simplicity) for as many users as possible. These first two aspects the authors consider to be works in progress, because many are trying to achieve what we suggest. However, a third aspect to consider in the realm of catering for as many diverse users as possible could be to develop navigation systems (menus and positioning) that allow for more tailoring by the end user. Some examples include: being able to easily enlarge a menu, change the colour and shading of menus to allow for different types of vision and allow users to be able to move a menu to some other part of the screen with the content of the application or web page adjusting itself automatically.

While some of these ideas would require more implementation effort (until more implementation libraries are available), it would nevertheless aid the goal of user interfaces being as universally designed as possible. The authors also suggest that while user-tailoring can be a good thing as suggested above, it can also be implemented in a way that could actually confuse users more. This would require designers and developers to strike a balance between user-tailoring and a greater learning curve and/or the need to use a lot more time to tailor options.

IV. Conclusions

This review paper has been written with the purpose of bringing together a body of research concerning menus and their positioning on the screen. Another purpose is to allow researchers, designers and developers a quick at a glance view of the most important work done in this field. Furthermore, this paper contributes by suggesting some ways forward for this research area.

Overall, the authors would like to see if there is some clear menu type that is better and preferred by users. However if this does not materialize, the authors feel that some improvements in universal design could go a long way to improve current menus and navigation.

References

[1] Statoil, http://www.statoil.no/no_NO/pg1334082177637/extra.html, 2016, Accessed March 2016.

[2] Oslo and Akershus University College of Applied Sciences, http://www.hioa.no, 2016, Accessed March 2016.

[3] Luleå University of Technology, http://www.ltu.se/?l=en, 2016, Accessed March 2016.

[4] J. M. Rubio and P. Janecek, "Floating pie menus: enhancing the functionality of contextual tools", UIST '02 - Adjunct Proceedings of the 15th annual ACM Symposium on User Interface Software and Technology, 2002.

[5] K. Samp and S. Decker, "Supporting menu design with radial layouts", Proceedings of the International Conference on Advanced Visual Interfaces, ACM, 2010.

[6] G. Bailly, E. Lecolinet and L. Nigay, "Flower menus: a new type of marking menu with large menu breadth, within groups and efficient expert mode memorization", Proceedings of the working conference on Advanced Visual Interfaces, AVI 2008, Napoli, Italy, May 28-30, 2008.

[7] P. Murano and I. N. Khan, "Pie menus or linear menus, which is better?", Journal of Emerging Trends in Computing and Information Sciences, Vol. 6, Issue 9, September 2015.

[8] K.A. Pittsley and S. Memmott, "Improving independent student navigation of complex educational web sites: an analysis of two

navigation design changes in LibGuides", *Information technology and libraries*, 31*(3)*, pp. 52 – 64, 2012.

[9] M.C.P. Melguizo, U. Vidya and H. van Oostendorp, "Seeking information online: the influence of menu type, navigation path complexity and spatial ability on information gathering tasks", *Behaviour & Information Technology*, 31*(1)*, pp. 59 – 70, 2012.

[10] P.J. Patsula, B.H. Detenber and Y.L. Theng, "Structure processing of web-based menus", *International Journal of human- computer interaction*, 26*(7)*, pp. 675 – 702, 2010.

[11] S. Leuthold, P. Schmutz, J.A. Bargas-Avila, A.N. Tuch and K. Opwis, "Vertical versus dynamic menus on the world wide web: eye tracking study measuring the influence of menu design and task complexity on user performance and subjective preference", *Computers in Human Behavior*, 27, pp. 459 – 472, 2011.

[12] P. Murano and K.K. Oenga, "The impact on effectiveness and user satisfaction of menu positioning on web pages", International Journal of Advanced Computer Science and Applications, 3: 9, 2012.

[13] P. Murano and T.J. Lomas, "Menu positioning on web pages. does it matter?", International Journal of Advanced Computer Science and Applications, Vol. 6, Issue 4, April 2015.

[14] M. L. Bernard, C.J. Hamblin and B.S. Chaparro, "Comparing cascading and indexed menu designs for differences in performance and preference", Proceedings of the Human Factors and Ergonomics Society 47th Annual Meeting – 2003.

[15] A. Burrell and A.C. Sodan, "Web interface navigation design: which style of navigation-link menus do users prefer?", Proceedings of the 22nd International Conference on Data Engineering Workshops, ICDEW 2006, 3-7 April 2006, Atlanta, GA, USA, IEEE Computer Society, 2006.

[16] J.D. McCarthy, M. A. Sasse, and J. Riegelsberger, "Could i have the menu please? An eye tracking study of design conventions", people and Computers XVII — Designing for Society, pp 401-414, 2004.

[17] X. Fang and C.W. Holsapple, "An empirical study of web site navigation structures' impacts on web site usability", Decision Support Systems, 43: 2, P.476-491, 2007.

[18] B. Yu and S. Roh, "The effects of menu design on information-seeking performance and user's attitude on the world wide web", Journal of the American Society for Information Science and Technology, 53: 11, P.923–933, 2002.

[19] J. Kalbach and T. Bosenick, "Web page layout: a comparison between left and right justified site navigation menus", Journal of Digital Information, Vol 4, No 1, 2003.

[20] [20] X. Faulkner and C. Hayton, "When left might not be right", Journal of Usability Studies, Vol 6, Issue 4, P. 245-256, 2011.

[21] [21] M.F. Story, "Maximizing usability: the principles of universal design", Assistive Technology: The Official Journal of RESNA, 10:1, 4-12, 1998.

[22] [22] WCAG (2016) https://www.w3.org/WAI/intro/wcag, Accessed March 2016.

[23] [23] Apple Inc. (2016) OS X Human Interface Guidelines, https://developer.apple.com/library/mac/documentation/UserExperience/Conceptual/OSXHIGuidelines/MenuAppearanceBehavior.html#//apple_ref/doc/uid/20000957-CH23-SW1, Accessed March 2016.

Edge Detection with Neuro-Fuzzy Approach in Digital Synthesis Images

Fatma ZRIBI
Ecole Nationale d'Ingénieurs de Tunis (ENIT)
Tunis, TUNISIA

Noureddine ELLOUZE
Ecole Nationale d'Ingénieurs de Tunis (ENIT)
Tunis, TUNISIA

***Abstract*—This paper presents an enhanced Neuro-Fuzzy (NF) Approach of edge detection with an analysis of the characteristic of the method. The specificity of our method is an enhancement of the learning database of the diagonal edges compared to the original learning database. The original inspired NF edge detection model uses just one image learning database realized by Emin Yuksel. The tests are accomplished in synthesis images with a noised one of 20% of Gaussian noise.**

***Keywords*—Neuro-Fuzzy; learning databases; Gaussian noise; synthesis images**

I. INTRODUCTION

On the basis Systems Artificial Intelligence Neuro Fuzzy Neural networks are forms with the benefits of parallel processing and the learning capacity inspired from the nervous system and fuzzy logic that allows the possibility of modeling language and Cognitive systems of human decision. [1]

The systems based on neural networks and fuzzy logic are Good approximations function from sample data. They do not Need mathematical model. They are pure estimators

But they bring something more compared to the statistical Approach : they do not require a priori knowledge about the inner workings of the system.

The main feature of neural and fuzzy systems is adaptive control and statistics. They are digital estimators and dynamic systems. The neural theory drowns in the mathematical semantics of dynamic systems. Fuzzy theory overlaps with these semantic and more with probability, mathematical logic, and theories of measurement.

In general neuronal systems and fuzzy one are used to improve the performance of real systems. [2]

Neural Networks and Fuzzy Logic; both approaches belong to the large class called structures of nonlinear systems that has some properties of extreme importance on knowledge a priori of a control system. Multiple nodes interconnections are achieved and performed a learning phase by adapting the weights of nodes interconnections [2]. This process is called "adaptation" of weight. Modify the weight changes data stored in the network, in fact, the weight values represent the "memory" of the network. Neural Networks (NN) Are special models of nonlinear systems modeling of the biological nervous system. Fuzzy Logic (FL) is a system modeled by language skills and human reasoning.

In this paper, we present related works in the second part of our paper. The third part we detail the architecture of our Neuro-Fuzzy edge detection system. The next step interpreter the result obtained for noisy images to detect edges. We finish this paper with a discussion and a conclusion.

II. RELATED WORKS

Neuro-Fuzzy edge detection was used widely in this last decades. In [6] authors propose an ANFIS model with 8-inputs and 1-output Neuro-Fuzzy system based in first-order-Sugano system. 2 triangular type membership functions are used for each input, and the output has a constant membership function. 256 rules were used with just one output. Authors use a Grid partition method on the contrary of subtractive clustering method. The Gray level image is used to detect edges, they firstly binarize the image, and then the binary image disintegrates to 3x3 windows and generates a set of the image pattern. The edge patterns in binary images were classified into 32 categories. Training the ANFIS on this category (patterns) classify the blank elements in each 3x3window in white pixels (value: 1) and dark pixels (value: 0).

In [7] authors classify image pixels into three sets of pixels contour, regular and texture using a model of Neuro-Fuzzy approach which takes the advantages of Neural Network's Learning characteristic and the fuzzy logic function. Spatial properties of the image features are the base of this approach. As every Neuro-Fuzzy system, a training set was used to create and train the classifier system. They assigned for every pixel a degree of membership for each of the three fuzzy subsets. They note that « this approach would be highly attractive in the biomedical field »

In [8] authors uses an ANFIS(Adaptative Neuro-Fuzzy Inference System) detector and « is a first-order Sugeno type Fuzzy inference system with 4-inputs and 1-output. Each input has three generalized bell type membership functions and the output has a linear membership function ». for the training they use Sugano first order model which characteristic is the combination of applying a mixature of the least-squares method and the backpropagation gradient descent method. This combination is to realize a given training data set.

Authors of the paper [9] use an ANFIS with 2x2 windows For edge detection. They allow four input values which are mapped to « the four fuzzy inputs in the fuzzy inference system ». the ANFIS system classify the output as an edge or as background.

In [10] authors use Neuro-Fuzzy system in the classification of biometric Multimodal Face, Iris, and Finger Fake they detect this characteristic with an ANFIS system. Another application of ANFIS is used in [11] to classify Texture Image. They compare Classification with Artificial Neural Networks and ANFIS one. Their results prove that's the ANFIS system gives better classification, and the methods could be applied to medical images or defense applications.

Neuro-Fuzzy system based in Takagi–Sugeno fuzzy inference system was used in [12] as a classifier for image retrieval. They use Canny detector for the learning database.

II. The Proposed Neuro-Fuzzy Edge Detection

A. Structure of the realised NF Edge Detector

Fig.1a shows the structure of Neuro- Fuzzy detector (NF) [3]. The detector consists of four NF networks operating as sub-detectors in the four directions: vertical, horizontal, right diagonal and left diagonal direction respectively. Each detector operates on a window of the size of 3x3 pixels given in Fig.1b. The NF sub-detector evaluates the relationship of the central pixel of the window with its neighborhood according to one of the four selected combinations of the contour direction. The topologies of the various selected combinations are shown in Fig.1c.

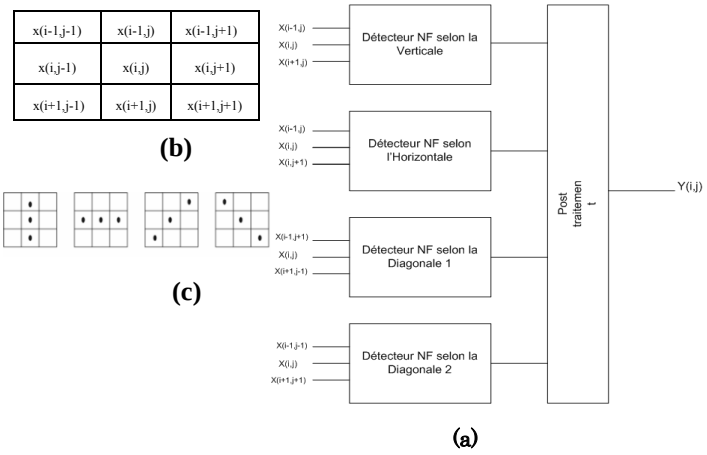


Fig. 1. (a) The structure of the neuro- fuzzy detector. (b) operator Filter window . (c) Different Topology orientations of selected contours

B. Architecture of Neuro- Fuzzy detector

A sub-NF-detector is a fuzzy inference system with 3-input and 1-Sugeno of the first order output Fig.1. The internal structures of each sub-detector are identical. Each entry has three functions of generalized Bell type of membership function. The input-output relationship of each sub-detector is governed by a base rule defined as:

X1, X2, and X3 are the inputs of the detector and Y is the corresponding output. The basic rule represents differently possible cases combinations between the three functions. Each function belong to each entry. It corresponds to a basic rule of 27 data laws as follows:

if (X1 is M11) and (M21 X2) and (X3 M31) then R1 = F1 (X1, X2, X3)

if (X1 is M11) and (M21 X2) and (X3 M32) then R2 = F2 (X1, X2, X3)

if (X1 is M11) and (M21 X2) and (X3 M33) then R3 = F3 (X1, X2, X3)

if (X1 is M11) and (M22 X2) and (X3 M31) then R4 = F4 (X1, X2, X3)

if (X1 is M11) and (M22 X2) and (X3 M32) then R5= F5 (X1, X2, X3)

if (X1 is M13) and (M23 X2) and (X3 M33) then R27 = F27 (X1, X2, X3)

Where: Mij is the j-th membership function of the ith entry, given by the generalized bell-type function (equation 1 and Fig.2).

$$M_{ij} = \frac{1}{1+\left|\frac{X_i - a_{ij}}{b_{ij}}\right|^{2c_{ij}}} \quad (1)$$

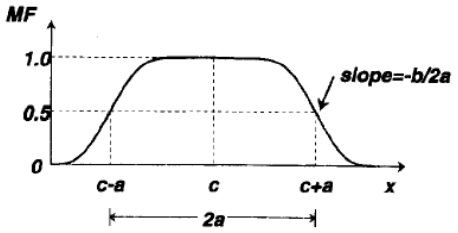


Fig. 2. Physics meaning of the parameter of The generalized Bell Function [4]

and Rk is the kth output of the rules and Fk is the kth output of the membership function, given by:

$$F_k(u_1,u_2,u_3) = d_{k0} + d_{k1}x_1 + d_{k2}x_2 + d_{k3}x_3 \quad (2)$$

Where: i=1,2,3 ; j=1,2,3 and k=1,……,27

The parameters a, b, c and d are constants that characterize the shape of the different functions of the membership. The optimum values of these constants are determined by the learning phase which will be detailed in paragraph *D.*

The output of the NF detector is the weighted average for the different weights of each output of different rules (Fig3). The weighting factor Wk. of each rule is calculated by evaluating the expression of belonging at the antecedent of the rule. This action is performed by the conversion, firstly, of the gray scale values of the input pixels in the fuzzy membership values through an initialization of the input membership Function, then the application of operator "and" to [1] Membership values. The operator "and" corresponds to a multiplication of different membership values of the entries pixels.

The weighting factors were calculated as follows:

$$w_1 = M_{11}(X_1).M_{21}(X_2).M_{31}(X_3)$$
$$w_2 = M_{11}(X_1).M_{21}(X_2).M_{32}(X_3)$$
$$w_3 = M_{11}(X_1).M_{21}(X_2).M_{33}(X_3)$$

$$w_{27} = M_{13}(X_1).M_{23}(X_2).M_{33}(X_3) \quad (3)$$

Once the weighting factors are calculated, the output of the sub NF detectors are calculated using the formula (4)

$$Y = \frac{\sum_{k=1}^{27} w_k R_k}{\sum_{k=1}^{27} w_k} \quad (4)$$

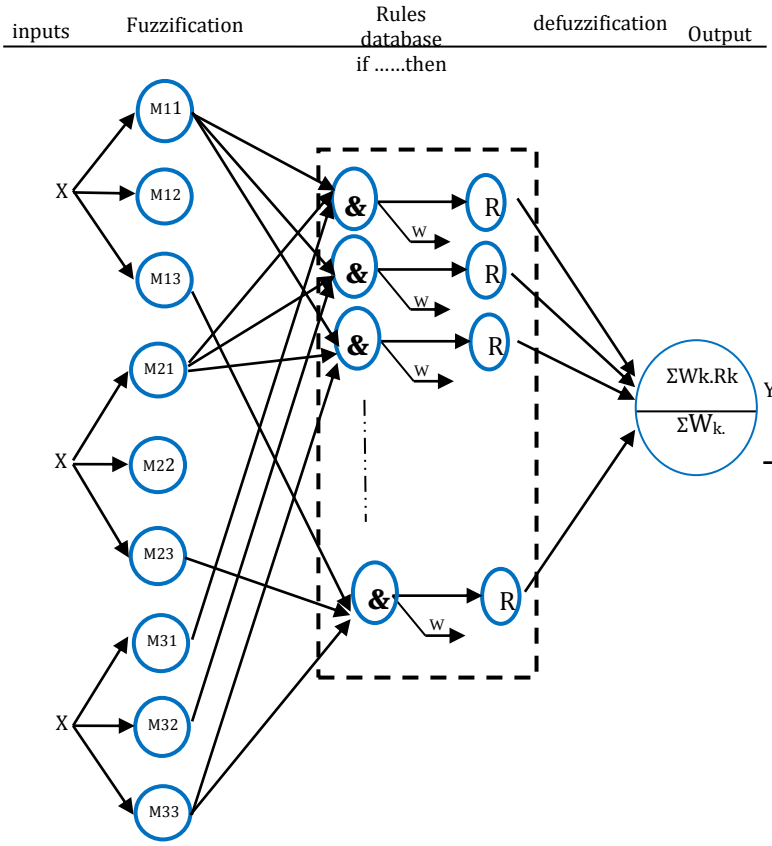


Fig. 3. The first order Sugeno model of 3 inputs and 1- output [5]

C. Post-processing: choice of edge pixel

This step involves a thresholding of the average of the Outputs of the various sub-detectors. An initial value of the Threshold 128 was used [3]. Following empirical tests the Threshold value selected from the results realized, and is 90. The input-output of the post-treatment phase relationship can be presented as follows:

Let the outputs respectively representing the four sub-detectors Forming the NF network. The final decision for a given pixel corresponds to the output of the post-processing Block (Figure 1a). This output was calculated in two steps. The first is the calculation of the average of the output of the various sub-detectors was given by equation 5,

$$Y_{Moy} = \frac{1}{K}\sum_{k=1}^{K} Y_k \quad (5)$$

the second step is thresholding performing the decision to set the belonging to the appropriate edge (Equation 6).

$$y(r,c) = \begin{cases} 0 & if \quad Y_{Moy} < 90, \\ 255 & else \end{cases} \quad (6)$$

$y(r,c)$ is the output of the post-processing block corresponding to the output of the NF detection operator.

D. The Learning of the various sub-detectors

The internal parameters of the NF detector were optimized by a learning phase which is carried out independently for each sub-detector of the overall structure (Figure 1a). The Figure 4 presents the diagrams of the learning process for a given NF-sub-detector.

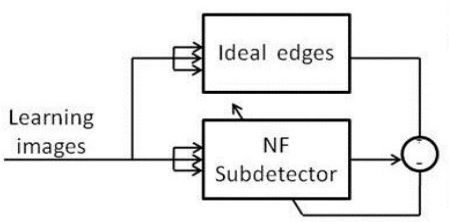
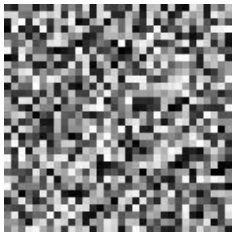
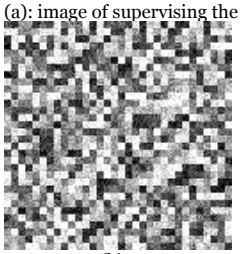


Fig. 4. The Learning an NF sub-detectors from [3]

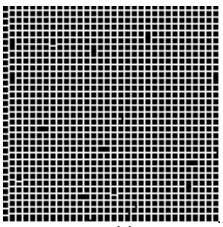
The selected learning database consists of two images: one for learning vertical and diagonal contours and the second for learning the edges located on the left and right diagonal. Here in the following statistical analysis of different existing patterns in the image of learning and degree of the occurrence of different patterns. Motifs circled in red in Figure 6 represent the most contours models based on the learning, their degree of occurrence percentages are given in Figure 7.



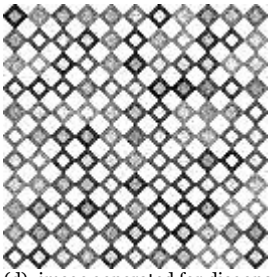
(a)
(a): image of supervising the horizontal and vertical Neuro-Fuzzy subdetectors



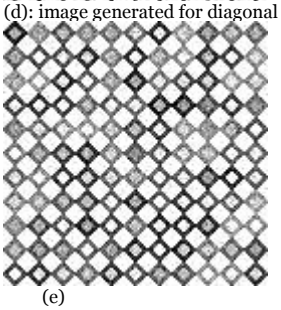
(b)



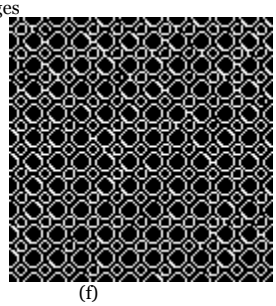
(c)
(b): image (a) blurred with Gaussian noise with 20% densities
(c): image of ideal edge (a) detected with the wavelet multi-scale product



(d): image generated for diagonal edges



(e)



(f)

(e) : Image (d) blurred with Gaussian noise with density of 0.02
(f): ideal edges detected with Sobel filter for the supervision of the diagonal subdetector

Fig. 5. images of learning databases: (a) and (d) Automatic images generated with grayscale randomly between 0 and 255. (b) and (e) the correspondent images of (a) and (d) corrupted with Gaussian noise of 0.02 intensity. (c) and (f) the edges correspondent to (a) and (d) used to supervise the Neuro-Fuzzy subdetectors: (b) and (c) for the vertical and horizontal subdetectors. (e) and (f) for the diagonal subdetectors

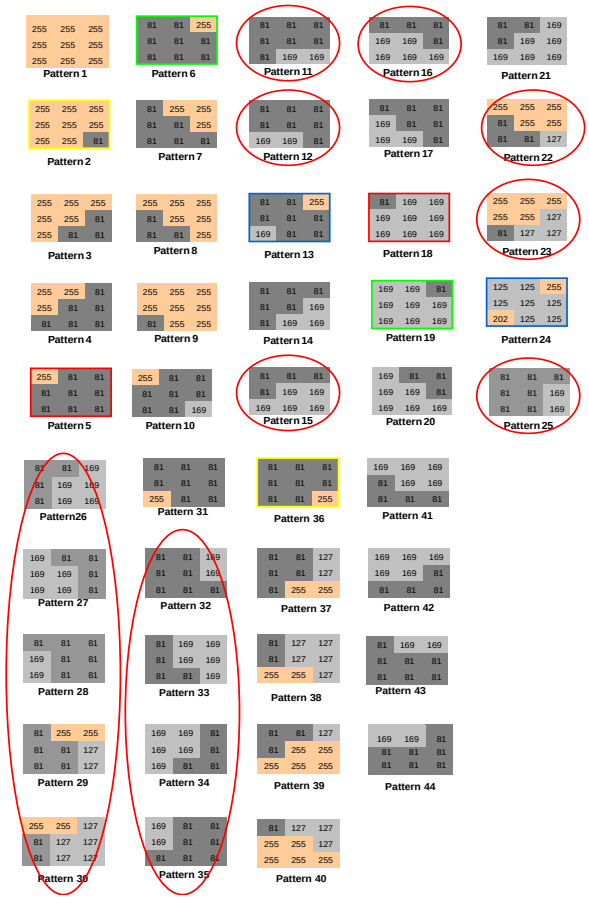


Fig. 6. Tthe most contours model based on the learning

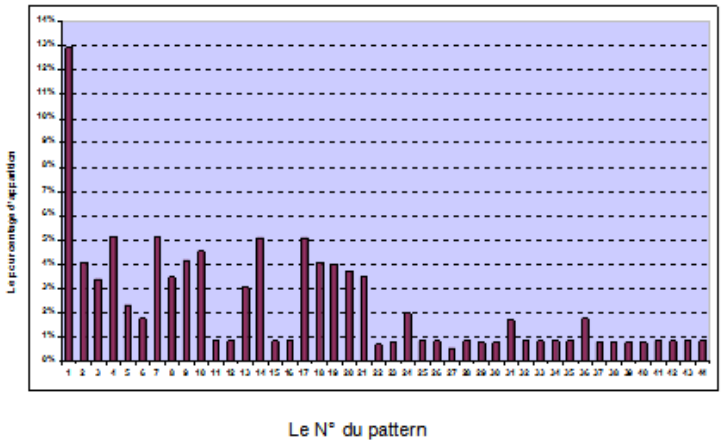


Fig. 7. The degree of occurrence percentages

Figure 6 and 7 present the degree of apparition of each pattern of our enhanced database. The pattern 1 represent the most pattern one which appears in the two databases horizontal and vertical one in one way and the two diagonal pattern of our the contribution of the database, the appear percentage is .13%. Patterns 4, 7,14 and 17 represent a rate of apparition of 5% Patterns number 14 and 17 has the ratio of apparition of 5% from the hole database vertical, horizontal and two diagonal. The patterns which have a rate of 1% are: 11,12,15,16,22,23,25,26,27,28,29, 30,32,33,34 and 35.

III. Results and Interpretation in Presence of Noise

The Neuro-Fuzzy detector was tested on synthetic images such as a circle, Lena, blobs, house and a cameramen (figure 8). All images are grayscale images coded on 8 - bit size 256x256 pixels. Tests were performed on the images without noise and the same figures noised by Gaussian noise of 14dB equivalent to 20% of the figure. The results of detections were compared with the detection of the same images by Sobel operator. The main feature of these detectors is revealed by the results. Indeed, the detection results on noisy images are clearly better than those given by the conventional Sobel detector.

IV. Discussion

The Neuro-Fuzzy edge detector in this paper is an example of edge detector belonging to a new family, which has the following advantages:

1) The preprocessing phase which was the subject of Several studies are no longer necessary with this detector. Just know the type of noise in the image and to learn the detector and thus, design the appropriate one.

2) Its structure is flexible. Indeed the number of sub sensor is variable depending on the accuracy in the direction of the Desired contour. Certainly, the number of sub-detector is important to the edges accuracy is fine, for cons, the running time and the computational complexity of the Algorithm is more important than the classical detector. Hence, a performance/precision compromise is needed and time depends on the empirical Results performed. Such is the

case of detection of ongoing work by the Neuro-Fuzzy approach with different numbers of detectors under (4.8 under detectors) on biomedical images: scanners and X-ray.

3) The structure of the sensor remains simple, since it consists of a Neuro-Fuzzy fundamental block of 3-inputs and 1-output.

4) The two bases are chosen of learning: is an image of the bibliography realized by Emin Yuksel and the second one proposed is our contribution for diagonal edges and it represent our contribution. These two images forming the learning base is easily generated on a machine and constitute The entire contour models possible on a 3x3 window. Analyzes and tests studying the different bases of learning will be the resultant of a paper being prepared.

V. CONCLUSION

We can conclude that this detector showed effectiveness are mainly for the treatment of noisy images while performing the actual detection. Many features of this sensor made so that it performs the filtering operation in parallel with the edge detection by the features of the Neuro-Fuzzy system. This characteristic was based on the property of the learning database which recognizes contours in spite of the presence of noise. Indeed, learning step is the main feature that makes learning The system settings that will recognize contours distorted by noise. These results improve that hybrid approach is better in an edge detection that conventional one especially in noisy images and it is possible to apply them in more complex images like medical ones or defense images.

REFERENCES

[1] F. L. Lewis, J. Campos and R. Selmic, "Neuro-Fuzzy Control of Industrial Systems with Actuator Nonlinearities (Frontiers in Applied Mathematics)", Society for Industrial and Applied Mathematics, April 2002.

[2] Y. Morere, "Les Réseaux Neuro-Flous", Thèse de doctorat, L.A.M.I.H, 17 may 2001.

[3] M.E Yüksel and E.Besdok, "A simple neuro-fuzzy impulse detector for efficient blur reduction of impulse noise removal operators for digital images", IEEE Transactions on fuzzy systems, Vol. 6,N.12, pp. 854-865, 2004.

[4] J. S Roger Jang and C. T. Sun, "Neuro-Fuzzy Modeling and Control", Proceedings of the IEEE, Vol. 83, N°. 3, March 1995.

[5] J.-S.R. Jang , C.-T Sun and E. Mizutani , "Neuro-fuzzy and soft computing.. a computational approach to learning and machine intelligence", Prentice-Hall, Upper Saddle River, NJ, (614 pages) 1997

[6] Suryakant and Renu Dhir , " Novel Adaptive Neuro-Fuzzy based Edge Detection Technique ", International Journal of Computer Applications (0975 – 8887) Volume 49– No.4, July 2012.

[7] R. J. Oweis and M. J. Sunna, " A COMBINED NEURO–FUZZY APPROACH FOR CLASSIFYING IMAGE PIXELS IN MEDICAL APPLICATIONS, Journal of ELECTRICAL ENGINEERING, VOL. 56, NO. 5-6, 2005, 146–150

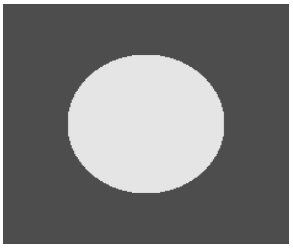
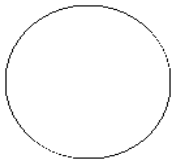
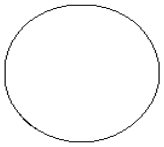
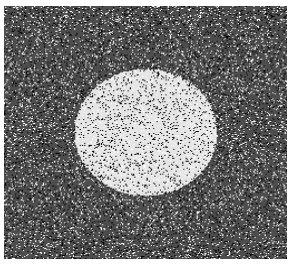
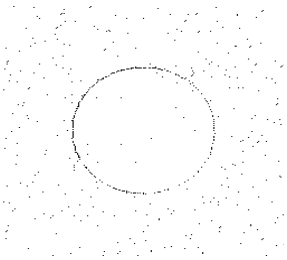
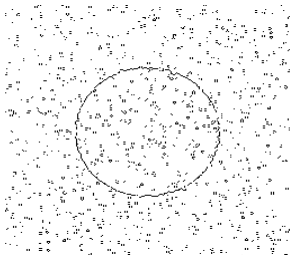
[8] Lei Zhanga Mei Xiaoa Jian Maa Hongxun Song, "A Novel Edge Detection Method Based on Adaptive Neuro-Fuzzy Inference System ", PIAGENG 2009: Intelligent Information, Control, and Communication Technology for Agricultural Engineering, edited by Honghua Tan, Qi Luo, Proc. of SPIE Vol. 7490, 74902S © 2009 SPIE · CCC code: 0277-786X/09/$18 · doi: 10.1117/12.836828, Proc. of SPIE Vol. 7490 74902S-1

[9] S. Anwar and S. Raj, "A Neural Network Approach to Edge Detection using Adaptive Neuro–Fuzzy Inference System", International Conference on Advances in Computing,Communicationsand Informatics (ICACCI), 2014.

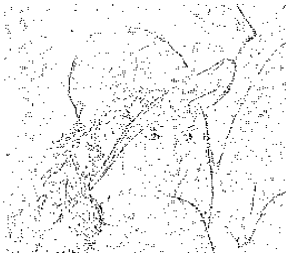
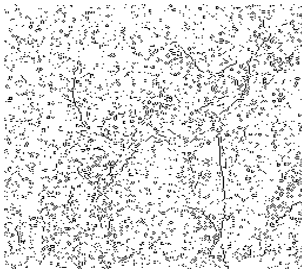
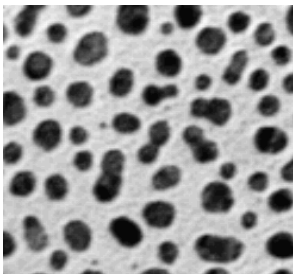
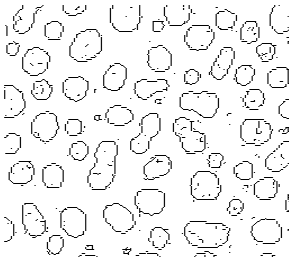
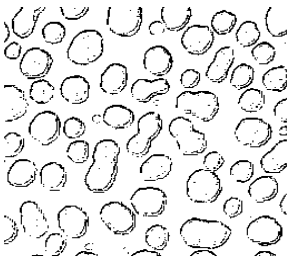
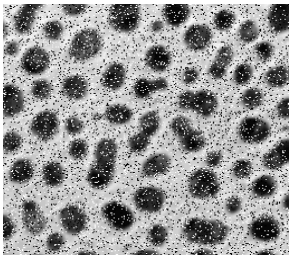
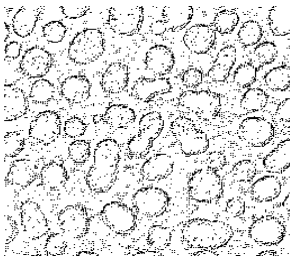
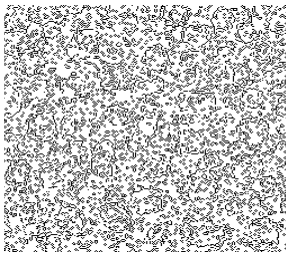
[10] S. Wilson and A. Lenin Fred, "An Efficient Biometric Multimodal Face, Iris and Finger Fake Detection using an Adaptive Neuro Fuzzy Inference System (ANFIS)", Middle-East Journal of Scientific Research 22 (6): 937-947, 2014.

[11] S. Panigrahi, T. Verma, "Texture Image Classification using Neurofuzzy Approach", International Journal Of Engineering And Computer Science ISSN:2319-7242 Volume 2 Issue 7, Page No. 2309-2313, July 2013.

[12] S.Asha, S.Ramya, M.Sarulatha, M.Prakasham, P.Priyanka," Neuro Fuzzy Classifier for Image Retrieval", International Journal of Innovative Research in Computer and Communication Engineering, Vol. 3, Issue 1, January 2015.



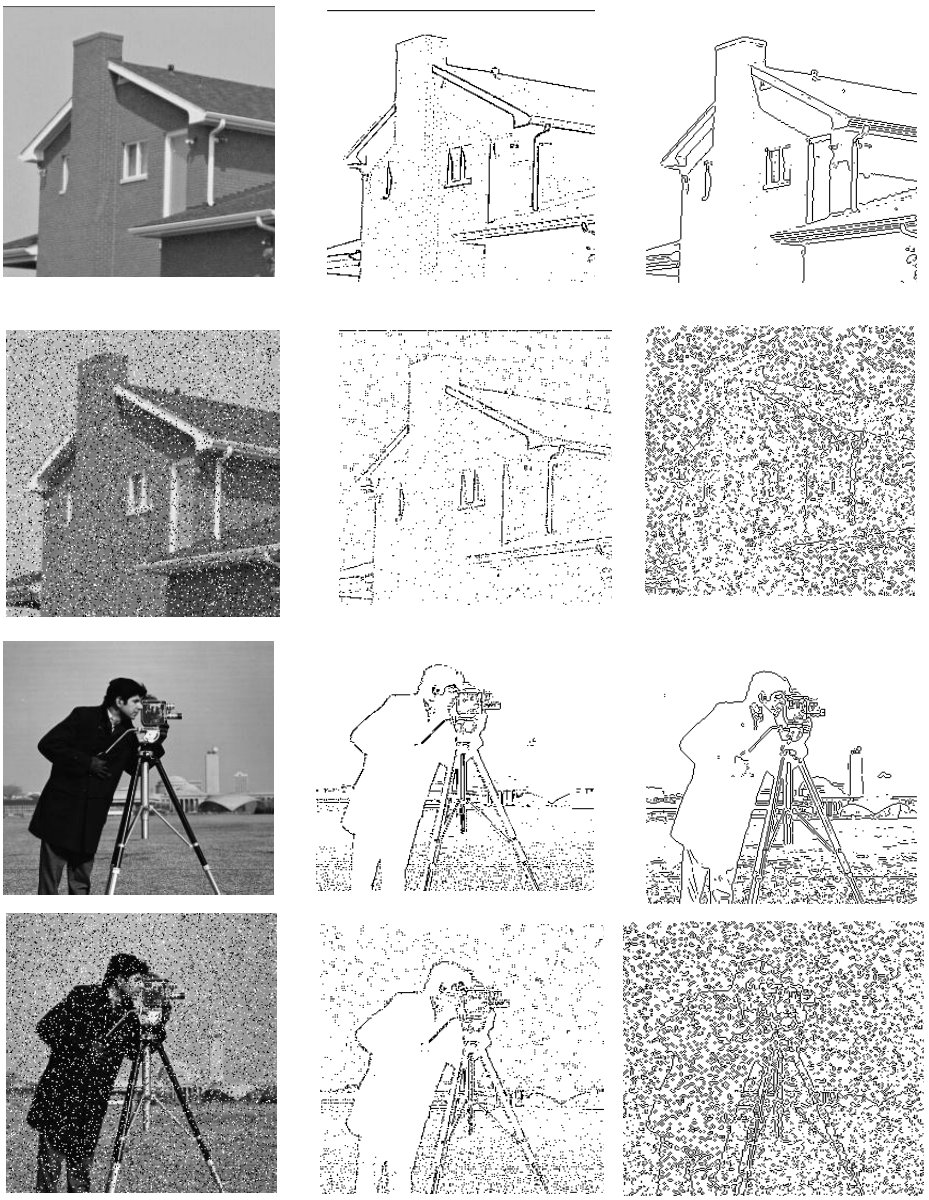


Fig. 8. Left : Original image tracking the same noisy image , medium: detection with the Neuro- Fuzzy detector , Right : Detection with Sobel detector

A Proposed Multi Images Visible Watermarking Technique

Ruba G. Al-Zamil
Department of Computer Information Systems
The Hashemite University
Zarqa, Jordan

Safa'a N. Al-Haj Saleh
Department of Software engineering
The Hashemite University
Zarqa, Jordan

***Abstract*—Visible watermarking techniques are proposed to secure digital data against unauthorized attacks. These techniques protect data from illegal access and use. In this work, a multi visible watermarking technique that allows embedding different types of markers into different types of background images has been proposed It also allows adding multiple markers on the same background image with different sizes, positions and opacity levels without any interference. The proposed technique improves the flexibility issues of visible watermarking and helps in increasing the security levels. A visible watermarking system is designed to implement the proposed technique. The system facilitates single and multiple watermarking as illustrated in the proposed technique. Experimental results indicate that the proposed technique applies visible watermarking successfully.**

Keywords*—Visible watermarking; multi-image watermarking; marker image; background image; image channels; opacity; Matlab

I. INTRODUCTION

Digital data distribution is widely spread due to the rapid growth of networks and internet communications in the recent years. The trend of moving from traditional libraries to digital libraries is growing day after day. The availability of digital data- such as images, sounds, videos, and e-books- may lead to slick illegal duplication or distribution; so it should be protected by contractual and intellectual property rights.

To regulate copyright issues and to prevent unauthorized use, the concept of digital watermarking has evolved over the past few years. It basically depends on embedding a digital signal with information that cannot be easily removed [12]. Generally, digital watermarking is the process of hiding any digital message –usually called marker or watermark- into a digital media. The embedded message may be either visible or invisible. When dealing with images, it is common to embed digital marker image or text into digital host image such as textual data about the author, or copyright image. The main goal of image watermarking is to protect confidentiality, integrity, availability and authenticity of images [15] taking into consideration that watermarking process should not affect the quality of original and watermarked image.

According to human perception, image watermarking techniques can be divided into three types [16]: Visible watermark, invisible watermark, and dual watermark. Visible watermarking is a semi-transparent text or image (i.e. the marker) that is overlaid on the original image where the watermark is visible to the viewer. The used watermark often contains name or logo of company or copyright information that identifies the owner of the original image. When adding a visible watermark, it is important to take into consideration that the marker should be difficult to remove; i.e. removing the watermark should be more costly than purchasing the original image from the owner. In invisible watermarking the watermark is added to the original image in a way that no modifications are noticed; i.e. the watermark is hidden in the image and can be retrieved by some mathematical calculations to prove ownership. Dual watermarking is the combination of visible and invisible watermark where invisible watermark is used as a backup for the visible watermark.

Some desirable properties should be applied on watermarking process to get the intended goals [8]. These properties include effectiveness (i.e. the watermark should be detective), quality of watermarked image, watermark size, and watermark robustness [15] against attacks.

Watermarking has many applications where the issues of security and ownership preservation are vital. It is applied in owner identification, copy and usage control [17] [4], broadcast monitoring, medical applications, fingerprinting, and military applications.

In this work, a visible watermarking technique for embedding different types of multi watermarks into background images is proposed.

The paper is organized as follows: Section 2 presents related work and previous proposed techniques. The proposed multi visible watermarking technique is discussed in section 3. Visible watermarking system that implements the proposed technique is presented in section 4. In section 5 the results are reported and appropriate discussion is made. Finally, the paper is closed with a conclusion in section 6.

II. RELATED WORK

Many techniques and methodologies were proposed for visible and invisible watermarking. In this section, we outline some visible watermarking techniques as mentioned in the literature. Mohanty et al. [14] presented a visible watermarking scheme that is applied to the original image in the DCT domain based on a developed mathematical model. They have also proposed a modification to increase the robustness of the watermark. Tamilselvi et al. [6] proposed a methodology based on DCT modification of original image with respect to watermark image. The watermarking is done in frequency domain using compound mapping function. The

results showed that the proposed watermarking process is secure against possible attacks. Parvathavarthini et al. [13] proposed a watermarking scheme using Hadamard transform based on a calculated scaling factor of the image. Hadamard transform was used due to its robustness against attacks. The value of scaling factor was controlled by a control parameter. The proposed scheme proved its efficiency by experimental results and performance analysis. Jose et al. [3] suggested a new approach for lossless visible watermarking using one to one compound mapping. Two types of applications are were proposed, opaque monochrome watermarks and non-uniformly translucent full color watermarks.

A new method was proposed by Bhaisare et al. [9] for visible watermarking for lossless image recovery capability based on one-to-one compound mappings. The methodology was designed to embed different types of visible watermarks into images. Majeed et al. [10] presented a large scale integration (VLSI) architecture for implementing two visible digital image watermarking schemes. The authors designed a watermarking chip that can be integrated within a digital camera. The integrated chip can have two different types of watermarking capabilities in spatial domain. Kumar et al. [7] presented a method based on the use of deterministic one-to-one compound mappings of image pixel values to watermark visible watermarks on original images. Park and Kim [5] proposed a watermarking technique using digital seal image (i.e. a binary image) as a marker. The authors showed how to construct a verifiable seal image and how to apply it to the original image. Biswas et al. [2] suggested a methodology for adding watermarks to medical images in Region of Non-Interest (RONI) to guarantee correct diagnosis. The selection of RONI was based on Fuzzy C-Means segmentation and Harris corner detection. Watermark embedding was based on alpha blending technique. Chen et al. [1] presented a wavelet based visible watermarking scheme that partitioned the original image to four similar images. The partitioned images were classified into two sets: fixed set (FS) and water marked set (WS). Wavelet transform was applied on the partitioned images for embedding watermark image.

Most of the presented related works do not present embedding different types of markers into different types of background images, and adding multi markers on the same background image.

III. The Proposed Multi Visible Watermarking Methodology

The proposed technique allows adding multiple markers on the same background image by selecting different positions for hiding the markers and preventing any interference between them. It allows hiding markers with different transparency degrees using different opacity levels. Repeating any marker with different size, position, and opacity is also allowed.

There is no restriction to use a specific type of images for marker and background images. Different types of images can be used (i.e. binary, gray, and RGB). Any combination of marker and background images is valid; for example RGB background may contain binary, gray, or RGB markers. On the other side, gray background may contain binary, gray, or RGB markers and so on.

Suppose that we tend to hide image 1(marker image) in image 2 (background image).The proposed technique is based on the following:

1) Store the state of images. The stored state indicates images types (i.e. binary, grayscale, or RGB).

2) Resize watermark image to suitable size smaller than the original image size.

3) The position where the marker image will be inserted is specified. This position represents the center point of the marker image. According to the center point, the start point (i.e. the top-left corner) and the end point (i.e. the bottom-right corner) are calculated. This is to check that the marker will not exceed the boundaries of the background image.

If the start or the end points are outside the background image, the insertion position will be modified to be within the boundaries of the background image.

4) Store number of channels for each image depending on its type; i.e. three channels for RGB image, and one channel for gray or binary image. This is because embedding marker image into background image is based on merging the corresponding channels.

5) Marker image is added with a specific level of transparency (i.e. opacity). Opacity values can be between 0-1. Transparency is usually not considered to be a blending mode. Transparency is just a combination of the multiply and addition blending modes; it takes a percentage of the foreground (i.e. marker image) and adds it to the complementary percentage of the background image. Thus, if we want the value of the foreground to be 75% opaque, multiply the foreground by 0.75 and the background by (1-0.75) and add the two values.

The proposed technique applies the following transparency equation (1) [11]:

$$K=n*A+ (1-n)*B \quad (1)$$

where K refers to the output image, n refers to opacity value, A refers to the marker image and B refers to the background image.

6) After applying this equation the final result (output image) will be converted to RGB image. This is to view and preserve all colors in the combined images (i.e. marker and background images). For example, suppose we have binary marker embedded into RGB image, if final image is stored as binary or gray, colors in RGB are not preserved. On the other side, suppose we have binary marker embedded into gray image, there is no problem to store final image as RGB as it will preserve colors in binary and gray images.

7) Finally, The width, the height, the size and the position of the marker image are stored. This helps in adding new markers without any conflict (i.e. markers interference). This will be explained in details in the next section.

IV. VISIBLE WATERMARKING SYSTEM

A visible watermarking system is designed using Matlab 7.8.0 (R2009a) to implement the proposed technique. The system facilitates loading marker and background images. It allows the user to choose marker position by specifying X and Y coordinates. The user can also determine the opacity and the size of the added marker image. Fig 1 shows the main screen of visible watermarking system.

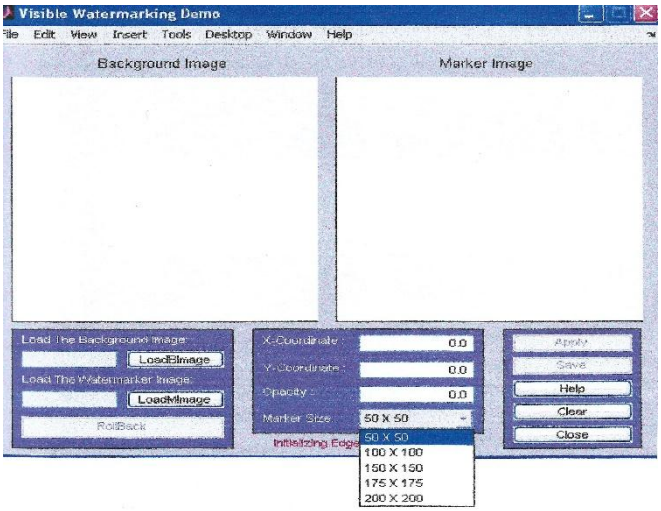


Fig. 1. Visible watermarking system

The system provides the following functionalities:

- Background and marker images selection and loading.

- Marker's position selection (i.e. center point). The user can determine the position using mouse. By this, the values of X and Y coordinates will be determined automatically. Coordinates values can be also entered manually.

- Opacity value selection within the range of (0-1). The value of 0, means that the marker will not appear in the background image at all (i.e. invisible watermarking). The value of 1 means that the marker will replace the selected part of the background image. The values between 0 and 1 will view the marker in a transparent mode.

- Marker size determination. User can choose one of the following sizes: (50X50, 100X100, 150X150, 175X175, and 200X200). If the selected marker size is greater than the background image, the user will be notified to choose another smaller size.

The user should apply all the specified options to get the final result (i.e. the watermarked image). When applying watermarking process, it is common that the user needs to rollback and go back to the previous state. To make this applicable, the system saves all watermarking steps in a stack. When the user clicks the rollback button, the system will remove the latest watermarking step. If the user chooses to roll back again, the watermarking step before the latest one will be removed and so on. When the user chooses another background image the stack will be cleared.

The user can add another marker(s) to the watermarked image. The system prevents markers overlapping by storing the position and the size for each added marker. When the user adds a new marker, the system compares the current position and size with all markers added before, if any interference occurs the user will be notified.

Markers overlapping can occur in two cases:

1) The user selects the same X and Y coordinates for an added marker on the background image.

2) The user selects different coordinates but unexpected part of the new marker overlaps an added marker. In this case, the user needs to make sure that the new selected position is away enough from other marker images around the selected position for the added marker.

V. RESULTS AND DISCUSSION

This section presents the experimental results of the proposed technique. The functionality of the implemented system is tested by performing watermarking on number of test images. Marker and background test images are of various types and sizes.

The proposed technique applies visible watermarking successfully on test images.

Fig 2 shows sample of some results of embedding different types of markers into background images where (a) shows RGB marker embedded into RGB background image, (b) shows RGB marker embedded into binary background, (c) shows grayscale marker embedded into grayscale image, and (d) shows binary marker embedded into grayscale image.

The proposed technique also allows embedding multiple markers into the same background image with no interference. Fig 3 shows an example. Also embedding repeated marker with different size and positions is also allowed. Fig 4 illustrates an example using camera man as a marker.

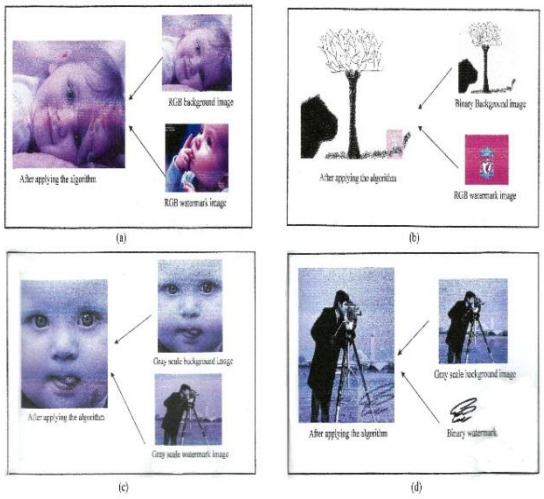


Fig. 2. Sample of some results of the proposed technique

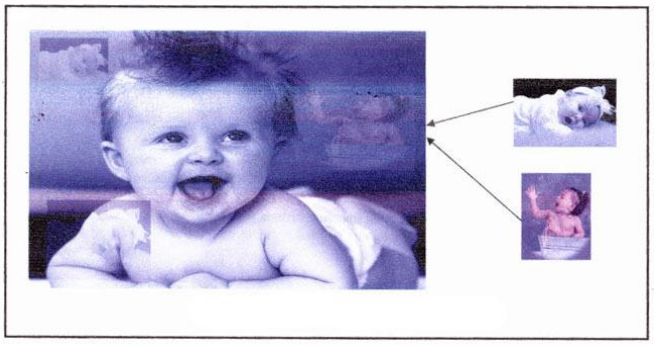


Fig. 3. Embedding multiple markers into the same background image

Fig. 4. An example of adding repeated marker with different size and positions

VI. CONCLUSION

A multi visible watermarking technique has been proposed. The proposed technique has presented adding different types of images into different types of background images. Embedding multiple markers on the same background image with no interference has been also proposed. Multi visible watermarking technique is based on storing the state of images (images types), resizing marker image to suitable size smaller than the original image size, determining insertion position of marker image, embedding marker image with background image based on number of channels using transparency equation, and storing width, height, and position of the marker image to prevent markers interference in the case of adding many markers. A visible system has been also designed to implement the proposed technique. The system provides many functionalities to allow single and multiple watermarking with different sizes, positions, and opacity levels.

For future work, we aim to embed animated watermarks. We also tend to expand the functionalities of the designed system to allow the user to enter the desired watermark as a text and embed this text into background image

REFERENCES

[1] C. Chen , and H. Tsai, "Wavelet-Based Reversible and Visible Image Watermarking Scheme", Springer-Verlag Berlin Heidelberg ,2011.

[2] D. Biswas, P. Das, P. Maji, N. Dey and S. Chaudhuri, "Visible Watermarking Within the Region of Non-Interest of Medical Images based on Fuzzy C-Means And Harris Corner Detection", Third International Conference on Computer Science, Engineering and Applications (ICCSEA 2013), Delhi, India, 24-26 May 2013.

[3] D. Jose, R. Karuppathal, and A. Kumar, "Copyright Protection using Digital Watermarking", National Conference on Advances in Computer Science and Applications with International Journal of Computer Applications (NCACSA), 2012.

[4] G. Kaur and K. Kaur, "Digital Watermarking and Other Data Hiding Techniques", International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-2, Issue-5, April 2013.

[5] H. Park and K.Kim, "Visible Watermarking using Verifiable Digital Seal Image", Symposium on Cryptography and Information Security, pp. 103- 108, Japan, 2001.

[6] J. Tamilselvi and S. Janakiraman, "A Visible Watermarking Scheme for DigitalImages in Frequency Domain", International Journal of Advanced Networking and Applications (IJANA), Vol. 4 Issue 3 pp. 1635-1639 ,2012.

[7] K.Madhu Kumar, M. Katta Swamy and B. Reddy, "Lossless Visible Watermarking Using Compound Mapping", International Journal of Engineering Research and Development, Volume 4, Issue 9, PP. 27-35,Nov. 2012.

[8] L. Saini and V. Shrivastava, "A Survey of Digital Watermarking Techniques and its Applications", International Journal of Computer Science Trends and Technology (IJCST) ,Volume 2 Issue 3, May-Jun 2014

[9] M. Bhaisare and V. Raut, "Generic Lossless Visible Watermarking: A Review" ,International Journal of Innovative Research in Computer and Communication Engineering, Vol. 3, Issue 1, January 2015.

[10] M. Majeed , S.C.Ramesh and R.Anuja, "Implementation of a Visible Watermarking in a Secure Still Digital Camera Using VLSI Design", International Symposium on Computing, Communication, and Control (ISCCC), 2009.

[11] M. Topkara, A. Kamra, and M. J. Atallah, "ViWiD: Visible Watermarking Based Defense against Phishing", Lecture Notes in Computer Science, Vol.3710, pp.470-483., 2005,

[12] P. Bidla, S. Gengaje, and R. Shelke, "Visible Image Watermarking Based On Texture and Luminance Blocks **In DCT Domain – A Review"**, International Journal of Emerging Technology and Advanced Engineering, Volume 2, Issue 4, April 2012.

[13] S. Parvathavarthini and R. Shanthakumari, "An Adaptive Watetmarking Process In Hadamard Transform", International Journal of Advanced Information Technology (IJAIT),Vol. 4, No. 2, April 2014.

[14] S. P. Mohanty , K. R. Ramakrishnan and M. S. Kankanhalli , "A DCT domain visible watermarking technique for images" , Proc. IEEE Int. Conf. Multimedia Expo. , vol. 2 , pp.1029 -1032 , 2000.

[15] S. Priya, B. Santhi and P. Swaminathan, "Image Watermarking Techniques - A Review", Research Journal of Applied Sciences, Engineering and Technology, July 2012.

[16] V. Singh, "Digital Watermarking: A Tutorial", Multidisciplinary Journals in Science and Technology, Journal of Selected Areas in Telecommunications (JSAT), Jan. 2011.

[17] Y. Perwej, F. Parwej and A. Perwej, "An Adaptive Watermarking Technique for the copyright of digital images and Digital Image Protection" ,The International Journal of Multimedia & Its Applications (IJMA) Vol.4, No.2, April 2012.

Miniaturized Meander Slot Antenna Tor RFID TAG with Dielectric Resonator at 60 Ghz

JMAL Sabri
Dept. of Physics,
El Manar University, Faculty of Sciences,
Tunisia

NECIBI Omrane
UR of Circuits and Elec. Sys. HF,
Dept. of Physics,
El Manar University, Faculty of Sciences,
Tunisia

TAGHOUTI Hichem
Departement of physics, chimical and processes Engineering
University of Carthage

MAMI Abdelkader
Dept. of Physics,
El Manar University, Faculty of Sciences,
Tunisia

GHARSALLAH Ali
UR of Circuits and Elec. Sys. HF, Dept. of Physics,
El Manar University, Faculty of Sciences,
Tunisia

***Abstract*—Day after day, recent advances in millimeter wave communications have called for the development of compact and efficient antennas. Furthermore, the greatest challenges in this area is to get a good performance and a miniaturized antenna. In addition, the design of antenna, in the Silicon technology, is one of the key challenges. In this way, this work will focus on the design of the free 60 GHz band, high gain and high efficiency on-chip antenna meandering slots for transponder Radio Frequency Identification (RFID). Further, the stacked dielectric resonators (DRS) will be arranged above the power element of on-chip antenna with an excitation of coplanar waveguide (CPW). We will use the Scattering Bond-Graph formalism like a new technique to design these proposed antennas and we will use the microwave Studio CST software simulation to validate the results. We have miniaturized the proposed antenna after having such a number of iteration and by applying the Bond Graph methodology, and the size of the antenna is about 1.2 * 1.1 mm^2.**

Keywords*—RFID TAG; Millimeter Wave Identification; Meander Slot Antenna; Dielectric Resonator Antenna (DRA); On-chip Antenna; Silicon; Scattering Matrix; Bond-Graph; Scattering Bond-Graph

I. INTRODUCTION

In the last few years, radio frequency identification (RFID) has become more and more popular in numerous applications, such as logistics, supply chains management, assets follow-up (active persons), and vehicles positioning [1]. Among a variety of RFID systems using radio frequencies, an RFID Ultra Height Frequency (UHF) system drew a lot of attention because of its numerous advantages, such as the cost, the size, and the long-range reading [2].

Millimeter Wave Identification (MMID) modernizes the RFID system to millimeter waves [3]. The higher frequency systems have several advantages: First, smaller antennas allow us to create miniaturized transponders. Second, antenna arrays can achieve a narrow beam for the reader antennas, which allow a well-organized transponder localization. Third, the regulations of relaxed radio at 60 GHz allow great data communication rate.

The most obvious advantage of the technology of 60 GHz is its strong capacity in terms of debit. Indeed, the Shannon theorem (1) shows that the capacity C of a communication channel (C = Max debit.) increases linearly with the bandwidth BW, while it increases following a logarithmic law according to the report signal with noise SNR:

$$C = BW.\log_2\left(1+SNR\right) \quad (1)$$

Thus, this relation leads us to conclude that to increase the debit of a radio channel, it is necessary to widen its bandwidth, rather than improve its report signal with noise. At these frequencies, antenna with dielectric resonator is the most voluminous component and its miniaturization constitutes one of the most important current challenges for the designers of communicating objects [4]. The use of dielectric resonators in the design of antennas for RFID tags in the EHF band has never been made. However, the miniaturization of antennas is usually accompanied by deterioration of its efficiency and bandwidth, because materials with high permittivity are generally used [5]. Therefore, miniature structures designed must present a compromise between bandwidth efficiency and physical size. Our objective is to conceive a structure of antenna with miniature dielectric resonators for RFID TAG in the V band (60 GHz) with correct performances.

On the other hand, the Bond Graph approach has showed [6], together with the Scattering formalism [7], its usefulness in studying and understanding the antennas [8].

A new way of applying the scattering bond graph technique will be used in this article. It is to show that the scattering bond graph methodology can serve as a new miniaturization approach to the design of RFID TAG in the V band.

In this paper, we focus only on one component of the MMID system, namely, the design of a very small antenna for an RFID TAG with a silicon technology.

In addition, in this work we are going to study the design and the simulation of an antenna with dielectric resonator of rectangular shape for RFID TAG in the EHF band. Then we are making a comparative study in terms of gain and volume of our structure with an UHF antenna of the trade and we finished our work with the validation of the simulation results by the Scattering Bond Graph methodology.

II. The Basic 60GHz RFID Antenna

In radio communication, antenna is a key element. We all know that antenna recovers the power and data signal traveling through the air. The quality and performance of the antenna are involved in the quality of wireless communication.

A. First Iteration

The configuration of the proposed antenna is shown in Figure 1.

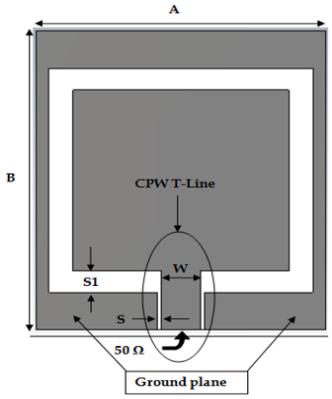


Fig. 1. This 3D on-chip antenna-top-sectional view of the feeding structure (x-y plane), where A =1.77, B =2, W =0.24, S =0.025, S1 =0.145 ;(all dimensions in mm)

The antenna structure design is based on the optimal length of the radial stub protrusion, which is in the range of 0.45-0.6 times the slot's width. Further, we have achieved the impedance matching by increasing the width of printed stub. We have also noted that the edge treatment of the stub has an impact on the bandwidth improvement. In addition, we have represented the slot region between the antenna input, the edge of the patch and the distal slot region, by Bond Graph elements by lumped admittances (capacitance and inductance). Moreover, this physical interpretation has been deduced by using the Bond Graph model shown in Figure 2.

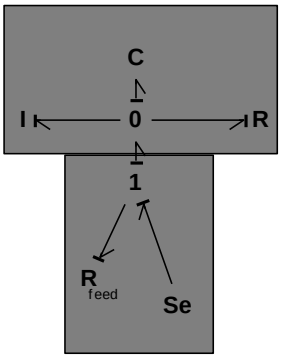


Fig. 2. 3D on-chip antenna-top-sectional Bond Graph model

Where the equivalent model of patch antenna is shown on the resistance R, inductance I and capacitance C interconnected by a ''0-junction''. A resistance R_{feed} and an effort source S_e modulate the excitation and the CPW line interconnected by ''1-junction''.

Thus, for the selected size of ground plane, we have a good input impedance-matching can be obtained and it is verified by our Bond-Graph methodology.

B. Second Iteration

The proposed antenna consists on a modulated loop element printed on the Rogers substrate. A detailed on-chip antenna is presented in Figure 3 with S1 is the rings width, S is the gap between the rings and W is the gap on both end of the rings. A coplanar waveguide (CPW) and a slot ring create the feeding structure [10]. It consists of a silicon based on-chip feeding structure.

The configuration of the proposed meander slot-based RFID loop antenna is shown in Figure 3.

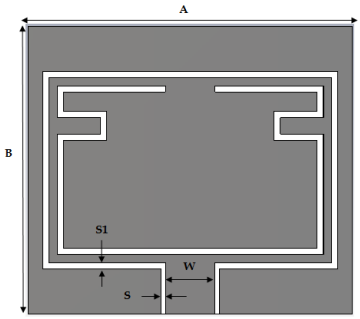


Fig. 3. 3D on-chip antenna-top-sectional view of the feeding structure (x-y plane), where A =1.57, B =1.4, W =0.24, S =0.02, S1 =0.03; (all dimensions in mm)

C. Simulation Results

This section presents the simulated results of the meander slot antenna. We use the scattering bond graph methodologies [11] and we use the CST Microwave studio software, for several performance parameters such as impedance bandwidth, and radiation patterns. For the scattering bond graph methodology, we take into account the equations demonstrated in the above work [12].

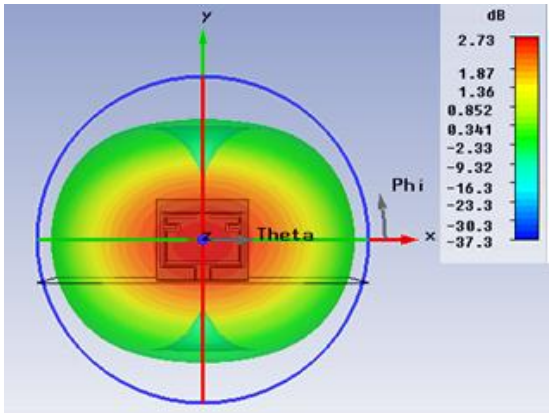


Fig. 4. Simulated gain of proposed antenna simulated by CST

We note that structures present important parameters, as shown in the simulation results: From the reflection coefficient S11, the resonance frequency is about 60 GHz for the two iterations. According to the far field, the two On-Chip-antennas present an acceptable gain but it decrease from 3.02dB to 2.73dB.

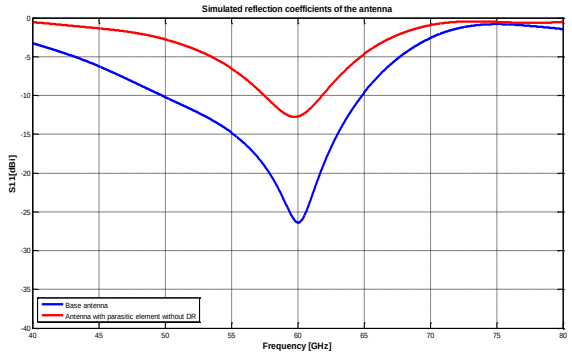


Fig. 5. Simulated reflection coefficient for the On-Chip antenna by Bond graph methodology

D. Above IC

To improve the performance of antennas on silicon, hybrid technology or "above IC" is proposed, which allow the optimization of the radiation efficiency by promoting the coupling between guided incident modes and radiated continuous modes.

Interests in silicon-based wireless system-on-chip (SOC) applications have become pervasive because of the advancement of its low-cost technology [13]. On-chip antennas eliminate the need for external off-chip connection and packaging process, which would incur a gain loss and overall size increase [14].

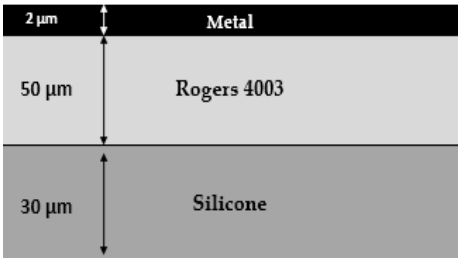


Fig. 6. Geometry of the proposed antenna (Button view)

We take into consideration the figure 6 shown blow we can conclude the bond graph model shown on the figure7.

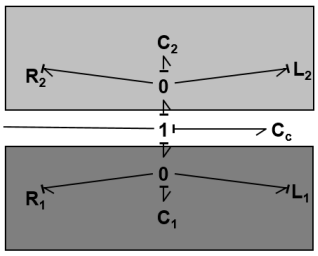


Fig. 7. Causal Bond graph model of proposed antenna

Where the values of the bond graph elements are deduced by the above work [15] That we have considered that any resonate element is considered a resistance R element, Capacitance element C and Inductance element I interconnected by 0 junction[16] .

To be sure, we can write the expression of the input impedance Z_{in} [17]

$$Z_{in} = \frac{R}{1+Q_t^2\left[\frac{f}{f_r}-\frac{f_r}{f}\right]^2} + j\left[X_L - \frac{RQ_t\left[\frac{f}{f_r}-\frac{f_r}{f}\right]}{1+Q_t^2\left[\frac{f}{f_r}-\frac{f_r}{f}\right]^2}\right] \quad (2)$$

The resistive element **R** represented the losses of the antenna in its planar environment. **R** is calculated by the following [18].

$$R = \frac{Q_t.h}{\pi f_r \varepsilon_{dyn}\varepsilon_0 WL}\cos^2\left(\frac{\pi x_0}{L}\right) \quad (3)$$

Where: f_r : Resonant frequency; Q_t : Quality factor; ε_{dyn} : Dynamic permittivity; x_0 : Distance between the feeding point and the patch; L, W: Dimensions of the patch. h: Substrate height.

Electrical energy storage modeled by the capacitance C. It is represented by the following equation [19].

$$C = \frac{Q_t}{2\pi f_r R} \quad (4)$$

To find *L*, we verify the relation represented in the following equation [20].

$$L = \frac{R}{2\pi f_r Q_t} \quad (5)$$

The coupling between the two layers is reflected only by a capacitive element C_c connected by 1junction. The expression of this capacitance represented in the [21].

III. The Proposed Tag Antenna with Dielectric Resonator

This study consists in analyzing and modeling the geometry of simple antenna excited under the $TE^X_{\delta 11}$ mode. The radiating element is formed by the concentration of two materials with different permittivity: it is a vertical superposition of the portions in a concentric way. The purpose is to excite the two resonators at the same frequency to have single-band functionality.

A. Rectangular Dielectric Resonator Antennas

The rectangular shape of the Dielectric Resonator (DR) offers additional advantages compared with other geometrical forms. Indeed, it possesses (Fig.6) two independent parameters (w/h and w/a), which allow it, in one hand, to have two freedom degrees to determine a specific resonance frequency to a given value of dielectric permittivity. On the other hand, a good selection of the dimensions of the resonator, that helps us to avoid the problem of the degeneracy

of modes, therefore an increase in cross-polarized antenna's implementation. The existence of the latter will degrade the antenna gain and will distort the radiation pattern [22]. Finally, the rectangular dielectric resonator antenna offers a greater flexibility in the optimization of the desired frequency band.

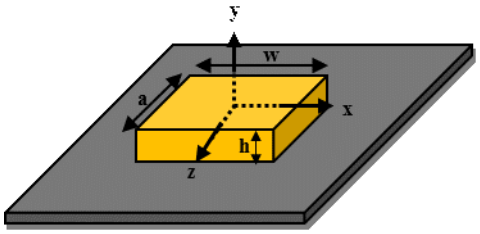


Fig. 8. Rectangular structures

After looking of this structure, we can deduce the bond graph model shown in Figure 9.

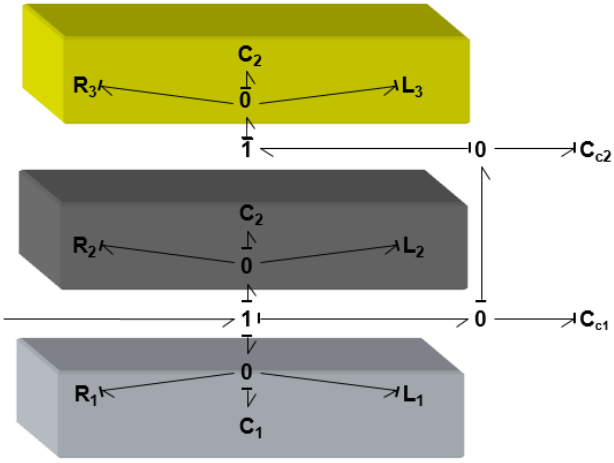


Fig. 9. Rectangular structures by Bond-Graph model

Analysis of the dielectric resonator antenna (DRA) can be made by considering the general classification methods proposed by Van Bladel for arbitrary forms of DR [23, 24]. According to Van Bladel, the modes of a dielectric resonator of high permittivity can be confined or not confined. For both types of modes, the electric component of the field, perpendicular to the dielectric/air interface, it must cancel what is explained in the following boundary condition

$$\vec{n}.\vec{E} = 0 \tag{6}$$

With $\vec{E}$ the electric field vector and $\vec{n}$ the normal direction to the separation surface DR / air.

The last relation is one of two boundary conditions that must be satisfied. The second condition is given as below:

$$\vec{n} \wedge \vec{H} = 0 \tag{7}$$

Where $\vec{H}$ represents the magnetic field.

As regards the nature of material that composes the DR, relation (6) could be no longer satisfied. Therefore, by confined modes, we consider that all modes satisfy both two boundary conditions ((6) and (7)). Only dielectric symmetrical shapes such as sphere and cylinder cannot support them. On the other hand, the not confined modes satisfy only the equation (6) [25] and they can supported by any arbitrary shape of the DR. Consequently, the DR does not present symmetry of revolution; it only supports the not confined modes.

Marcatilli demonstrated that the existence of TM modes in a rectangular dielectric wave-guide is questionable [26], because they do not satisfy the boundary condition (6) consequently, for the analysis of modes in a rectangular dielectric resonator antenna, only $TE^{X}_{\delta 11}$ modes are considered.

The $TE^{X}_{\delta 11}$ mode of the DRA is particularly interesting for antenna applications because it has the lowest mode rank and has the lowest quality coefficient.

B. Coupling Techniques

In dielectric resonators antenna design, the coupling scheme has a very important role. Indeed, the coupling system and its location affect the performance of antenna in terms of bandwidth, radiation pattern and polarization. According to the basic electromagnetic theory and the reciprocity theorem of the Lorentz law [27], coupling k between the source (magnetic or electronic) and the fields inside the dielectric resonator can be determined by the following relationships:

$$k \propto \int (\vec{E}.\overrightarrow{J_e}) dV \tag{8}$$

$$k \propto \int (\vec{H}.\overrightarrow{J_m}) dV \tag{9}$$

Where $\vec{E}$ and $\vec{H}$ are the vectors of the intensity of electric and magnetic field, $\vec{J}_e$ and $\vec{J}_m$ are the electric and magnetic currents, respectively.

To have a strong coupling between the source of electric current (magnetic) and dielectric resonator antenna, the source should be located where there is an important electric field (magnetic).

C. Operating Principle of Series-Fed DRA Array

Figure 10 shows the configuration of the proposed on-chip stacked DRA array. It consists on a feeding structure and one or more DRs. A dielectric resonator (DR_1) is mounted on the surface of the substrate, forming the traditional on-chip DRA investigated in [27] and [28].

A second dielectric resonator, DR_2 is arranged above the antenna in the vertical direction (z-axis), forming a series-fed linear array.

In the proposed array design, all the DRs resonate at the same frequency f_0 of the dominant mode $TE^{X}_{\delta 11}$. With the specified design frequency, the dimension of each DR having a dielectric constant of ε_r can be predicted using the traditional truncated dielectric waveguide model [29] or a simplified engineering formula presented in [30].

The operating mechanism of the proposed series-feed

antenna array can be explained as follows: most of the energy is initially coupled to DR1 from the on-chip feeding structure. DR1with dominated mode ($TE^{X}_{\delta 11}$) radiates with high efficiency there by acting as a traditional on-chip DRA. Then, part of the energy is serially coupled to the upper DRs (DR_2) to excite the mode.

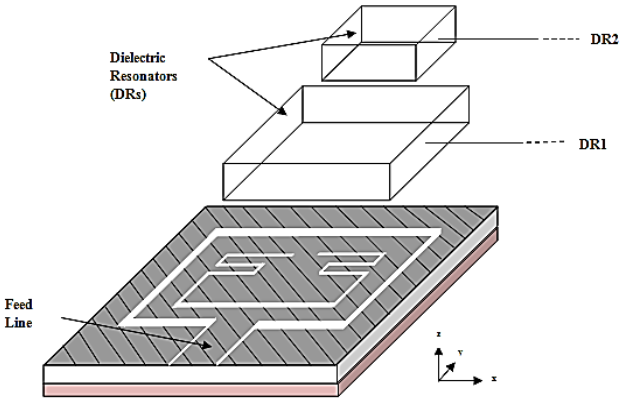


Fig. 10. On-chip stacked dielectric resonator antenna array: 3D structure of on-chip stacked dielectric resonators antenna array

The series coupling excitation leads to 180-phase difference of the electromagnetic (EM) waves between the two DRs.

D. The Antenna Structure

The proposed antenna is shown in the following figure.

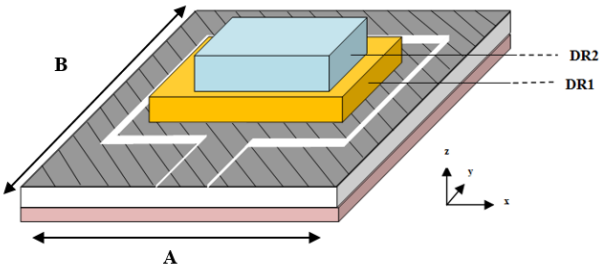


Fig. 11. 3D on-chip antenna with double dielectric resonators where A =1.2, B =1.117; (all dimensions in mm)

The physical interpretation of the proposed antenna using the Scattering Bond Graph methodology give us the following Bond Graph model which is shown in the following figure.

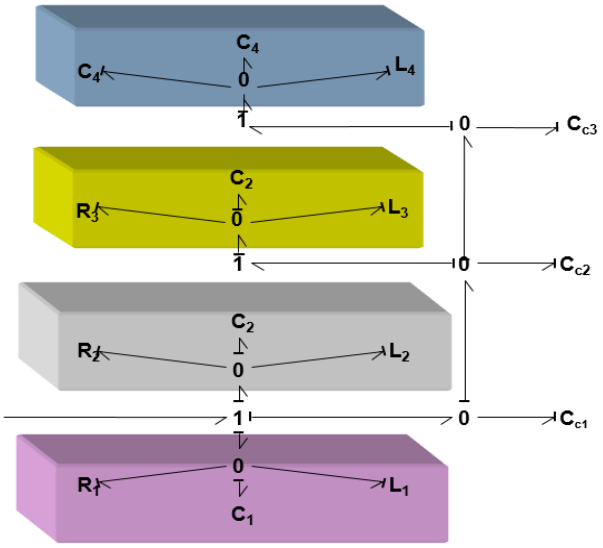


Fig. 12. On-chip antenna by bond graph methodology

A detailed double-DR arrangement of the on-chip antenna array is presented in Figure 11. It consists of a silicon based on-chip feeding structure and two DRs. A grounded coplanar waveguide (GCPW) and a meander slot create the feeding structure.

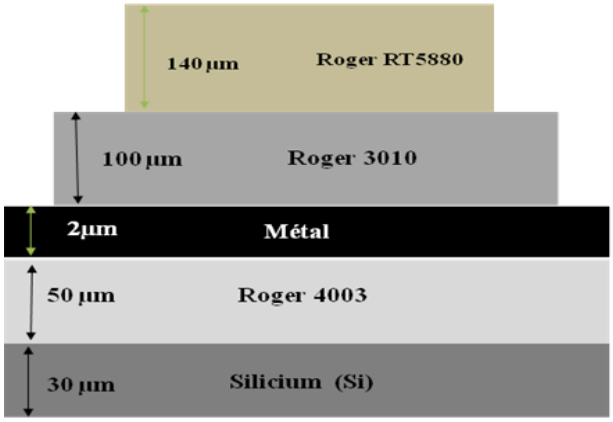


Fig. 13. Geometry of the proposed antenna with DRs (Button view)

The meander slot is implemented on the 2-µm thick top metal. It is shielded by the bottom metal and connected to a 50Ω GCPW with signal line width of w = 240 µm. The thickness of the Roger substrate is 50 µm.

For this study, the dielectric constant, ε_{r1} of the DR1 was selected to be 10.2 and the dielectric constant ε_{r2} of the DR2 was selected to be 2.2.

E. Design Procedure

The design procedure of the proposed array is described in the following paragraphs.

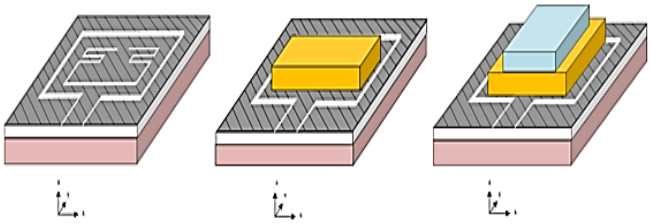


Fig. 14. Antenna design steps

In Step 1, the dimension of DR1 is first predicted using the simplified equations in [30], at f_0=60 GHz , and adjusted by 3D EM simulation. A meander slot feeding structure is designed to excite the dominant mode of DR1. In order to increase the bandwidth of the antenna, the slot feeding structure is designed as a radiator [31] at a resonant frequency that is close to that of DR1.

In Step 2, DR2 is also deliberated to resonate at f_0=60 GHz. The placement and dimension of DR2 will not affect the field distribution inside DR1. This also implies that DR2 has no effect on the input impedance of the antenna then we have validated the observation through simulations.

F. Simulation Results

This section presents the simulated results of the meander slot antenna. CST has been used to simulate the antenna for several performance parameters.

It is evident in Figure 15 that the antenna adaptation increases with the number of DRs, but reaches a tray with three or more DRs. Considering the gain performance and the assembly difficulties; the compromised number of DRs is two or three for each on-chip antenna. In this paper, we will focus only on the two-DR configuration.

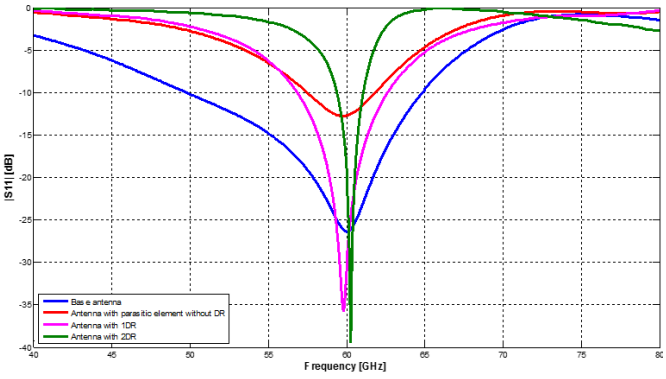


Fig. 15. Simulation results of the proposed antenna

The characteristics of the four antennas can be summarized as follows: according to the simulation results, we can see that the entire proposed antenna presents a resonance at 60GHz. The main advantage of using dielectric resonator is the possibility of reducing the size of the antenna by more than 40% of based antenna.

The final structure have an interesting reflection coefficient simulation. We have validate our work with scattering bond graph methodology by obtaining the simulation shown on the Figure 16.

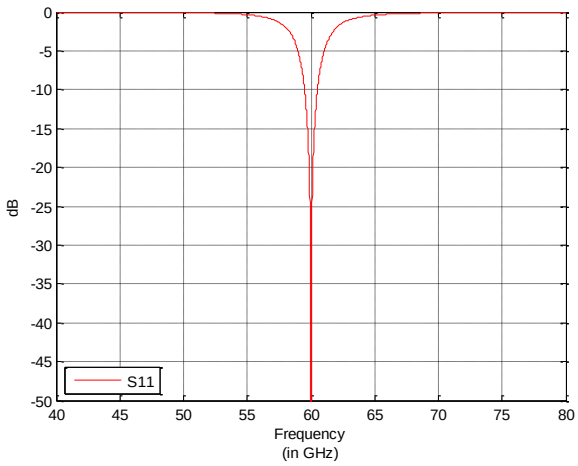


Fig. 16. Simulation of reflection coefficient of final antenna by Scattering Bond Graph methodology

Besides, the antenna is more adapted and the decreasing of the gain is little is from about 3dB to 2.35 dB. Antennas with multi-DRs are also investigated.

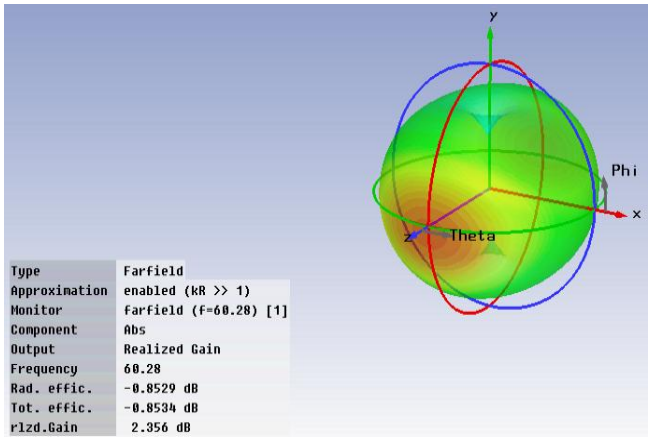


Fig. 17. Simulated gain of proposed antenna

G. Comparative Study with a UHF Antenna Trade

We are comparing our structure with the RFID antenna reader for Tag Product Company in terms of space and gain. Table 1 shows the study.

TABLE I. A COMPARATIVE STUDY BETWEEN THE STUDIED ANTENNA AND THE BUSINESS ANTENNA

	Patch antenna (Tag Product)	Proposed antenna
Frequency (GHz)	0.869	60
Gain (dB)	6	2.35
Size (L*W*h) (mm^3)	260*260*36	1.2*1.1*0.322
Volume (dm^3)	2.43	0.322

From the table above, we found that:

- The basic structure of the dielectric resonator has a total volume of 0.322 dm^3 compared to patch antennas, which have a trade volume of 2.43 dm^3. We notice a significant miniaturization.
- The gain of the DRA is lower of 3dBi compared to the patch with Tag Product antenna, which seems logical to us because we used a high-permittivity dielectric resonator ($\varepsilon_r = 10.2$).

IV. CONCLUSION

A novel antenna for an RFID TAG was simulated and evaluated. The proposed antennas have a resonant frequency equal to 60 GHz and it presents a very compact size and an acceptable gain.

In the first part, we presented the design of a patch antenna operating at 60 GHz linearly polarized then we added in this structure the loop meander slot. This form allowed us to reduce antenna size by 38%.

In the second part, we focused on designing a DRA in which the radiating element is formed by the arrangement of rectangular DR portions of different permittivity. The goal was to excite the two resonators to the desired frequency (60 GHz).

The use of dielectric resonator was a very useful and fast tool to ameliorate the antenna. This work is one of the first steps in the millimeter wave identification (MMID) and we desire in the future to use our antenna in MMID TAG Chip-Less.

The scattering bond graph methodology, used in this paper as a new design technique has shown its effectiveness in helping to understand and to interpret the studied antenna. This methodology can be used as a substitute for the iterative method since we can move from the first iteration to the last iteration without going through intermediate iterations because the scattering bond graph model can be simplified easily.

ACKNOWLEDGMENT

The preferred spelling of the word “acknowledgment” in America is without an “e” after the “g”. Avoid the stilted expression “one of us (R. B. G.) thanks ...”. Instead, try “R. B. G. thanks...”. Put sponsor acknowledgments in the unnumbered footnote on the first page.

REFERENCES

[1] A. Lehto, J. Nummela, L. Ukkonen, L. Sydänheimo and M. Kivikoski,. Passive UHF RFID in paper industry: Challenges, benefits and the application environment. Transactions on Automation Science and Engineering, IEEE, 6(1), (2009), pp. 66-79.

[2] E. Abdulhadi, and A. Ramesh "Design and experimental evaluation of miniaturized monopole UHF RFID tag antennas." Antennas and Wireless Propagation Letters, IEEE 11 (2012): 248-251.

[3] P. Pursula, T. Vähä-Heikkilä, A. Müller, D. Neculoiu, G. Konstantinidis, A. Oja, J. Tuovinen, “Millimetre Wave Identification — new radio system for low power, high data rate and short range,” IEEE Trans. on Microwave theory tech., vol. 56, no. 10, pp. 2221-2228, Oct. 2008.

[4] Lu, Kai, Kwok Wa Leung, and Yong Mei Pan. "Theory and experiment of the hollow rectangular dielectric resonator antenna." Antennas and Wireless Propagation Letters, IEEE 10 (2011): pp.631-634.

[5] Q. H. Lai, C. Fumeaux, W. Hong, and R. Vahldieck, “60 GHz aperture coupled dielectric resonator antennas fed by a half-mode substrate integrated waveguide,” IEEE Trans. on AP, vol.58, no.6, pp.1856-1863, 2010.

[6] S. JMAL. ‘’Etude d’une méthodologie Bond Graph pour la modélisation des circuits hauts fréquences’’. Master thesis, faculty of sciences of Tunis, Tunisia. (2012)

[7] H. Taghouti and A. Mami. "How to Find Wave-Scattering Parameters from the Causal Bond Graph Model of a High Frequency Filter."American Journal of Applied Sciences (2010).

[8] R. MEHOUACHI, H. TAGHOUTI and MAMI, A., "Analysis of microstrip patch antenna array by using a new Bond Graph technology ". American Journal of Applied Sciences, 11 (8): 1436-1449, June, 2014

[9] Balanis, Constantine A. Antenna theory: analysis and design. Vol. 1. John Wiley & Sons, 2005.

[10] T. Jang, J. Choi and S. Lim, “Compact Coplanar Waveguide (CPW)-Fed Zeroth-Order Resonant Antennas With Extended Bandwidth and High Efficiency on Vialess Single Layer”, IEEE Transactions on Antennas and Propagation Vol 59 , no. 2 , pp. 363 – 372, Feb. 2011.

[11] S. Jmal, H. Taghouti, A. Mami, Design of a Ka Band Antenna by a New Methodology Based on Bond Graph Approach, International Journal on Communications Antenna and Propagation (IRECAP), 5(5), (2015) pp. 301-306.doi:http://dx.doi.org/10.15866/irecap.v5i5.7565

[12] H.Taghouti, , A. Mami, and S. Jmal, Nouvelle Technique de Modélisation et Simulation par Bond Graph: Applications aux Circuits Hauts Fréquences et Antennes Patch. (2014) Éditions universitaires européennes.

[13] B. Razavi, “Design of millimeter-wave CMOS radios: A tutorial,” IEEE Transaction Circuits Systems. I, Reg. Papers, vol. 56, no. 1, pp. 4–16, Jan. 2009.

[14] Y. C. Ou and G. M. Rebeiz, “On-chip slot-ring and high-gain horn antennas for millimeter-wave wafer-scale silicon systems,” IEEE Trans.Microwave Theory Tech., vol. 59, no. 8, pp. 1963–1972, Aug. 2011.

[15] S. JMAL, H. TAGHOUTI and A. MAMI, “Modeling and simulation of a patch antenna from its Bond Graph model”. International Conference on Control, Decision and Information Technologies (CoDIT), 2014. IEEE, 2014. p. 609-614.

[16] S Jmal, H.Taghouti and A. Mami. “A new modeling and simulation methodology of a patch antenna by Bond Graph approach”. International Conference on Electrical Engineering and Software Applications (ICEESA), 2013, pp. 1-6. IEEE.

[17] TK. Chen and GH. Huff,” On the Constant Input Impedance of the Archimedean Spiral Antenna in Free-Space”, Vol 62, no. 7, pp. 3869 – 3872, July 2014.

[18] J. J.Wang, , Y. P.Zhang, , K. M. Chua and Lu, A. C. W. “Circuit model of microstrip patch antenna on ceramic land grid array package for antenna-chip codesign of highly integrated RF transceivers”. Antennas and Propagation, IEEE Transactions on, 53(12), 3877-3883. (2005).

[19] L. A Belov, S.M. Smolskiy, and V. N. Kochemasov, “Handbook of RF, Microwave, and Millimeter-wave Components”. Artech House. (2012)

[20] I. J. Bahl. “Lumped elements for RF and microwave circuits”. Artech house. (2003)

[21] K. C. Gupta, R. Garg, L. Bahl, and P. Bhartia. “Microstrip Lines and Slotlines”. Second edition . Arthech House.

[22] KX. Wang and H. Wong, “A Circularly Polarized Antenna by Using Rotated-Stair Dielectric Resonator”, IEEE Antennas and Wireless Propagation Letters, Vol 14, pp 787 – 790, 2015.

[23] X. Sheng Fang, K. Wa Leung "Linear-/Circular-Polarization Designs of Dual /Wide-Band Cylindrical Dielectric Resonator Antennas” IEEE Transactions On Antennas And Propagation, Vol. 60, no. 6, JUNE 2012.

[24] Y. Ding, K. Wa Leung and K. Man Luk, "Compact Circularly Polarized Dual-band Zonal-Slot/DRA Hybrid Antenna Without External Ground Plane", Antennas and Propagation, IEEE Transactions, Vol. 59, no. 6, pp. 2404 - 2409, 2011.

[25] S. Dhar, R. Ghatak, B. Gupta, D.R Poddar, “A Wideband Minkowski Fractal Dielectric Resonator Antenna "Antennas and Propagation, IEEE Transactions on", Vol 61, no. 6, pp. 2895 - 2903, 2013.

[26] J. Avella Castiblanco, D. Seetharamdoo, M. Berbineau, M. Ney and F. Gallee, "Electromagnetic modeling and definition of antenna specifications and positions for radio system deployment in confined environments", IOP Conference Series: Materials Science and Engineering, 2013.

[27] D. H. Neil H. E. Weste, CMOS VLSI Design, 3rd ed. Addison Wesley, 2004.

[28] P. V. Bijumon, Y. Antar, A. P. Freundorfer, and M. Sayer, "Dielectric resonator antenna on silicon substrate for system on-chip applications," IEEE Trans.on AP, vol. 56, no. 11, pp. 3404–3410, Nov. 2008.

[29] M. R. Nezhad-Ahmadi, M. Fakharzadeh, B. Biglarbegian, and S. Safavi-Naeini, "High-efficiency on-chip dielectric resonator antenna for mm-wave transceivers," IEEE Transon AP, vol. 58, no. 10, pp. 3388–3392, Oct. 2010.

[30] D. Hou, Y. Z. Xiong, W. Hong, W. L. Goh, and J. Chen, "Silicon-based On-chip Antenna Design for Millimeter wave/THz Applications," in Proc. IEEE Electrical Design of Advanced Package. Systems Symp, pp. 1–4, Dec. 2011.

[31] SY. Yang, M. Ng Mou Kehn, "A Bisected Miniaturized ZOR Antenna with Increased Bandwidth and Radiation Efficiency", IEEE Antennas and Wireless Propagation Letters, Vol. 12, pp. 159 – 162, 2013.

Word Sense Disambiguation Approach for Arabic Text

Nadia Bouhriz
Dept. of Mathematics and computer science
Faculty of Sciences Ben M'sik
Hassan II University
Casablanca, Morocco

Faouzia Benabbou
Dept. of Mathematics and computer science
Faculty of Sciences Ben M'sik
Hassan II University
Casablanca, Morocco

El Habib Ben Lahmar
Dept. of Mathematics and computer science
Faculty of Sciences Ben M'sik
Hassan II University
Casablanca, Morocco

***Abstract*—Word Sense Disambiguation (WSD) consists of identifying the correct sense of an ambiguous word occurring in a given context. Most of Arabic WSD systems are based generally on the information extracted from the local context of the word to be disambiguated. This information is not usually sufficient for a best disambiguation. To overcome this limit, we propose an approach that takes into consideration, in addition to the local context, the global context too extracted from the full text. More particularly, the sense attributed to an ambiguous word is the one of which semantic proximity is more close both to its local and global context. The experiments show that the proposed system achieved an accuracy of 74%.**

***Keywords*—Word Sense Disambiguation; Arabic Text; local context; global context; Arabic WordNet; Semantic Similarity**

I. Introduction

WSD is a natural language processing (NLP) field. It aims at determining the appropriate sense of an ambiguous word occurring in a given context [1] [2]. It is a task which allows a better understanding, and consequently a better exploitation of the processed linguistic material. It is therefore very essential task for NLP applications, such as Machine Translation (MT), Information Retrieval (IR), Text classification…etc.

The oldest WSD approach proved that two words before and after the ambiguous word are sufficient for its disambiguation [3]. For the Arabic language, the information extracted from this local context is not always sufficient.

To solve this problem, an Arabic WSD system was proposed in this paper that is not only based on the local context, but also on the global context extracted from the full text. The objective is to combine the local contextual information with the global one for a better disambiguation.

More particularly, the proposed system uses the resource Arabic WordNet (AWN) to select word senses. The sense attributed then to an ambiguous word is the one that possesses the closest semantic proximity to the local context, as well as to the global one. This proximity is measured based on the semantic hierarchy offered by WordNet.

The rest of the paper is organized as follows: Section II presents the architecture of WSD systems. Section III exposes some Arabic WSD systems. Section IV displays the description of the proposed system. Section V contains experiments and the obtained results. The last section gives conclusion and some perspectives.

II. WSD Systems Architecture

In 1949, Weaver [4] discussed the necessity of WSD for MT, and he explained that to realize this process, the ambiguous word must be taken from the context where it occurred. In 1950, Kaplan [3] made experiences to determine in which size the context should be, in order to disambiguate a word. It proved that two words at the right and at the left (size =2) of the ambiguous word are sufficient for its disambiguation; Masterman [5] confirmed this result for the Russian language, while Choueka and Lusignan [6] confirmed it for the French.

Over the years, WSD systems were developed according to different approaches. Actually, these systems have generally an architecture structured around three main steps:

- Sense inventory: consists on selecting the senses of the words.
- Context representation: represents senses and contexts in a formal manner.
- Disambiguation Process: attributes for every ambiguous word its correct sense according to its context.

The sense inventory step is the one that makes the difference from one system to another depending on the adopted approach. Generally, two approaches exist:

The first one, called Knowledge-based approach, is based on the use of external lexical resources. These resources are containing all the words of a language with their senses. These resources can be dictionaries [7], thesaurus [8], or ontologies [9] [10].

Unlike the first approach, the second one doesn't use external lexical resources, but it acquires the necessary information to define words' senses from a corpus; it's called a Corpus-based approach. This information is obtained by the application of statistical language models on this corpus. Three approaches are distinguished in this category, supervised approaches that require annotated corpus [11] [12] [13], unsupervised approaches [14] [10] that require unannotated corpus, and a semi-supervised approaches that require both of the annotated and the unannotated corpus [15].

III. ARABIC WSD SYSTEMS

A. Challenges

Arabic presents several challenges for WSD, due essentially to the particularities of this language and also to the lack of resources necessary to the disambiguation process.

Diacritics' missing in Arabic texts is the most challenging characteristic for WSD; because it increases the number of a word's possible senses and consequently makes the disambiguation task more difficult. For example, the word without diacritics (Swt صوت) have 11 senses according to the AWN, while the use of diacritics for the same word (Saw~ata صَوَّتَ), cuts down the number of senses to two.

On the other hand, the Arabic language is very rich morphologically. This causes an ambiguity during the lexical segmentation, and influences consequently the detection of the words' correct sense during disambiguation process. For example, the word (وجد Wjd) have two possible segmentations; the first one considers that the letter (و W) is a prefix of (جد Jad), while the second considers it as a letter in the word, which gives two totally different words.

B. State of the Art

The first WSD systems were mostly concerned by Latin languages like English and French since several decades ago. The Arabic language, as for it, didn't get the attention until the last decade.

The first Arabic WSD system was proposed by Mona Diab in 2002 [10]. The author introduced in this work an unsupervised method to annotate Arabic words by their sense using English WordNet and an English-Arabic parallel corpus.

Another contribution was proposed by Elmougy [13] where a Naïve Bayes Classifier was used to disambiguate Arabic words without diacritics.

Merhbene [16] was based on the semantic trees and a measure of collocation to choose the most appropriate sense to an ambiguous word.

Zouaghi [17] have proposed a system of WSD by combining the information retrieval measures with the Lesk algorithm to estimate the most appropriate sense of the ambiguous word.

The most recent work was proposed by Menai [18], in which the author was based on the genetic algorithms. His objective is to exploit the power of these algorithms in the Arabic WSD.

All of the previously mentioned works used only one contextual information to disambiguate. The proposed system, as for it, leans on two contextual informations. The first one is extracted from the local context of the ambiguous word and the second from its global context.

IV. THE PROPOSED SYSTEM

Before describing the system process, the structure of Arabic WordNet is firstly given.

A. Arabic WordNet

The Arabic WordNet (AWN) [19] [20] [21] is a lexical resource for modern standard Arabic. It was constructed according to the Princeton WordNet content. It's organized around elements called Synsets, which are a set of synonyms and pointers connecting it with other synsets. So, the AWN is a lexical network in which synsets represent its nodes and the connections between synsets represent its edges.

This resource counts at present 23,481 words organized into 11,269 synsets. A word can belong to one or more synsets.

In this work, the senses of a word are defined by the Synsets to which it belongs in the AWN. Below, some words synsets (i.e. senses) extracted from AWN are presented:

TABLE I. EXAMPLE OF AWN SENSES

words	Senses (synsets)
بحر	Sense 1 = [بحر] Sense 2 = [محيط, بحر]
شعر	Sense 1 = [شعر, قصيدة] Sense 2 =[شعر] Sense 3 = [شعر] Sense 4 = [أحس, شعر] Sense 5 = [حس, أحس, شعر] Sense 5 = [أحس, شعر]
مال	Sense 1 = [فلوس, ثروة, دراهم, مال] Sense 2 = [نقود, مال] Sense 3 = [مال] Sense 3 = [تمايل, ترنح, مال] Sense 4 = [انحدر, مال] Sense 5 = [نزع إلى, مال] Sense 6 = [أقنع, أمال, مال] Sense 7 = [انحرف, انحنى, مال]

B. Description of the proposed system

1) Sense inventory

In this step, a preprocessing phase is applied; it contains a text segmentation process, a stop words removal process, and finally a stemming process to remove words' affixes (prefixes and suffixes).

Afterwards, the obtained words are classified, according to the AWN, into two categories:

- Non ambiguous words: belonging to one Synset, i.e. possessing one sense.
- Ambiguous words: belonging to several Synsets, i.e. possessing several senses.

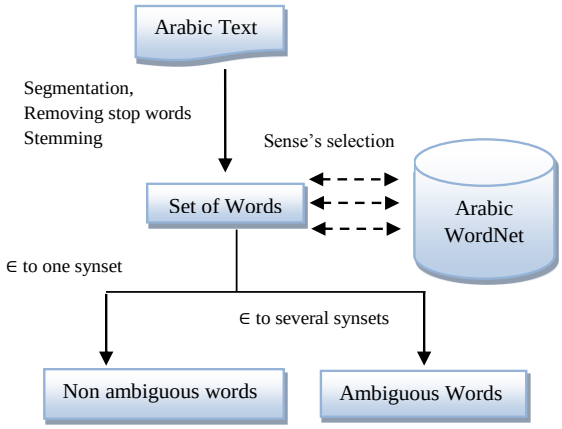


Fig. 1. Sense inventory process

Sense inventory Algorithm

Input: Arabic Text T
Output: List of Ambiguous Words AW and Non Ambiguous Words NAW.

1: **Segment** the Text
2: **Remove** stop words
3: **Apply** Stemming process for all obtained words
4: **For each word** w **do:**
5: **If**: w is belonging to one synset in AWN
6: **Then**: Add w to NAW list.
7: **Else if**: w is belonging to a several synsets in AWN
8: **Then**: Add w to AW list.
9: **End**

2) Context representation

This step consists of representing words' senses as vectors. For this purpose, the set of all non ambiguous word senses(S_1, S_2, S_3,…, S_n) is firstly considered, afterwards, the vector space, spanned by the standard basis $B = \{e_i\}_{i=1..n}$, where $e_1 = (1,0,0,\dots,0)$, $e_2 = (0,2,0,\dots,0)\dots, e_n = (0,0,0,\dots,1)$ are respectively the unit vector of the sense S_i, is built.

Using this space, words' senses will be represented by the vector $V = \sum_{i=1}^{n} a_i\, e_i$ where a_i is the i^{th} coordinate representing the semantic distance between the word sense and the sense S_i in AWN. To calculate this distance, the Wu and Palmer (wu-p) measure is used [2].

The global context will be afterwards defined by the sense vectors set of non ambiguous words present in the full text: $Contx_{Global} = \{V_1, V_2, \dots, V_n\}$, while the local context will be defined by the sense vectors set of non ambiguous words present only locally: $Contx_{Local} = \{V_i$ where V_i is the vector sense of i^{th} non ambiguous word present locally $\}$.

Finally, an ambiguous word aw that has m senses will be represented by the set of its sense vectors:

$aw = \{W_1, W_2, \dots, W_m\}$.

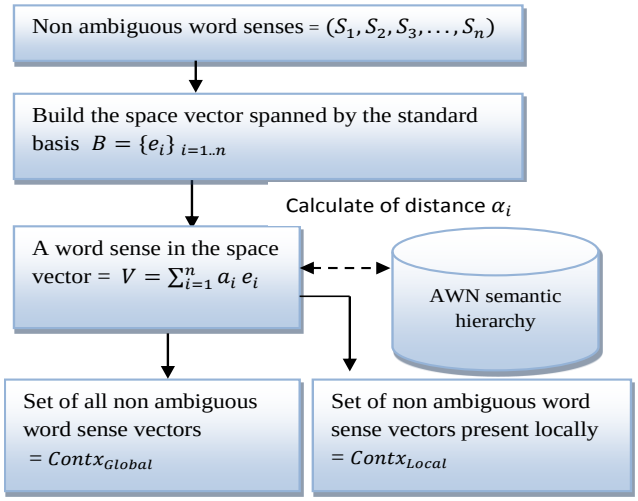


Fig. 2. Context representation

Context representation Algorithm

Input: list of AW and NAW
Output: $Contx_{Global}$, $Contx_{local}$,

1: **For all** words in NAW **do:**
2: **Extract** associated senses (S_1, S_2, S_3,…, S_n)
3: **End**
4: **For each** word **do:**
5: **For each** senses **do:**
6: **For each** S_i **do:**
7: **Calculate** the wu-p semantic distance a_i between word sense and the sense S_i.
8: **End**
9: **Calculate** word sense vector $V = \sum_{i=1}^{n} a_i\, e_i$
10: **End**
11: **End**
12: **Construct** $Contx_{Global}$
13: **Construct** $Contx_{Local}$

3) Disambiguation process:

This last step consists of attributing for each ambiguous word its appropriate sense. This is done by choosing the sense with the closest semantic proximity to its local and global context.

Sense semantic proximity with a context is defined by the percentage of vectors in this context that are similar to the vector of this sense.

Similarity measurement between two vectors $V=(v_1, v_2, \dots, v_n)$ and $W = (w_1, w_2, \dots, w_n)$ can be calculated by three distances which are; dot product, cosines, and Jaccard defined respectively as follows:

$$\text{dotProduct}(V.W) = \sum_{i=1}^{n} v_i.w_i$$

$$\cos(V,W) = \frac{\sum_{i=1}^{n} v_i.w_i}{\sum_{i=1}^{n} {v_i}^2 . \sum_{i=1}^{n} {w_i}^2}$$

$$\text{Jaccard}(V,W) = \frac{\sum_{i=1}^{n} v_i.w_i}{\sum_{i=1}^{n} {v_i}^2 . \sum_{i=1}^{n} {w_i}^2 - \sum_{i=1}^{n} v_i.w_i}$$

According to the previous definitions, the local and global semantic proximity are measured for each ambiguous word sense; as a result, a pair of percentages representing respectively each of the semantic proximity is obtained. The sense with the better average of its two percentages will be assigned finally to the ambiguous word.

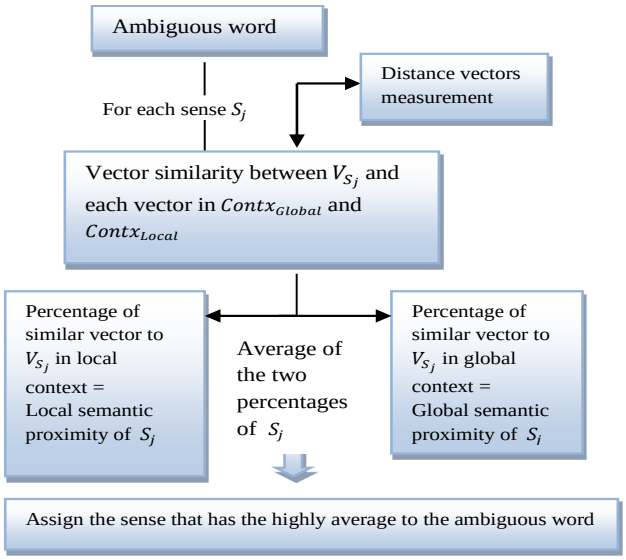


Fig. 3. Disambiguation Process

Disambiguation process Algorithm

Input: ambiguous word
Output: sense of the ambiguous word

1: **For each** sense S_j of the ambiguous word **do:**
2: **Calculate** local semantic proximity of S_j
3: **Calculate** global semantic proximity of S_j
4: **Calculate** average semantic Proximity of S_j
5: **End**
6: $BestSense = S_1$
7: **For each** sense S_i of the ambiguous word **do:**
8: **If: (** $average(S_i) > average(S_1)$**)**
9: **Then:** $BestSense = S_i$
10: **End**
11: **Assign** the sense $BestSense$ to the ambiguous word.

V. Experimentations and Results

To evaluate the proposed system, a test corpus is constructed by collecting texts from various fields (news, sport, medicine, religions, etc.); afterwards, each word is annotated manually by its correct sense according to the AWN.

The Java language was used to implement the system, and to access the XML AWN database the 'Java API for Arabic WordNet'[1] was used. Finally, the application of the stemming process is based on SAFAR platform[2].

For measuring the system's efficiency, the precision measurement was used; it consists of the number of words correctly disambiguated divided by the number of all ambiguous words. Experiment results have shown a precision of 74%.

Another experiment have shown that the use of a stemming process during the sense inventory phase increases the system's efficiency. More particularly, results (Table II) show firstly that the use of this process increases efficiency by 30%, moreover, they have shown that the use of AlKhalil Analyzer [23] is better than Buckwalter [24] by 4%:

TABLE II. Importance of the Stemming Process

Morphological analyzer	Without Stemming	Buckwalter	AlKhalil
System precision	40%	70%	74%

The table below (Table III) show some disambiguated words from this piece of text:

وفي سياق متصل قال المتحدث الرسمي في الرئاسة العامة للأرصاد وحماية البيئة حسين القحطاني انه يوجد بالرئاسة مركز للبلاغات ورقم مجاني (988) لاستقبال بلاغات الكوارث الطبيعية والبحرية، كما أن هناك خطة وطنية للاستجابة ومكافحة التلوث تضم في عضويتها الجهات ذات العلاقة بالتلوث البحري.

TABLE III. Example of Words Disambiguated

Ambiguous word	Senses	Local Semantic Proximity	Global Semantic Proximity	Average Semantic Proximity	Sense selected
رئاسة	رئاسة,إدارة	0%	1.88%	0.94%	قيادة,دور قائد,رئاسة,دور رئيس
	قيادة,دور قائد,رئاسة,دور رئيس	16.6%	11.3%	13.95%	
	فترة,رئاسية,رئاسة,إدارة	0%	0%	0%	
	قيادة,زعامة,رئاسة	0%	0.9%	0.45%	
حماية	حماية,جهاز,وقاية,جهاز حماية,سلامة	16.6%	18.8%	17.7%	حماية,جهاز,وقاية,جهاز حماية,سلامة
	حماية,تركيب دفاعي,دفاع	16.6%	18.8%	17.7%	
	حماية,رعاية,حراسة,عناية	0%	1.88%	0.94%	
	حماية,رعاية,وقاية	16.6%	16.9%	16.75%	
	حماية,حفظ	0%	0%	0%	
	حماية,اهتمام,رعاية,وقاية	16.6%	11.3%	13.95%	
مركز	مركز,موقع,مكانة	0%	0%	0%	مركز
	مركز,بؤرة	0%	11.32%	5.66%	
	مركز,موقع	0%	18.8%	9.4%	
	مركز	16.6%	20.75%	18.67%	
	مركز,مكان	0%	0%	0%	
	مركز,بؤرة,محرق	0%	11.32%	5.66%	
	مركز,منتصف,وسط	0%	11.32%	5.66%	
	موضع,مركز,موقع,علاقة مكانية	0%	3.77%	1.88%	
	موضع,مركز,مكان,موقع	0%	11.32%	5.66%	
	مادة,تركيز,المنطقة,مركز	0%	5.66%	2.83%	
رقم	رقم,رقم,تعيين الهوية	0%	5.66%	2.83%	رقم,عدد
	رقم,عدد	0%	11.32%	5.66%	
	رقم	0%	0%	0%	
	علم,وضع علامات الترقيم,رقم,نقط	0%	0%	0%	
خطة	استراتيجية,خطة,مخطط	0%	0%	0%	استراتيجية,خطة,مخطط
	نظام,برنامج,خطة	0%	0%	0%	
	خطة	0%	0%	0%	
	مؤامرة,خطة	0%	0%	0%	
استجابة	إجابة,رد,استجابة,جواب	0%	18.8%	9.4%	إجابة,رد,استجابة,جواب
	رد فعل,استجابة	0%	0%	0%	
	تلبية,إجابة,رد,استجابة,جواب	0%	11.32%	5.66%	

The last experiment results show that the proposed approach is better by 0.34% than the classical method (based on local context). This is due to some challenges described as follows:

- The non-recognition of named entities (persons' names, locations, organizations…etc.). These last should not be separated during segmentation process. Experiments show that words like: عبد الله, أبو ظبي have not been recognized as a named entity.

[1]https://sourceforge.net/projects/javasourcecodeapiarabicwordnet/
[2] http://sibawayh.emi.ac.ma/safar/publications.php

- Another similar challenge that decreases the system efficiency is the incapability of multiword expression recognition such as قاعدة بيانت, الأمم المتحدة…etc.
- The absence of a component that allows disambiguating senses with the same average semantic proximity.
- The absence of a part-of-speech tagging that allows categorizing words in verbs and names allowing consequently studying names and verbs in a separate way.
- The last challenge is relying on the lexical resource used. The AWN doesn't cover all Arabic words, which has consequently an impact on the system efficiency. For example the word منطق doesn't belong to the AWN structure.

VI. Conclusion

In this paper, a WSD system for Arabic texts was presented. The proposed system, unlike other systems, takes into consideration two types of context during disambiguation process. The first one is the local context defined by the words in the neighborhood of the ambiguous word, and the second is the global context defined by the full text.

Experiments have shown an accuracy of 74% for the proposed system.

The incorporation of a named entities and a multiword expression component in the process will be necessary done in the future for a better results, as well as a raise of all the challenges previously mentioned.

As for the future, this method will be integrated in a semantic indexing process to help enhancing Arabic information retrieval system.

References

[1] E. Agirre, P. Edmonds, ''Word sense disambiguation: algorithms and applications''. Springer, 2006.

[2] R. Navigli, ''Word sense disambiguation: a survey''. ACM Comput Surv 41(2):1–69, 2009.

[3] A. Kaplan, ''An experimental study of ambiguity and context''. Mechanical Translation 2(2), 39–46, 1955.

[4] W. Weaver, ''Translation''. in Machine Translation of Languages, MIT Press, Cambridge, MA, 1949.

[5] M. Masterman, ''Semantic message detection for machine translation, using an interlingua". International Conference on Machine Translation of Languages and Applied Language Analysis, Her Majesty's Stationery Office, London, 437-475, 1961.

[6] Y. Choueka, S. Lusignan, ''Disambiguation by short contexts''. Computers and the Humanities, 19, 147-158, 1985.

[7] M.E. Lesk, ''Automatic sense disambiguation using machine readable dictionaries : How to tell a pine cone from a nice cream cone''. In Proceedings of the SIGDOC Conference, Toronto. 1986.

[8] D. Yarowsky. ''Word-Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora''. Proceedings of the 14th International Conference on Computational Linguistics (COLING-92), 454 –460, 1992.

[9] P. Resnik. ''Disambiguating noun groupings with respect to WordNet senses'', in S. Armstrong, K. Church, P. Isabelle, S. Manzi, E. Tzoukermann & D. Yarowsky, eds, 'Natural Language Processing Using Very Large Corpora', Kluwer Academic Publishers, Boston, M.A, pp. 77–98, 1999.

[10] M. Diab, P. Resnik ''An unsupervised method for word sense tagging using parallel corpora''. in Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL-02), Philadelphia, PA, pp. 255–262, 2002.

[11] E.F. Kelly, P.J. Stone. ''Computer recognition of english word senses''. North-Holland Publishing. North-Holland, Amsterdam. 1975.

[12] T. Pedersen. ''Learning Probabilistic Models of Word Sense Disambiguation''. PhD thesis, Southern Methodist University, Dallas 1998.

[13] S. Elmougy, T. Hamza, H.M. Noaman. ''Naive Bayes classifier for Arabic word sense disambiguation''. In: Proceedings of INFOS 2008, Cairo, pp 27–29, 2008.

[14] H. Schutze. ''Automatic word sense discrimination. Computational Linguistics''. Special Issue on Word Sense Disambiguation, 24 (1), 97–123, 1998.

[15] D. Yarowsky. ''Unsupervised word sense disambiguation rivaling supervised methods''. In 33th Annual Meeting of the Association for Computational Linguistics, pp 189-196, 1995.

[16] L. Merhbene, A. Zouaghi, M. Zrigui. ''Approche basée sur les arbres sémantiques pour la désambiguïsation lexicale de la langue arabe en utilisant une procédure de vote''. proceeding de 21ème conférence sur le Traitement Automatique des Langues Naturelles, Marseille 2014.

[17] A. Zouaghi, L. Merhbene, M. Zrigui. ''Combination of information retrieval methods with LESK algorithm for Arabic word sense disambiguation''. Artif Intell Rev 38:257–269, 2012.

[18] M.E. Menai. ''Word sense disambiguation using evolutionary algorithms – Application to Arabic language''. Computers in Human Behavior 41 : 92–103, 2014.

[19] W. Black, S. El-Kateb. ''A Prototype English-Arabic Dictionary Based on WordNet''. http://www.fi.muni.cz/gwc2004/proc/95.pdf . 2004.

[20] C. Fellbaum, W. Black, S. Elkateb, A. Marti, A. Pease, H. Rodriguez, P. Vossen. ''Constructing Arabic WordNet in Parallel with an Ontology''. http://www.globalwordnet.org/AWN/meetings/meet20050901/Fellbaum .ppt , 2005.

[21] S. Elkateb, W. Black, P. Vossen, D. Farwell, A. Pease, C. Fellbaum. Arabic WordNet and the Challenges of Arabic. http://www.mt-archive.info/BCS-2006-Elkateb.pdf, 2006.

[22] Z. Wu, M. Palmer. ''Verb semantics and lexical selection''. Proceedings of 32nd annual Meeting of the Association for Computational Linguistics, Las Cruces, New Mexico, pp 27-30, 1994.

[23] A. Boudlal, A. Lakhouaja, A. Mazroui, A. Meziane, O.B.M Ould Abdallahi, and M. Shoul. ''Alkhalil Morpho SYS1: A Morphosyntactic Analysis System for Arabic Texts''. In the proceedings of the 11th International Arab Conference on Information Technology. Benghazi, Libya 2010.

[24] T. Buckwalter. ''Buckwalter {Arabic} Morphological Analyzer Version 1.0'' , 2002.

A Format-Compliant Selective Encryption Scheme for Real-Time Video Streaming of the H.264/AVC

Fatma SBIAA
University of South Brittany
Laboratory of Information Science and Technology, communication and Knowledge (Lab-STICC)
Lorient, France

Medien ZEGHID
University of Monastir
Laboratory of Electronics and Microelectronics
Faculty of Sciences
Monastir, Tunisia

Mohsen MACHHOUT
University of Monastir
Laboratory of Electronics and Microelectronics
Faculty of Sciences
Monastir, Tunisia

Sonia KOTEL
Department of Informatics
Higher Institute of Computer Science and Communication Techniques of Hammam Sousse
Hammam Sousse, Tunisia

Rached TOURKI
University of Monastir
Laboratory of Electronics and Microelectronics
Faculty of Sciences
Monastir, Tunisia

Adel BAGANNE
University of South Brittany
Laboratory of Information Science and Technology, communication and Knowledge (Lab-STICC)
Lorient, France

***Abstract*—H.264 video coding standard is one of the most promising techniques for the future video communications. In fact, it supports a broad range of applications. Accordingly, with the continuous promotion of multimedia services, H.264 has been widely used in real-world applications. A major concern in the design of H.264 encryption algorithms is how to achieve a sufficiently high security level, while maintaining the efficiency of the underlying compression process. In this paper a new selective encryption scheme for the H.264 standard is presented. The aim of this work is to study the security of the H.264 standard in order to propose the appropriate design of a hardware crypto-processor based on a stream cipher algorithm. Since the proposed cryptosystem is mainly dedicated to the multimedia applications, it provides multiple security levels in order to satisfy the requirements of various applications for different purposes while ensuring higher coding efficiency. Different performance analyses were made in order to evaluate the new encryption system. The experimental results showed the reliability and the robustness of the proposed technique.**

Keywords—component; Video coding; Data encryption; Data compression; H.264/AVC

I. INTRODUCTION

Different multimedia applications have become increasingly popular due to the fast development of communication technologies. Since communications across public networks can easily be intercepted, privacy becomes a major concern for commercial uses of multimedia communication. Encryption is an important tool for providing the security services in different fields of applications. Thus, since the 1990s, many research efforts have been devoted to the development of certain video encryption algorithms. Therefore, many algorithms have been proposed to ensure the confidentiality of video data.

Multimedia data requires either full encryption or selective encryption depending on the application requirements [1]. For example military and law enforcement applications require full encryption. However, there is a large spectrum of applications that demands a lower security level. These applications require the development of a cryptosystem using a selective encryption.

To clearly identify the characteristics of video encryption algorithms, the encryption algorithms can be divided according to their association (or not) with the video compression process. We distinguish the encryption algorithms joint compression and others independent of the compression. In fact, there are three different approaches which combine encryption and compression. As shown in "Fig. 1", an encryption algorithm could be placed before, during, or after the compression process [2][3].

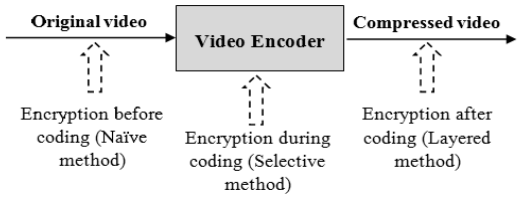


Fig. 1. Video Encryption Techniques

The video encryption algorithms placed before or after compression are called encryption algorithms independent of the compression while those executed during the compression process are called encryption algorithms joint compression.

The direct approach consists in the encryption of the entire compressed video stream using a conventional cryptographic method, such as the Advanced Encryption Standard (AES). This approach is called the naive approach. However, conventional cryptographic algorithms, which are generally designed to encrypt text data, are not well suited for video encryption because they can't treat the large volume of video data in real time. In addition, it is almost impossible to adapt

them to specific paradigms of video applications which pose specific requirements that are never encountered during the encryption of text data. These requirements are related to the efficiency of encryption, the security needs, the code conformity of the video stream, the compression efficiency, the respect for the syntax and the perception. They can be ensured using the selective encryption. In fact, this kind of encryption treats a part of the plaintext and presents two main advantages. First, it reduces the computational requirements, since only a part of plain-data is encrypted. Second, encrypted bitstream maintains the essential properties of the original bitstream. It just prevents abuse of the data. In the context of video encryption, it refers to destroying the commercial value of the video stream to a degree which prevents a satisfying viewing capability. The H.264/AVC-based selective encryption schemes have been already presented on CAVLC and CABAC [3]. These two previous methods fulfill real-time constraints by keeping the same bitrate and by generating completely compliant bitstream.

This paper presents a new selective encryption method for the H.264/AVC videos. The second section is devoted to introduce the H.264/AVC standard and the related encryption schemes. The third section will discuss the system specification, the choice of algorithms and the cryptographic techniques (scenarios). The fourth section is devoted to the design and the implementation of proposed cryptosystem. The next step is the Hardware/software validation on FPGA platform taking into account the real-time aspect.

II. H.264/AVC –Based Video Encryption

In this section, we will present the H.264/AVC video coding as well as its bit stream syntax structure. Then, we will discuss some key parameters which are imperative to design a format-compliant encryption scheme. Finally, some related works will be evaluated.

A. Overview of H.264/AVC

In terms of classification, video encryption algorithms respect in a proportional manner certain criteria such as the efficiency of encryption, the security level, the conformity to standard video codecs and the compression efficiency. The latter two are closely related to the video compression process. In fact, the Standardized video compression technologies such as MPEG-1 (ISO/IEC, 1993) [5], MPEG-2 (ISO/IEC, 2000) [6], H.261 (ITU-T, 1993) [7], H.263 (ITU-T Recommendation H.263, 1998) [8], and MPEG-4 / H.264 AVC (Advanced Video Coding) (ITU-T Recommendation H.264, 2007, ISO/ IEC, 2005) [9] are widely deployed to economically store digital video on storage devices having limited ability or to effectively communicate on networks with limited bandwidth.

Most video coding standards use hybrid coding approach that consists on compressing the video data using simultaneously the "intra" and the "inter" encoding. Although there are differences among the applied coding algorithms, compression standards are built on the same set of basic operation elements.

H.264/AVC, known also as MPEG-4 Part10, has an enormous improvement in term of the compression performance. Thus, the compressed sequence is usually 30 to 50% shorter when compared to the previous MPEG-4 Part2 standard [4]. The block diagram of the H.264/AVC encoder/ decoder is presented in "Fig. 2".

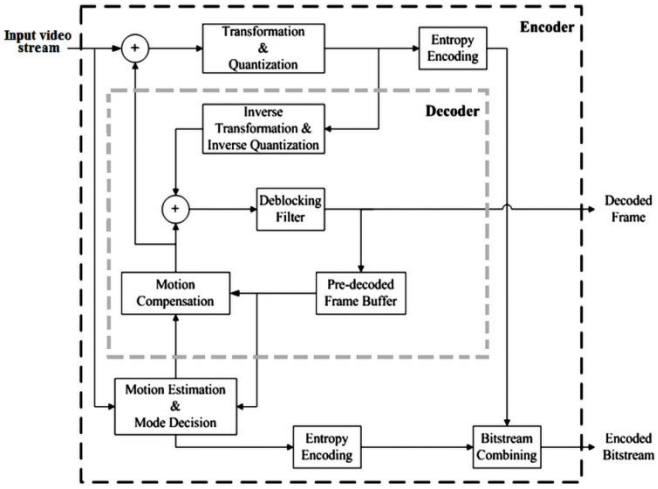


Fig. 2. Video Encryption Techniques

B. H.264/AVC Bitstream Syntax Structure

The main aim of the present research is to find a compromise between the speed of transfer and the preservation of a significant security level of multimedia data, while respecting the constraints that are imposed by the dedicated application (occupation, time, consumption ...). Accordingly, a mixed approach of encryption and compression is chosen in the present work. Thus, the cryptosystem must ensure not only confidentiality but also low power consumption and a very small occupation on FPGA. Furthermore, to ensure its integration into a compression sequence, different key parameters of the compression standard must be evaluated. This section is devoted to study the design constraints and various properties of the H.264 standard.

In fact, in a video stream, the data is presented in a hierarchical way. First, the video begins with a start code sequence (header). It contains one or more groups of pictures (GOP), and ends with an end code sequence.

The group of pictures (GOP) consists of a periodic sequence in the compressed images. In reality, there are three types of compressed images. The I-image (Intra) is compressed independently of the other pictures. The P-image (predictive) is coded using prediction of a previous image of type I or P. Finally, the B-images (Bidirectional) are encoded by double prediction using as reference a previous and next image of I or P type. A group of pictures starts with an I-frame, contains a periodic sequence of P-frames separated by a constant number of B-frames (see "Fig. 3") [8][9].

Fig. 3. The structure of a GOP

A GOP structure is defined by two parameters. These are the number of images and the distance between I-images and P-images. In fact, an I-image is inserted every 12 frames.

An image consists of three matrices where each matrix element represents a pixel. The YUV model defines a color space with three components. The first is the luminance and the others present the chrominance. The U and V matrices have smaller dimensions than the matrix Y (relatively to the used format). The most important information of the picture is stored in the matrix Y [8][9].

The image is cut in slices whose purpose is to limit the errors propagation in image transmission/storage. A slice is a sequence of macros blocs. A macro-block represents a portion of the image of 16 × 16 pixels size. A block is a 4 × 4 matrix of coefficients each one represents one of the three components of a pixel, Y, U or V [8][9].

"Fig. 4" below describes the hierarchical aspect of a video sequence from the GOP to the 4x4 blocks.

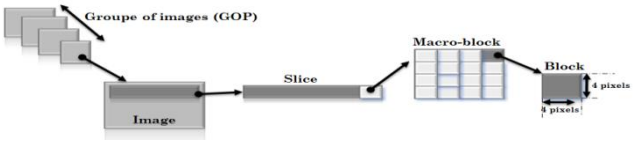


Fig. 4. Data hierarchy in a video stream

C. Key Parameters for a Selective Video Encyption

The process of video compression involves three processes: Discrete Cosine Transform (DCT), quantization and coding. To achieve the best choice of the location of the designed cryptosystem in the chain of compression, it is indispensable to take in consideration the execution time, the level of security, and the complexity of the system.

Observing the structure of a video encoder, we realize that if the proposed cryptosystem is placed after the DCT transformation, a decryption system is needed to be added in the decoder which aims to build on the temporal redundancies of a video streaming. The principle is to predict the content of an image and to encode only the error made in this prediction. Thus the existence of a cryptosystem increases the processing time and affects the complexity of the encoder. However, a cryptosystem inserted after the quantization step will not require an additional time for a decryption process.

In fact, the DCT is used to move the spatial domain to the frequency domain and also to collect as much information as possible in a small number of frequency coefficients. The DC coefficient shows the average of samples processed and presents the most important details in the raw of an image (lower spatial frequency). The AC coefficients represent the fine details of the image (higher spatial frequencies) [10].Thus, the DC coefficients carry more useful information than the high frequency components. Moving away from DC components of the image, not only the coefficients tend to have low values, but also, they become less important for the description of the image.

"Fig. 5" shows that the number of the DC coefficients represent (1 / 16) of all coefficients in a macro-block that contains 24 DC coefficients and 384 AC coefficients. Therefore, DC coefficients of an image I present (1 / 192) of the total coefficients. In consequence, if we assume that TG represents the required time to encrypt a video stream, hence the required time to encrypt only the I-frames of this flow will be reduced to (TG / 12) while maintaining a considerable security level. Moreover, if only the DC coefficients of I-frames are encrypted the required time for encryption process will be (TG / 192).

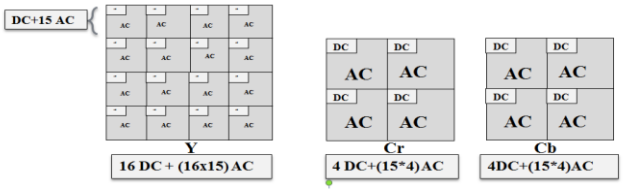


Fig. 5. The structure of 4 : 2 : 0 macro-blocks

Before defining the encryption scenarios, it is required to know the maximum number of different types of coefficients processed per second. It helps us to choose the most appropriate cryptographic algorithm. In this context, all the necessary calculations for the design of the proposed cryptosystem were performed. The following table I summarizes the obtained results.

TABLE I. KEY PARAMETERS FOR VIDEO ENCRYPTION

Group	max size of one image (mb)	DC coef Nbr in an image	AC coef Nbr in an image	max I images per seconde	max DC coef Nbr of I images (s)	max AC coef Nbr in I images (/s)	max DC coef Nbr (/s)	max coef Nbr (/s)
1	99	2376	35640	4	9504	142560	35640	570240
1b	99	2376	35640	4	9504	142560	35640	570240
1.1	396	9504	142560	2 or 9	19008 or 85536	285120 or 1283040	72000	1152000
1.2	396	9504	142560	6	57024	855360	144000	2304000
1.3	396	9504	142560	6	57024	855360	285120	4561920
2	396	9504	142560	6	57024	855360	285120	4561920
2.1	792	19008	285120	6	114048	1710720	475200	7603200
2.2	1620	38880	583200	5	194400	2916000	486000	7776000
3	1620	38880	583200	5	194400	2916000	972000	15552000
3.1	3600	86400	1296000	5	432000	6480000	2592000	41472000
3.2	5120	122880	1843200	4	491520	7372800	5184000	82944000
4	8192	196608	2949120	4	786432	11796480	5898240	94371840
4.1	8192	196608	2949120	4	786432	11796480	5898240	94371840

4.2	8704	208896	3133440	4	835584	12533760	12533760	200540160
5	22080	529920	7948800	5	2649600	39744000	14155776	226492416
5.1	36864	884736	13271040	5	4423680	66355200	23592960	377487360

D. Review of the Related Work

In this section, we will describe the currently known encryption algorithms for MPEG video streams in order to evaluate them with respect to three metrics: security level, encryption speed, and encrypted MPEG stream size.

In fact, several selective encryption schemes have been previously discussed in the recent past. In [11], an efficient encryption system for the H.264/Scalable Video Coding (SVC) codec is presented. The proposed selective encryption scheme is suitable for video distribution to users who have subscribed to differing video qualities on medium-to-high computationally capable digital devices. Another idea of a selective encryption on SCV is proposed in [12]. It involves the encryption of signs of coefficients, sign of motion vectors, and the alteration of DC values to ensure three different security levels. Although the sign encryption has no effect on the compression efficiency and the bitrate, the alteration of the DC values changed the video statistics and affected the compression efficiency.

In [13], the proposed scheme encrypts the video by scrambling the Intra-Prediction Mode (IPM) of intra macro-blocks. The main limit of this scheme is that it offers less security level due to the length of the pseudo number sequence. In [14], two fold video encryption techniques applicable to H.264/AVC are presented. In fact, the authors proposed an encryption of the DCT coefficients which affects the statistical characteristics of data. In addition, the compression ratio is affected which consequently increases the bitrate.

This paper proposes a combination of pseudo-random key generator and permutation code algorithm. The main objective is to enhance the security of H.264 video. In the next section, the proposed scheme is discussed in detail along with the generation of pseudo-random keys.

III. The Proposed Selective Encryption Scheme

The purpose of this work is the design of a cryptographic processor mainly dedicated to multimedia applications. The obtained cryptosystem will be placed on a prototyping platform based on FPGA to encrypt video transmissions in real-time conditions. In this context, the H.264 AVC part 10 standard is chosen. It is defined in most multimedia applications such as video conferencing, Internet video, media players, video mobile, and some satellite channels.

The design of the cryptosystem can be studied in two directions: The first one consists on proposing cryptographic protocols that should be appropriate for applications presenting time and security constraints. In the second direction, it is essential to realize the implementation of the system in a compression sequence that presents the constraints of the target application.

A. Design Flow

Designing systems with high architecture performance requires the choice of the most appropriate algorithms. Similarly, the definition of the design flow from functional level to physical level is a crucial step. It greatly affects the time of conception and the realization of the target system.

The proposed design flow is based on five strategic points. First, the definition of the requirements and the specification of the encryption techniques is an important step that consists on setting the goals of the project and studying the various constraints. The latter are related to target applications in order to ensure the conception coherence. Secondly, according to the study of the constraints imposed by the target applications, different cryptographic protocols will be proposed in order to achieve a hierarchy of security levels. Then, modeling the security IP requires architectural optimizations in order to adopt the cryptosystem to both application needs and used platform. Fourth, the logic synthesis and the performance evaluation of the designed cryptosystem ensure the validation of the proper functioning of the IP under real- time constraints. Finally, the hardware/software validation (Co-simulation) of the proposed cryptosystem verifies the architecture of the final prototype in a hardware environment. This will enable us to achieve real-time evaluation of system performance in terms of execution time and throughput. The tools provided by the reconfigurable platform and the electrical measurements allow us to evaluate the energy consumed by the proposed cryptosystem.

B. Proposed Cryptographic Scenarios

As mentioned before, encrypting the entire video is not always reasonable. This is mainly due to the large size of videos. Thus this kind of encryption approach is not recommended for embedded systems where the energy capabilities are limited. In such cases, saving time and energy consumption becomes an important issue. Hence, a selective encryption is compulsory. Accordingly, in this paper, four different encryption scenarios were proposed. They consist on encrypting only the most important data. In order to deal with the constraints of a real-time transmission, the least significant information will be switched while the most important data will be encrypted using a sufficiently secure algorithm. Therefore, the proposed scenarios are described below:

- The first scenario consists in encrypting the DC coefficients of the I-frames using an algorithm A. As shown previously, the images I carry the most useful information of the video stream. Hence, this scenario guarantees a high security level.
- The second scenario encrypts the I-frames. Thus, the DC coefficients of the I-frames are encrypted using an algorithm A while the AC coefficients are enciphered using an algorithm B. Therefore, this scenario has greater security level compared with the first scenario although it requires more execution time.
- The third scenario encrypts all the DC coefficients in the video stream using an algorithm A. Since the DC coefficients present the most important information of an image, this scenario provides a better security level.

- The fourth scenario consists in the encryption of the DC coefficients of all the images by an algorithm A and the AC coefficients of the I-frames by an algorithm B. This scenario provides a very high security level. However, it needs much execution time due to the large number of coefficients to be treated.

The table II summarizes the different proposed scenarios. It illustrates the speed, the security level, and the influence of encryption on the compression rate.

TABLE II. THE PROPOSED ENCRYPTION SCENARIOS

Scenarios	Treatments	Security level	Required execution time	Influence on the compression ratio
Scenario1	Only the DC coefficients of the I-images are encrypted.	**	*	*
Scenario2	The DC and AC coefficients of the I-images are encrypted	***	***	***
Scenario3	The DC coefficients of all the images are encrypted.	*****	**	**
Scenario4	The DC coefficients of all the images and the AC coefficients of the I-images are encrypted	******	****	****

Since the influence of the encryption on compression ratio depends only on the quantity of the encrypted data, the choice of encryption algorithms does not affect this parameter. However, while selecting the encryption algorithms, it is indispensable to take in consideration the coefficient nature and the desired security level which affect the encryption time and the compression ratio. Thus, in order to respect the constraints imposed by the characteristics of different levels and profiles, the choice of the encryption algorithms (A and B) must consider the speed of processing. Therefore, it is to guarantee a balance between the speed, the compression ratio, and the security level.

The table III shows the minimum speed needed to ensure the application of different scenarios. The minimum speed required for each treatment is equal to the maximum number of coefficients to encrypt multiplied by the size of a single coefficient (in bits).

TABLE III. THE MINIMUM REQUIRED SPEED FOR THE TREATMENT OF EACH SCENARIO

Scenarios	min speed required for the treatment (Mbit/s)
Scenario 1	53.084160
Scenario 2	849.346560
Scenario 3	283.115520
Scenario 4	1079.377920

C. Choice of the Encryption Algorithms

While encrypting a video stream, the transmission speed is a fundamental criterion. Therefore, the symmetric key algorithms are suggested to be used. In fact, the main disadvantage of asymmetric algorithms is that their treatment is slow. In addition, they require a lot of calculation. Therefore, their use becomes impossible for real-time applications. Concerning security, they present problems related to the structure of the public key systems. In fact, to ensure adequate security, the generated keys are larger in size compared to the symmetric key.

The main types of private key cryptosystems used today can be classified into two categories. These are the block ciphers that treat data blocks of fixed size and the stream ciphers that treat the data bit by bit. For the block cipher, good security is defined by a long key. This implies some drawbacks. In fact, the large blocks are safer but are heavier to implement. However, stream ciphers are very fast. The hardware implementation of the latter needs few gates, so they are suitable for real-time applications and often used to protect multimedia data. Generally, they are presented as a generator of pseudorandom numbers. A bit XOR is operated between the generator output and a bit from the data. However, the XOR is not the only operation possible.

In order to choose the appropriate key generator, a comparison between the most known stream ciphers has been made. We synthesized using the "Synplify Pro" component packages and the target component Virtex2 XC2v2000-6ff896. The table IV below summarizes the obtained results.

TABLE IV. COMPARISON BETWEEN PSEUDO-RANDOM GENERATORS

ciphers	Key size (bits)	Initialization Vector size(bits)	Frequency (MHz)	Occupation (Luts)	Consumption (mW)
A5/1	64	114	250.376	110	46.33
W7	128	128	188.590	777	111.77
CA 16×16	256	16	308.550	683	52.75
Grain-80	80	64	230.9	355	13.72
Grain-128	128	96	238.5	495	19.22

According to the table IV, we note the following observations:

- A5/1 has an acceptable speed and occupation rate (2%), and a relatively low consumption ratio. These results justify the use of this generator in GSM applications.
- The W7 frequency is the lowest. Whereas, its period is greater than that obtained by the other generators. Thus, it ensures a good security level.
- Grain consumption is the least compared to the other pseudo random generators. The frequency and the occupation values are acceptable for the real time

applications. However, its security level has to be checked.

Thus, randomness is very important to evaluate the quality of the generated keys. It presents one of the most critical points of configuring a crypto processor. In fact, to test quantitatively the randomness of the generated keys, the National Institute of Standards and Technology (NIST) announced, in 2001, a standard called FIPS 140-2. It covers four types of tests, namely, Monobit test, frequency test, Runs test and Longest test runs. A sequence is considered to be random if the probability P-value for each test is greater than 1% (0.01). The results of the various tests applied to the algorithms A5/1, W7, CA and Grain are presented in the following table V.

TABLE V. SECURITY TESTS OF PSEUDO-RANDOM GENERATORS

	Monobit test	**Frequency test**	**Runs test**	**Longest run test**
A5/1	$0.0026\ 2^{64}$	$0.0028\ 2^{64}$	$0.0049\ 2^{64}$	$0.0021\ 2^{64}$
W7	$0.0022\ 2^{1024}$	$0.0016\ 2^{1024}$	$0.0046\ 2^{1024}$	$0.0025\ 2^{1024}$
CA	$0.0025\ 2^{256}$	$0.0018\ 2^{256}$	$0.0045\ 2^{256}$	$0.0010\ 2^{256}$
Grain Standard Version	$0.0109\ 2^{1024}$	$0.0101\ 2^{1024}$	$0.0131\ 2^{1024}$	$0.012\ 2^{1024}$

The obtained results show that Grain provides a higher security level compared to the other well-known ciphers such as A5/1, W7, and CA. Grain provides higher security while maintaining a small hardware complexity. Accordingly, grain-80 will be used to ensure the key generation in the proposed cryptosystem.

As mentioned before, the cryptosystem will be integrated after the quantification step. Therefore, the AC and DC coefficients, resulting from the quantization step will be treated with appropriate systems and crypto-coded in order to achieve a crypto-compressed video. Thus, the DC coefficients will be encrypted using the key generated by Grain-80 while the AC coefficients will be switched. Fig. 6 illustrates the new Crypto-Compression Process.

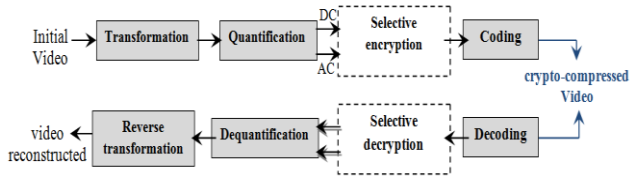


Fig. 6. Crypto-Compression Process

IV. DESIGN AND IMPLEMENTATION OF THE PROPOSED CRYPTOSYSTEM

From the system specification and cryptographic techniques, developed previously, it results the selection of the appropriate cryptographic algorithms as well as the location of the proposed cryptosystem in the compression process.

The implementation of the designed system is based on the complementarily of four different blocks. These are the Algorithm A (key generator: Grain-80), the configuration processor, the encryption processor, the re-configuration unit, and the permutation tables.

This structure allows for a good distribution of tasks between the blocks so that the proposed system can be adaptable to various applications. First, the key generation algorithm A and the permutation tables are defined with respect to the need. In addition, the function performed by the Encryption-module can be easily modified.

"Fig. 7" shows the general structure of the proposed system. In order to achieve the scenarios described above, Grain has been chosen as encryption algorithm to process the DC coefficients. The AC coefficients will be swapped using predefined permutation tables.

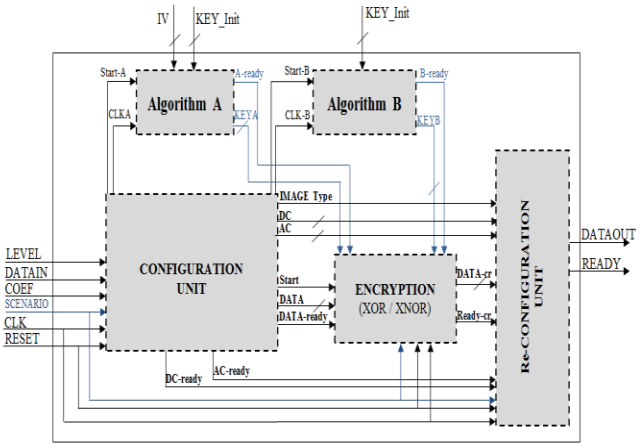


Fig. 7. The structure of the poposed cryptosystem

A. Grain Implementation

Grain is a stream cipher algorithm that appeared in 2005. It is designed to be very small and efficient in material implementation [15]. Gain family currently consists of two types of encryption. The first uses a key of 80 bits while the other uses a 128-bit key. Grain uses two registers. These are the LFSR (Linear Feedback Shift register) and the NFSR (Nonlinear Feedback Shift register). The output result is generated through a non-linear filter that takes two inputs of the shift registers. The following figure "Fig. 8" describes the structure of the Grain Stream Cipher.

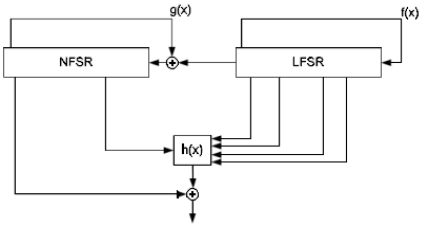


Fig. 8. The Grain cipher

The implementation and simulation of the Grain algorithm was achieved in VHDL. The Key Initialization phase ensures the initialization of the cipher using the initial key and the init-IV vector. This step is crucial before generating the key stream.

Grain is intended to be used in environments where gate counts and the power consumption as well as the memory needs to be very small. In fact, several ciphers are designed

with better software efficiency compared to Grain. In fact, they are more appropriate when high speed in software is required.

In reality, the basic implementation has 1 bit/clock rate. The speed of a word oriented cipher is typically higher since the rate is 1 word/clock. Grain is a bit oriented cipher but it has compensated this problem by the possibility to increase the speed. Accordingly, a designer could choose the appropriate speed of the cipher according to the amount of hardware available. The following "Fig. 9" illustrates the cipher process when the speed is doubled.

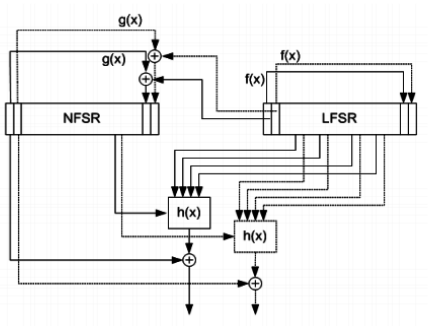


Fig. 9. The cipher process when the speed is doubled

We implemented all the possible versions of Grain-80 in order to choose the appropriate speed and performances for the target application. The synthesis results presented in table VI proved that the speed changes proportionally to the occupancy. In addition, the consumption ratio becomes increasingly significant from one version to another. For example from the standard version of Grain to Grain-16 version (where the speed is multiplied by 16), the change in consumption is negligible compared to the evolution of the speed (≈7x230.9 = 1652.8 Mb/s). The generic version gives the opportunity to choose the version that is compatible with dedicated applications, but it has a loss in speed, frequency and occupation. For example, compared to the original version the frequency of the generic version (with N=1) decreases from 230.9 to 39.9 MHz, while the occupancy reaches a value equal to 4533 Luts (≈12x 336).

TABLE VI. SYNTHESIS RESULTS OF GRAIN STREAM CIPHER 80

Grain Version	Frequency (MHz)	Occupation (Luts)	Consumption (mW)	Throughput (Mbps)
1	230.9	336 (<1%)	13,72	205,1
2	167,2	369	13,01	334,4
4	154,7	424	13,87	618,8
8	144,8	562	14,81	1158,4
10	148,3	664	14,84	1483
16	101,1	1035	18,58	1617,6
N	39,9	4533(9%)	15,95	(39,9*n)

According to these results, it is clear that each version has its own characteristics. Thus, choosing the appropriate version is based on the constraints of the target application. The proposed cryptosystem is dedicated to the real-time video application. Thus, the version Grain-V4 was chosen where the speed is multiplied by four.

B. Configuration and Re-Configuration Units

The scenarios configuration and assignment are carried out by the configuration module. It ensures three important functions. These are the level identification, the scenario specification, and the classification of images and coefficients. "Fig.10" illustrates the process of this unit.

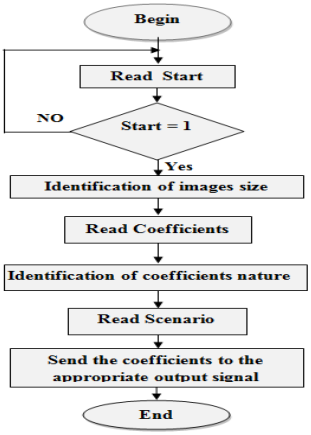


Fig. 10. The configuration unit process

The configuration module is fundamental in order to ensure synchronization between the other modules. Similarly, the reconfiguration module's role is to restore the flow of input coefficients and to reconstitute the encrypted video streaming.

C. Encryption Unit

This block is responsible for performing an XOR or an XNOR operation between the key generated by Grain-V4 and the coefficients to be encrypted (in case of DC coefficients). The AC coefficients are swapped using predefined permutation tables.

The encryption key generated by Grain is 80-bit size. Therefore, it can serve to encrypt 6 different DC coefficients. To improve the robustness of the proposed cryptosystem, two different functions were chosen to be performed. These are the XOR and the XNOR.

Since Grain takes 20 clock cycles to manage its first key, it is needed to manage the first coefficients reaching this block before the generation of the cipher key. However, after 20 clock cycles only 2 DC coefficients and 18 AC coefficients are ready for encryption. Thus, two registers have been defined to ensure this task.

D. Permutation Unit

As previously mentioned AC coefficients are switched following permutation tables that were defined for this purpose. Only 16 permutation tables were chosen to meet the design requirements. First, it is important to reduce the used memory in order to consume less in terms of occupation. Secondly, the key generated by GRAIN can be used to define only 6 different addresses (if the number of tables increases, more than 4 bits will be needed to represent the table number). In fact, 50 different tables were generated (based on Grain keys). Then, four different cryptographic tests were applied in order to evaluate the cryptographic properties of the generated tables. These are the nonlinearity, the strict avalanche criterion (SAC), bijection, and the BIC (output bits independence criterion). In fact, the generated tables satisfy the requirement of bijectivity since they have different output values. In

addition, the average value of nonlinearity of the 16 generated tables is equal to 102. Furthermore, the mean value of the dependence matrix (SAC) of the chosen tables is equal to 0.5281 which is very close to the expected value 0.5. All these results justify the choice of the used permutation tables.

The following "Fig. 11" illustrates how the Grain key is used to choose the permutation table for the encryption process. In fact, the same table cannot be used to encrypt two successive blocks of data.

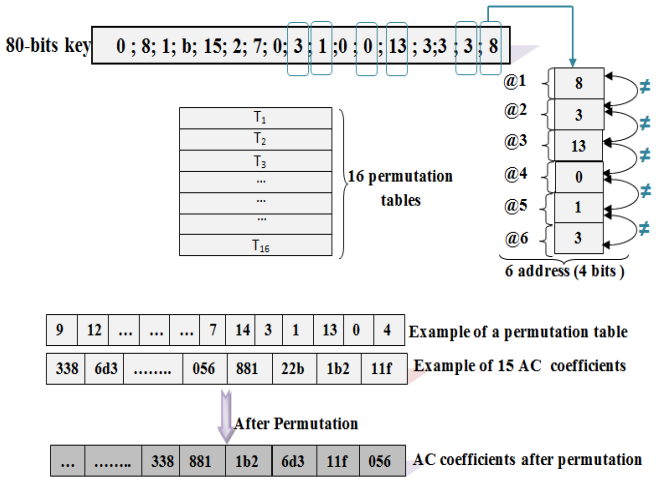


Fig. 11. The choice of the permutation tables

E. Synthesis of the Proposed Cryptosystem

Synchronization between system units is an imperative operation. In fact, the management of the clock has a fundamental role in system performance such as its total consumption. In this context, the Grain cipher is activated all throughout the treatment, although it is used only for specific times to encrypt DC coefficients. This gave us the idea to design a second version in order to optimize the used resources.

The Grain process was examined in order to be activated only when a key is needed. The management of the activation and deactivation of this generator allows us to use all the produced keys and to benefit of the provided security level. In the same context, the "Encryption" block can be activated in need. To manage the activation and deactivation of these two blocks, we used a clock generation processor which was implemented in the configuration block.

Moreover, different improvements have been carried out in order to optimize the used resources in the proposed cryptosystem. To evaluate the impact of these modifications, the synthesis of the proposed cryptosystem was performed using the component packages "Synplify Pro 9.6" and the target component Virtex5-XC5VLX50-FF676. The obtained results are presented in table VII.

TABLE VII. SYNTHESIS RESULTS OF THE DIFFERENT UNITS OF THE PROPOSED CRYPTOSYSTEM

	Configuration Unit	Grain V4	Encryption unit	Reconfiguration Unit
Occupation (Luts)	256 (1%)	282 (<1%)	157 (<1%)	28 (<1%)
Frequency (Mhz)	155.1	623.6	348.8	532.3

It is clear that the occupation of the different blocks is very small. This is due the division of labors and the use of procedures. For example, the synthesis of block "Encryption" gave the value 18 262 LUTs (95%) as occupation. However, after using the proposed modifications, the occupation has become equal to 157 LUTs (<1%). Furthermore, the frequency increases from 102.8 MHz to 160.3 MHz due to the proposed improvements. The following table VIII summarizes the obtained results.

TABLE VIII. SYNTHESIS RESULTS OF THE PROPOSED CRYPTOSYSTEM

	Optimized Cryptosystem	Original Cryptosystem
Occupation (Luts)	722(2%)	18 827(97%)
Frequency (Mhz)	160.3	102.8

To conclude, we can claim that the optimized cryptosystem has good performance in terms of occupation, frequency and consumption. It increases with the amount of information processed and the complexity of the applied scenarios. The results justify the implementation the optimized version in the validation phase.

V. THE HARDWARE/SOFTWARE VALIDATION

The objective of this section is to check that the hardware and software specifications are valid. This involves testing and studying the evolution of the cryptosystem in the presence of environmental constraints such as the throughput, the implementation costs, and the execution time. The verification includes the examination of the running of the designed system. In fact, it can be simulated and tested at the behavioral level through an ordinary simulation tool such as "ModelSim". Then, the obtained synthesizable IP (Intellectual Property) can be frozen in hardware (FPGA or ASIC). Accordingly, the implementation of the proposed cryptosystem on the reconfigurable platform will allow for an assessment of the occupied area as well as the real-time constraints.

Since, the integration of an IP Core in a real-time hardware design is a complex task; an efficient methodology for the real-time implementation on a reconfigurable platform is required. In fact, the flow consists in developing and synthesizing the appropriate IP to be integrated through the Xilinx System Generator tool in the EDK flow which is used to transform the RTL implementation into a complete FPGA implementation [16-18]. Once the IP is valid, it is integrated and exported as a PCORE to the Platform Studio Project. Finally, the communication between the Micro Blaze processor and the PCORE has to be made. It provides a Hardware/Software Co-Simulation environment to test the embedded system design. This communication often occurs over shared bus connectivity.

The following "Fig.12" illustrates the conception flow of a real-time design for the proposed cryptosystem.

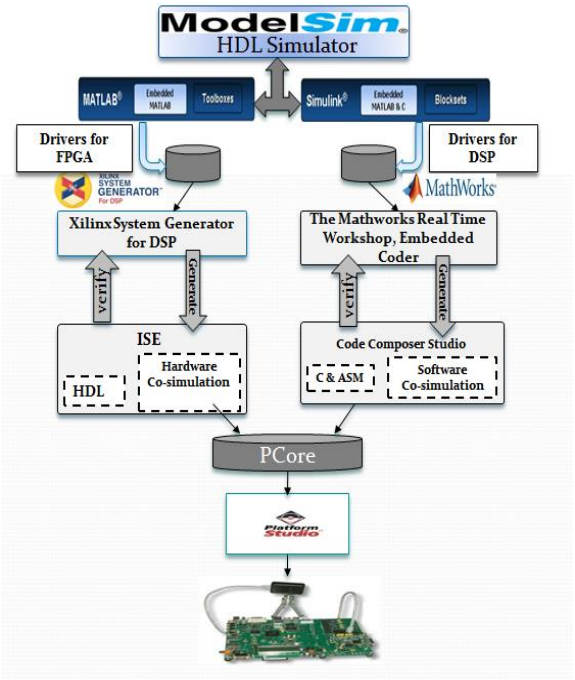


Fig. 12. The choice of the permutation tables

The System Generator provides a hardware co-simulation to incorporate an architecture running on the FPGA directly in a Simulink simulation. The video model tested and verified in the previous step, must be compiled for hardware co-simulation. The selection of the target platform for the compilation must be made. In fact, Spartan 3A DSP 3400 Platform offers us the opportunity to implement and verify the hardware implementation results.

A. Integration of the Proposed Cryptosystem in the H.264 Encoder

Zexia provided H.264 encoder implemented in VHDL [20]. It is designed as a modular system with small and efficient components using low power ressources. The proposed cryptosystem was integrated in the Zexia-H.264-encoder in order to validate its process. The following figure 13 shows the structure of the obtained crypto-encoder.

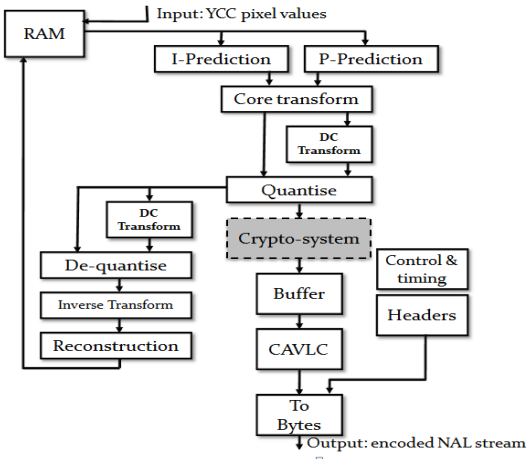


Fig. 13. Architecture of the new crypto-compression system

The proposed cryptosystem is adapted in order to be integrated into the compression process. "Fig.14" shows the simulation results of the obtained crypto-compression system. It presents the major signals of the different blocks when the fourth scenario is applied to encrypt the video stream.

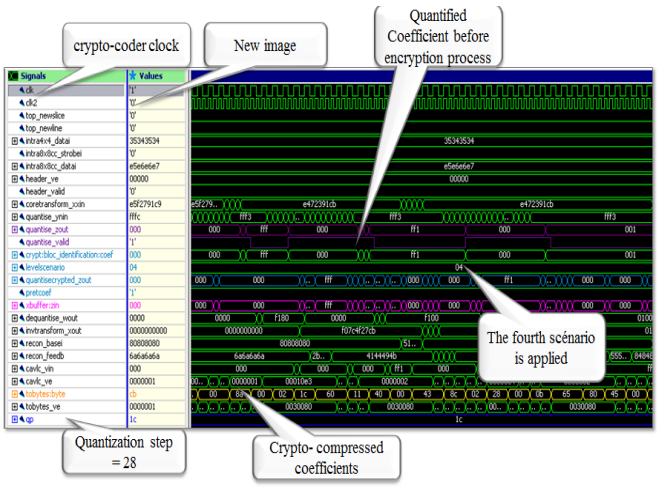


Fig. 14. Simulation results of the optimized crypto-compression system

B. Integration of cryptosystem model of Camera Design

This section presents the integration of the proposed cryptosystem, developed in VHDL, using the System Generator Black Box in the model of camera design. In fact, the reference design was used. It includes a VSK-Camera-VOP Bayer filter to restore the image in RGB format. The generated PCORE is exported as a new EDK-PCORE in the proposed project. The design shown in "Fig.15" consists of the Starter Kit video (VSK) Spartan 3A DSP FPGA XCSD3400A. This card is used to decode the data that came through the serial port interface LVDS Camera.

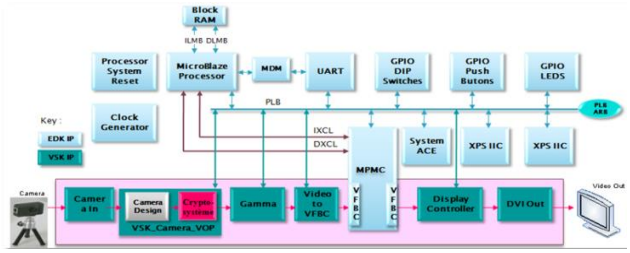


Fig. 15. Architecture of the integration of hardware cryptosystem in the Design of Camera Frame Buffer

"Fig.16" shows the external structure model VSK-Camera-VOP and "Fig.17" details its internal structure.

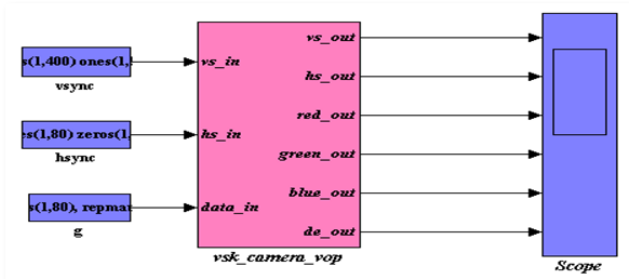


Fig. 16. external structure of VSK-Camera-VOP model

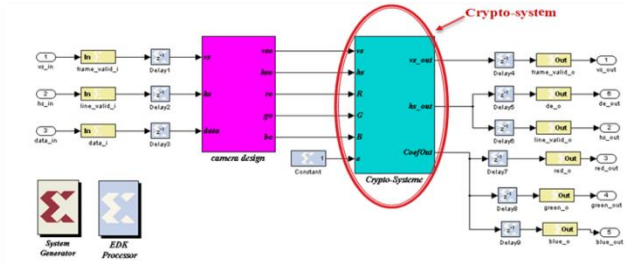
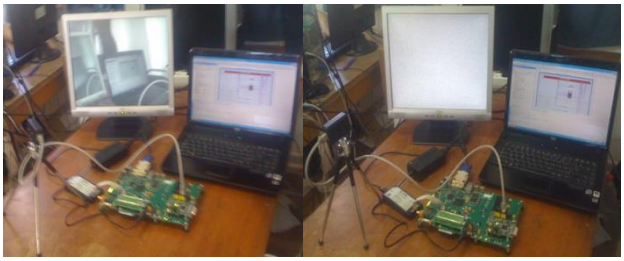


Fig. 17. internal structure of VSK-Camera-VOP model

C. Real time validation on Spartan 3A DSP platform

In the Hardware Co-simulation of real-time cryptosystem, the string contains the entire cycle of acquisition, processing and retrieval of a video signal from a video source (camera). The results of the Hardware Co-simulation presented in the following "Fig.18", allow us to verify the efficiency and the robustness of the proposed model HDL.



(a) original video (b)encrypted video

Fig. 18. Real time validation on Spartan 3A DSP platform

Image processing in real time requires the use of fast electronic circuits that are capable of handling large amounts of information generated by the video source. That's why FPGAs are ideal for this kind of application.

D. Security analysis

In order to analyze the security of the proposed cryptosystem against most known attacks, security tests were conducted on Foreman video (352x288, 164 frames). Then, the entropy values, the PSNR (Peak Signal-to-noise ratio), and the Horizontal and vertical correlation coefficients were observed.

- The correlation provides a quantitative representation of the similarities between the original and the encrypted frames. In fact, low correlation coefficient indicates that there is less similarity between the original and encrypted video, which shows the efficiency of the encryption scheme.
- The PSNR is the most widely used metric to estimate image distortion measure. This metric compares the visual quality between the plain image and the ciphered one. The PSNR is based on the Mean Squared Error value (MSE) that delivers the error between two images.
- The information entropy is one of the most important features of randomness. In fact, the source is considered to be truly random if the information entropy of the ciphered image is close to eight.

The following table IX presents the different analysis results. They justify the efficiency of the proposed cryptosystem.

TABLE IX. SECURITY ANALYSIS OF THE DIFFERENT PROPOSED SCENARIOS

	Horizontal correlation	**Vertical correlation**	**PSNR value (dB)**	**Information Entropie**
Scenario 1	0.0883	0.2201	16.42	7.4928
Scenario 2	0.0844	0.0974	14.8842	7.5580
Scenario 3	0.0732	0.0824	12.5028	7.5595
Scenario 4	0.0585	0.0778	9.9642	7.6956

VI. CONCLUSION

In this paper, a new cryptosystem dedicated for multimedia applications is proposed. It is designed to be integrated into the H.264 encoder. It provides four different encryption scenarios. The proposed structure is essentially based on a pseudo-random generator, a configuration unit and an operator performing an XOR/XNOR between the generated keys and the appropriate data which are identified by the configuration processor. This operator is also responsible, of the data swapping based on highly nonlinear permutation tables.

The choice of cryptographic algorithms was based on the study of environmental constraints imposed by the targeted applications such as the real-time transmission, the speed, the influence on the compression ratio and the desired security level. Hence, Grain-80-V4 was chosen to encrypt the DC coefficients which have the most important information of the video stream. The permutation was elected to encrypt the AC coefficients that are more numerous than the DC coefficients.

In order to deal with the real-time multimedia applications, we chose the joint compression and encryption approach that does not require too much time for encryption/decryption

process while maintaining a considerable amount of compression ratio.

Several perspectives emerge as a result of the present research. In fact, it is important to study the resistance of the proposed cryptosystem against certain types of attacks such as the fault injection attacks. Appropriate counter-measures should be proposed if necessary. In addition, the chaos-based selective encryption is a new and an efficient approach used for the multimedia application. It is attracting an increasing research effort due to its favorable properties such as the good pseudo randomness and the high sensitivity to the initial values.

REFERENCES

[1] B. Furht, D.Kirovski, ''Multimedia Security Handbook'', CRC Press LLC in December 2004.

[2] S. Lian, Multimedia Content Encryption: Techniques and Applications (Taylor & Francis Group, LLC, 2009).

[3] Z. Shahid, M. Chaumont, and W. Puech, "Fast protection of H.264/AVC by selective encryption of CAVLC and CABAC for I and P frames," IEEE Trans. Circuits Syst. Video Technol., vol. 21, no. 5, pp.565-576, May 2011.

[4] Li, Y., Liang, L., Su, Z., Jiang, J., "A new video encryption algorithm for H.264", Fifth International Conference on Information, Communications and Signal Processing (ICICS), pp. 1121–1124, IEEE 2005.

[5] ISO/IEC 11172-2. Coding of moving pictures and associated audio for digital storage media at up to about 1.5 Mbit/s: Video, 1993.

[6] ISO/IEC 13818-2. Information technology–generic coding of moving pictures and associated audio information: video, 2000.

[7] ITU-T Recommendation H.261. Video code for audiovisual services at px64 Kbit/s, March 1993.

[8] ITU-T Recommendation H.263. Video coding for low bit rate communication, Feb. 1998.

[9] ISO/IEC 14496-10. Information technology–coding of audio–visual objects–Part 10: advanced video coding, 2005.

[10] Richardson, E.G., "H.264 and MPEG-4 Video Compression–video coding for nextgeneration multimedia", 1st ed. John Wiley & Sons, New York, 2003.

[11] M. Asghar, M. Ghanbari, M. Fleury, and M. Reed, "Efficient Selective Encryption with H.264/SVC CABAC Bin-Strings", 19th IEEE International Conference on Image Processing, IEEE,2011.

[12] G.B. Algin, and E.T. Tunali, "Scalable video encryption of H.264/AVC codec," J. of Visual Commun. and Image Representation, vol. 22, no. 4, pp. 353-364, 2011.

[13] Siu-Kei Au Yeung et al, "Partial Video Encryption Based on Alternating Transforms", IEEE Signal Processing Letters, vol. 16, No. 10, pp. 893–896, October 2009.

[14] T.Chattopadhyay and Arpan Pal, "Two fold video encryption technique applicable to H.264 AVC", IEEE International Advance Computing Conference, pp. 785 – 789, March 2009.

[15] S.S. Mansouri, E. Dubrova, "An Improved Hardware Implementation of the Grain Stream Cipher", 13th Euromicro Conference on Digital System Design: Architectures, Methods and Tools (DSD), pp. 433–440, IEEE 2010.

[16] A. Rukhin, J. Soto, J. Nechvatal, M. Smid, E. Barker, S. Leigh, M. Levenson, M. Vangel, D. Banks, A. Heckert, J. Dray, San Vo, "A statistical test suite for random and pseudo-random number generator for cryptographic applications", NIST Special Publication 800-22.

[17] Xilinx, "Embedded System Tools Reference Manual", EDK 10.1, Service Pack 3, 19 September 2008, www.xilinx.com/support/documentation/sw_manuals/edk10_est_rm.pdf.

[18] Xilinx, "System Generator for DSP Getting Started Guide", UG639 (v11.4), December 2, 2009, www.xilinx.com/support/documentation/sw_manuals/xilinx11/sysgen_gs.pdf

[19] Xilinx, "Spartan-3A DSP FPGA, Video Starter Kit" UG456 (v2.1) March 15, 2010, www.xilinx.com/support/documentation/boards_and_kits/ug456.pdf.

[20] Zexia Access Ltd © 2008 - H.264 Hardware Encoder: http://hardh264.cvs.sourceforge.net/viewvc/hardh264/hardh264/src

Off-Line Arabic (Indian) Numbers Recognition Using Expert System

Fahad Layth Malallah
Faculty of Computer Science, Cihan University
Sulaimaniya, Iraq

Mostafah Ghanem Saeed
Faculty of Computer Science, Cihan University
Sulaimaniya, Iraq

Maysoon M. Aziz
Department of Mathematics, College of Computer Sciences and Mathematics, University of Mosul
Mosul, Iraq

Olasimbo Ayodeji Arigbabu
Department of Computer and Communication Systems Engineering, Faulty of Engineering, Universiti Putra Malaysia, Serdang Malaysia

Sharifah Mumtazah Syed Ahmad
Department of Computer and Communication Systems Engineering, Faulty of Engineering, Universiti Putra Malaysia, Serdang Malaysia

***Abstract*—This paper proposes an effective approach to automatic recognition of printed Arabic numerals which are extracted from digital images. First, the input image is normalized and pre-processed to an acceptable form. From the preprocessed image, components of the words are segmented into individual objects representing different numbers. Second, the numerical recognition is performed using an expert system based on a set of if-else rules, where each set of rules represents the categorization of each number. Finally, rigorous experiments are carried out on 226 random Arabic numerals selected from 40 images of Iraqi car plate numbers. The proposed method attained an accuracy of 97%.**

***Keywords*—Arabic numeral character recognition; Image Processing; Pattern Recognition; Feature Extraction; Object Segmentation; Expert System**

I. INTRODUCTION

Automatic character recognition is becoming very important in many practical applications such as postcode identification and car plate number recognition. A traffic police officer may want to document the license plate numbers of approaching vehicles. Manually performing this task would obviously be very laborious, and incur significant amount of time. Conversely, an automated process that involves the application of a camera to capture the plate numbers and recognize them using a predictive model, would not only be beneficial in terms of computational time, but also ease the amount of human effort required for such task. Furthermore, these systems can be used for surveillance or monitoring of specific events using numbers.

However, most research studies in this area are mainly concentrated on Latin character recognition. In this paper, an effective method for Arabic character recognition is presented, which is also applicable to Kurdish and Persian languages [1].

Arabic is the first language in all Arabic countries. In total, the estimated population of these countries is 280 million and some other countries that consider Arabic as a second language have an estimated population of 250 million. Moreover, Arabic language is ranked fifth out of the most commonly used languages in the world.

Due to the wide usage of Arabic language, it is highly desirable to develop an effective and automatic Arabic character recognition system.

Therefore, in this paper, a technique to recognize Arabic Numerals by using handcrafted features and expert system for decision making is proposed. This method involves extracting geometric features for each object to be further classified using expert system, which is discussed in-depth in the subsequent sections. It is worth noting that in this paper, Indian Numerals are basically denoted as Arabic, Kurdish and Persian numbers. The basic Indian numbers are 10, ranging from 0-to-9as 0,1,2,3,4,5,6,7,8,9. The organization of the rest of the paper is as follows. Section 2 describes the review of previous research works in this domain. Section 3 outlines the methods used for developing the proposed numeral character recognition system. Section 4 describes the experiment implementation. Section 5 provides the analysis of results and discussions on the main findings of the paper. Finally, this research is summarized in Section 6.

II. LITERATURE REVIEW

Many research studies have been conducted on automatic Arabic numeral recognition. Some studies are focused on hand-written recognition, while others concentrate on printed materials. A recognition system for offline handwritten Arabic numerals that exploited the properties of Hindi (Arabic) numerals as powerful set of features is proposed in [3], This method is mainly based on image processing operations and a decision making stage that uses if-else statements to determine the appropriate character output [3].

Also, a Latin number recognition system for number plate localization & segmentation is presented, here, the authors adopted skeletonization method for feature extraction and recognition of the characters is based on Support Vector Machine (SVM). It is claimed that the method is invariant to translation and illumination variation [4,5].

Olasimbo Ayodeji Arigbabu contributed to this work while he was a graduate student at Universiti Putra Malaysia.

An efficient shift and scale invariant approach for offline machine-printed decimal digit recognition that computes the correlation factor between the reference and test image to perform recognition, is described in [6]. Also, in [7], a technique utilizes a number of statistical methods to perform machine print recognition. In addition, several approaches that are based on Neural Networks and Support Vector Machine (SVM) have been investigated for recognition of on-line and off-line handwritten Arabic and Hindi numerals [8-15]. Likewise, Hidden Markov Models have also been adopted for recognition of off-line handwritten numerals [16]. Furthermore, in [17] a genetic programming is used to perform the recognition of hand-written digits, however, it has lack in terms of recognition rate. In [18] translational motion estimation has been examined for the recognition of offline machine-print Hindi digits.

III. METHODOLOGY

A. Framework Design

The overall steps involved in the proposed method are illustrated in Figure 1.

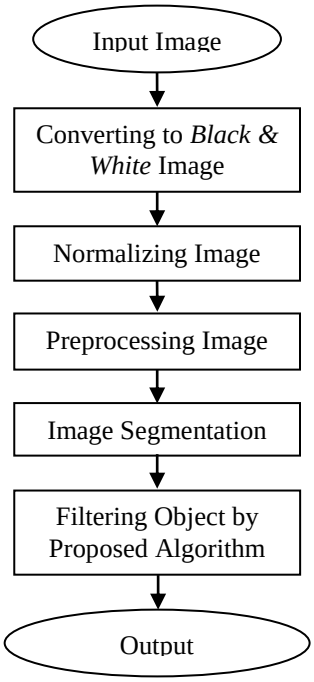
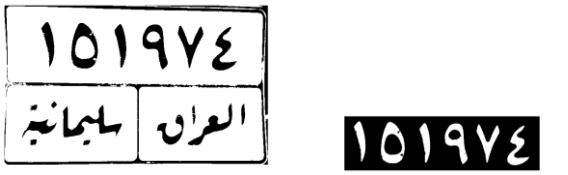


Fig. 1. Flowchart of the proposed offline print-written number recognition system

The process starts with an input image that contains object numbers. Initially, the image is converted to black and white. Then, image normalization (crop) is performed to localize the region of interest (ROI), which is mainly composed of the number section. In addition to that, image complement operation is applied to the normalized image to enable further processing on image pixels with value of 1.



Before Normalization **After Normalization**

Fig. 2. Car number plate image, before normalization and after normalization operation

Further, preprocessing operations including removal of noise and image enhancement are performed. Afterward, image segmentation is applied using region labeling method [19], and finally, the proposed algorithm for number recognition is implemented to obtain the decision of each number's identity.

B. Preprocessing

The normalized input image is converted to binary and the complement of the binary image is derived, as shown in Fig 3a. Then, median filter as a 3 x 3 kernel size for noise removal is used, as depicted in Fig 3b. Closing morphological operation is performed on the filtered image using a 7 x 7 structure element and opening operation is used to remove unwanted small pixels from the binary image. Figure 3 depicts the outputs of the four operations adopted in this research for preprocessing before proceeding to number recognition.

C. Expert System

An expert system is a computer system that emulates the decision-making ability of a human expert [22]. Expert systems are suitable tools for implementing structural pattern recognition techniques and it helps to solve difficult pattern recognition problems. More rules and human experience can be added easily using rule-based systems, especially in closed-system applications with precise inputs and logical outputs [23, 24]. Expert systems have a number of major system components and interface with individuals who interact with the system in various roles as shown in figure 4. In rule-based expert systems, there are two basic techniques; Forward chaining and Backward chaining inference. The domain knowledge is represented by a set of IF-THEN production rules and the data is represented by a set of facts about the current situation. The matching of the rule IF parts to the facts produces inference chains. The inference chain indicates how an expert system applies the rules to reach a conclusion. The inference engine must decide when the rules have to be fired [24]. An inference engine using forward chaining searches the inference rules until it finds one, where the IF clause is known to be true. Forward chaining is used in this paper because of the similarity to the methodology that depends on the data-driven reasoning. The reasoning starts from the known data and proceeds forward with that data. Each time, only the topmost rule is executed, and when fired, the rule adds a new fact to the database. Any rule can be executed only once and the match-fire cycle stops when no further rules can be fired [25, 26].

(a) Black & White with Complement

(b) After Median Filter

(c) After Closing Image Operation

(d) After Area Opening Operation

Fig. 3. Preprocessing steps (Black & White, Median filter, closing filter, area opening operation)

D. Feature Extraction and Recognition

This section discusses the features that are useful for recognition of each numerical object, as well as how the features are obtained or extracted. Prior to that, it is essential to mention that the recognition operation is performed by processing each object once at a time. Therefore, object segmentation operation is considered very crucial in the proposed method. Details of the segmentation operation are elaborated in [27]. For instance, figure 5 shows some examples of segmented Arabic numbers for feature extraction, where each number will be processed separately.

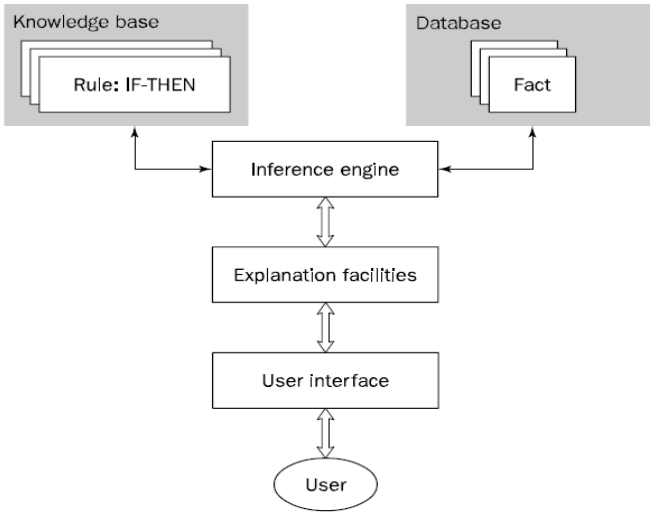


Fig. 4. Components of an expert system [24,25]

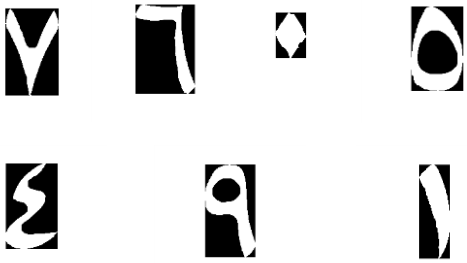


Fig. 5. Some examples of the segmented Arabic object numbers.

The feature extraction and recognition of each number are performed sequentially. In other words, the features of a particular number are extracted using the proposed algorithm, and then, the decision of the number's identity is performed based on the set of rules in the expert system. The process examines every possible match of the facts provided in the inference engine to determine the expected identity of the number. For instance, a random number can be predicted by first checking whether the feature properties align with the facts about number 5 in the inference engine. If an alignment is not found, the system examines the possibility with the facts about number 9. The process is repeated till it reaches number 6, but if a match is found then the expected identity will be presented to the user.

Number 5 Recognition:

In order to assign an identity '5' to an object, two conditions (facts) should be satisfied. Firstly, the Euler number should be zero. Euler number describes the relation between the number of contiguous parts and the number of holes on a shape. Let S denote the number of contiguous parts and N be the number of holes in a shape. Thus, the Euler number is determined as in (1):

$$Euler\ No. = S - N \quad (1)$$

For example, the Euler number for Shape (B) is -1, Shape (9) is 0, and shape (3) is 1. Secondly, the number of flips should be greater than or equal to 3 flips. Flips number is

computed by scanning the object from left to right at the mid-level of the image, as shown in figure 6. The pseudo code for number five is as follows:

If [(Euler Number No. =0) &&
(Flip_number_mid_horizonatally) >= 3]
Then
The Object is '5';

In figure 6, it can be seen that the arrow indicates three flips by scanning from left to right at the mid-level of the image. Number of flips is simply a count of the alternating transitions of pixel values from "1" to "0" or vice versa. Below is a pseudo code for extracting the number of flips.

```
first_value=(1,mid_y);
for x=min: x_max
   if (  first_value ~= Object_array(x,mid_y))
      first_value= Object_array(x,mid_y);
      flip_num=flip_num+1;
   end
end
```

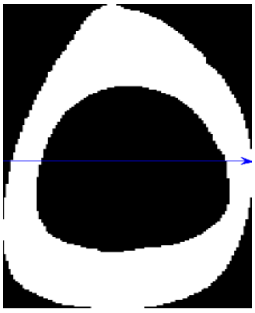


Fig. 6. Shows scanning Arabic object number five to count number of flips

Number 9 Recognition:

In case the conditions for number '5' are not satisfied, the object will be examined with the conditions for number '9'. The first condition is that the Euler number should be zero. The second is that the aspect ratio (calculated by dividing the minor axes by the major axes as shown in figure 7) should be more than 0.6. The Aspect ratio formula is specified in (2):

$$Aspect\ Ratio = \frac{Minor\ Axes}{Major\ Axes} \quad (2)$$

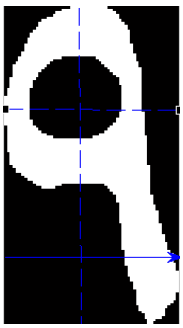


Fig. 7. Shows scanning Arabic object number nine to count number of flips

The third condition is the flips count resulting from object scanning from left to right direction at the lower part of the object, should be less than or equal to 2. As illustrated in figure 7, the arrow indicates the scanning direction to count the number of flips, which in this case is equals to 2. Pseudo code for number 9 recognition is:

If [(Euler Number =0) && (Aspect ratio > 0.6) && (flips-lower-horizontal <=2)]
Then
The Object is '9';

Number '8' Recognition:

Three conditions are used to examine whether the object is number '8', when the conditions for number '5' and '9' are not satisfied. Firstly, the Euler Number should be equal to one. Secondly, the widths of the object at the upper, middle, and lower segments are checked. Basically, the width at the upper segment should be less than the width at the middle and lower segments. The final condition is that the middle width should be less than the width at lower segment of the object. The pseudo code is as follows:

If [(Euler Number No.=0) && (lower_dist > middle_dist > upper_dist)]
Then
The Object is '8';

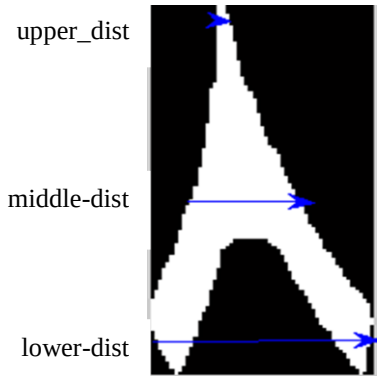


Fig. 8. Shows three scanning positions of Arabic object number eight

Number '7' Recognition:

Three conditions are also considered to determine whether the preprocessed object is number 7, when the conditions of '5', '9', and '8' are not satisfied. The Euler Number of the object should be to one. The width of the object at the upper segment should be greater than the width of the middle and lower segments. The final condition is that, the width at the middle segment should be larger than the lower segment width. The pseudo code is as follows:

If [(Euler Number =0) && (lower_dist < middle_dist < upper_dist)]
Then
The Object is '7';

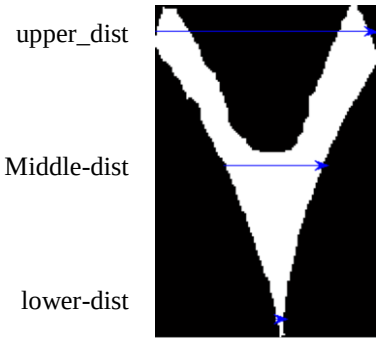


Fig. 9. Shows three scanning positions of Arabic object number seven

Number 3 Recognition:

When the conditions for number '5' '9', '8' and '7' are not satisfied, the object will be examined with the following conditions to determine whether the number is '3'. Firstly, the Euler Number should be equal to 1. Then, the algorithm calculates the number of flips for two separated parts. Firstly, the part located in the top quarter of the number object, as highlighted with the upper arrow in figure 10. Here, the number of flips must be greater than or equal to 6. This kind of feature is quite discriminating in comparison to the features of other number since 3 is the only object whose number of flips is equal to 6 in the mentioned position. The third condition is achieved by calculating the flips in the lower quarter of the object and the result should be equal to 2. The pseudo code for the number three object is as follows:

If [(EulerNumber= 1) && (flip_top_quarter >=6) && (flip_bottom_quarter=2)]
Then
The Object is '3';

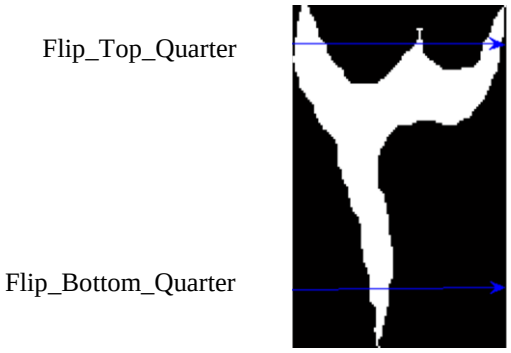


Fig. 10. Shows two scanning positions of Arabic object number three

Number 2 Recognition:

Now, conditions are described to determine the identity of the input image as 2, when the conditions for number '5' '9', '8', '7' and '3' are not satisfied. However, prior to that, it is imperative to mention that morphological processing based on skeletonization [28] as shown in figure 11, is adopted in order to extract features that are peculiar to number 2, and also enhance the processing of further feature computations.

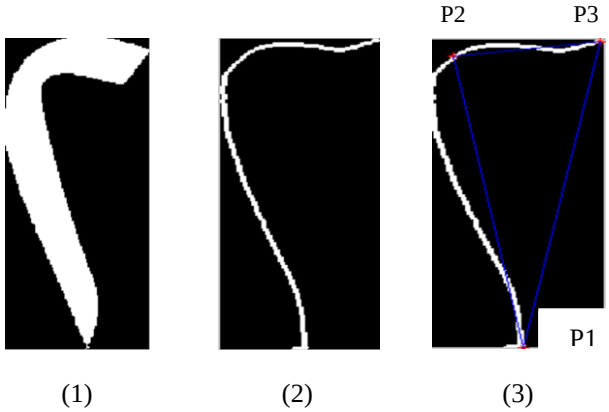


Fig. 11. Shows steps of Arabic object number two

The four conditions are as follows: the Euler Number should be equal to one. The ratio fraction of the object should be greater than or equal to 0.25. This factor is determined by dividing the width (W) distance by height (H) distance. In order to compute each of the mentioned distances, three basic points are located on the object as illustrated in figure 11, which are denoted by the following: P1 is positioned at the bottom of the object, P2 is positioned at the upper left, and P3 lays on the upper right. Since, each point has its coordinates x and y as P(x,y), Height (H), width (W) and slop distance are determined according to the Euclidean distance:

$$W_{dist} = \sqrt[2]{(y_3 - y_2)^2 + (x_3 - x_2)^2} \quad (3)$$

$$H_{dist} = \sqrt[2]{(y_2 - y_1)^2 + (x_2 - x_1)^2} \quad (4)$$

$$Slop_{dist} = \sqrt[2]{(y_3 - y_1)^2 + (x_3 - x_1)^2} \quad (5)$$

Then the ratio is computed as:

$$Width\_Height\_Ratio = \frac{W\_dist}{H\_dist} \quad (6)$$

Third condition is that, the angle between the slop (P1-P3) and x-axes must be positive (larger than zero) to indicate that the object is number '2', let P3(x3,y3) and P1(x1,y1) denote two points of the slop line as shown in figure 11, then the angle is computed using (7) and (8):

$$\theta = \frac{abs(y_3 - y_1)}{x_3 - x_1} \quad (7)$$

$$Angle = tan^{-1}(\theta) \quad (8)$$

The fourth condition is achieved by counting the number of flips, which should be equal to 2 in order to differentiate the number from number three. The following is the pseudo code for the four conditions:

If [Euler Number = 1 && Width_height_ratio > 0.25 && slop_ orientation > 0 && flip_num_top_horiz <=2]
Then
The Object is '2';

Number 0 Recognition:

In case conditions for number '5' '9', '8', '7' '3' and '2' are not satisfied, the object will be examined with the conditions for number '0' which are described as follows: Firstly, the Euler Number should be equal to one. Secondly,

solidity (*S*) factor should be greater than 0.9, (*S* > 0.9). Solidity (*S*) is a scalar specifying the proportion of the pixels object in the convex hull that are also in the region as shown in figure 12, it is computed in (9) as follows:

$$S = \frac{Area\_S}{Convex\ Hull\ Area} \quad (9)$$

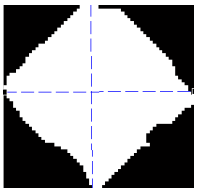


Fig. 12. Shows the Arabic object number zero

Where Area of object (*Area_S*) is calculated according to the conditional equation in (10):

$$Area\_S = \begin{cases} \sum_{y=1}^{m}\sum_{x=1}^{n} pixel(x,y), pixel(x,y) = 1 \\ 0 \quad , \ pixel(x,y) = 0 \end{cases} \quad (10)$$

And the convex hull is calculated as follows:

$$Convex\ Hull\ Area = \sum_{y=1}^{m}\sum_{x=1}^{n} pixel(x,y) \quad (11)$$

The third condition is that, the aspect ratio, which is the division of the minor axis by the major axes, should be more than 0.5. This factor is chosen as both of the mentioned axes sum up to 1, thus 0.5 is considered to take the worst case. The pseudo code of three conditions is:

If [Euler Number = 1 && Solidity > 0.9 && Aspect_Ratio > 0.5]
Then
The Object is '0';

Number 4 Recognition:

In case conditions for number '5' '9', '8', '7', ''3, '2' and '0' are not satisfied, the object will be examined with number '4' conditions which are two conditions. Firstly, Euler number should be equal to one. Secondly, the number of flips should be greater than or equal to 4. To retrieve the number of flips, the object is scanned from the top middle point to the bottom of the image, as shown in figure 13.

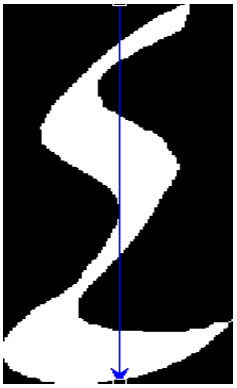


Fig. 13. Shows the Arabic object number four

The pseudo code is as follows:

If [Euler Number =1 && flip_num_middle_vertically >= 4]
Then
The Object is '4';

Number '1' Recognition

By examining the object with the aforementioned conditions, if the output of the object cannot be accurately decided, then the following conditions describing the facts about number '1' will be considered. Similarly, in this case skeletonization is utilized to enable extraction of detailed information about number 1. Afterwards, the three conditions considered are: Firstly, Euler Number must equal to one. Secondly, the ratio fraction of the object shall be less than factor (0.25) (opposite of the number 2 and 6 recognition). This factor is determined by dividing width distance on the height distance. In order to compute each mentioned distance, three basic points have to be located in the object as in the figure 14, as following: P1 is positioned in bottom of the object, P2 is positioned upper left, and P3 lays upper right.

As usual each point has its trajectories x and y as P(x,y). Height, width and slop distances are determined according to Euclidean distance and depicted in the figure 14. The determinations have been explained as in equations in (3), (4), (5) and (6). Additionally, third condition, the angle between the slop (P1-P3) and x-axes must be negative (less than zero) to indicate that this object is number '1', let P3(x3,y3) and P1(x1,y1) are two points of the slop line as shown in figure 14. Then, the angle is computed by taking *Tan* inverse to the theta θ as explained in equation (7) and (8). Now, the pseudo code for the three conditions is as follows:

If [(Euler Number =1 && Width_Height_Ration < 0.25 && Slop_Orientation < 0]
Then
The Object is '1';

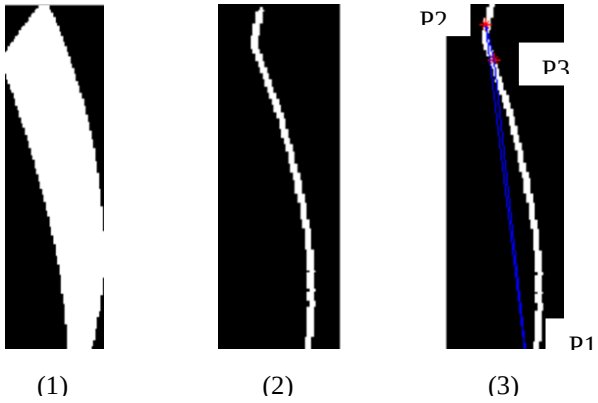


Fig. 14. Shows the Arabic object number one

Number 6 Recognition:

Finally, if conditions for numbers '5' '9', '8', '7', '3', '2', '0', '4', and '1' have not been achieved successfully, the object will be examined with facts about number '6' which is described in the following paragraph.

Also, in this case skeletionalization is initially used to preprocess the object. Afterwards, the three conditions are: Firstly, Euler Number must equal to one. Secondly, the ratio fraction of the object should be greater than or equal to factor (0.25) (opposite of number '1' recognition). This factor is determined by dividing width distance by the height distance.

In order to compute each mentioned distance, three basic points are located on the object as depicted in the figure 15. P1 is positioned in bottom of the object, P2 is positioned upper left, and P3 lays upper right. Also, since each point has its trajectories *x* and *y* as P(x,y) thus the Height distance (H), width distance (W), slop distance, width height ratio are determined based on Euclidean distance using the following equations (12), (13), (14) and(15) :

$$H_{dist} = \sqrt[2]{(y_3 - y_1)^2 + (x_3 - x_1)^2} \quad (12)$$

$$W_{dist} = \sqrt[2]{(y_2 - y_3)^2 + (x_2 - x_3)^2} \quad (13)$$

$$Slop_{dist} = \sqrt[2]{(y_2 - y_1)^2 + (x_2 - x_1)^2} \quad (14)$$

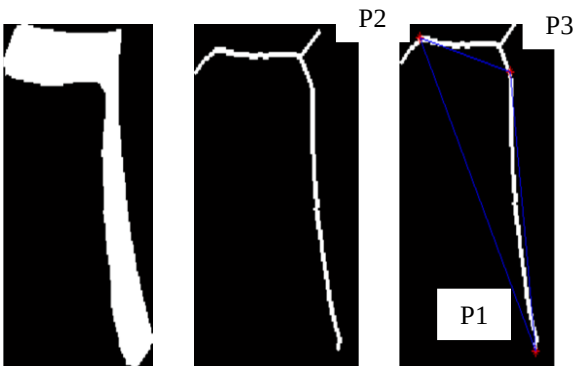
$$Width\_Height\_Ratio = \frac{W\ dist}{H\_dist} \quad (15)$$

The third condition is that, the angle between the slop (P1-P2) and x-axes should be negative (less than zero) to indicate that this object is number '6', which is computed as follows:

$$\theta = \frac{abs(y_2 - y_1)}{x_2 - x_1} \quad (16), \qquad Angle = tan^{-1}(\theta) \quad (17)$$

The pseudo code for the three conditions is as follows:

If [Euler Number =1 && Width Height Ration > 0.25 && slop orientation < 0]
Then
The Object is '1';



(1) (2) (3)

Fig. 15. Shows the Arabic object number six

IV. Experiment and Implementation

To test the proposed algorithm and ascertain its ability to generalize to any random input number, thus, evaluating the algorithm with car plate numbers is considered in this experiment. The number of Arabic numerals is 226 characters collected randomly from 40 images of Iraqi car plates. It is important to mention that, the images are captured in real world conditions where several imaging factors such as illumination, shadow, camera view and incidental lighting are not constrained. Normally, in the verification or identification comparison, there are two possible error measures: False Accept Rate (FAR), which results from the forged template that accepted by the computer system falsely during testing. and False Rejection Rate (FRR), which results from the genuine template that the system recognizes as the query template wrongly [29, 30]. Finally, the total accuracy of the system is calculated by subtracting the average error rate from 100% as in (18):

$$Accuracy\% = 100\% - \frac{FAR+FRR}{2} \quad (18)$$

In this research, FAR error does not exist, as there are no forge templates in this experiment. Therefore, FAR is mainly equal to zero. However, FRR is largely used for the testing measure to estimate the recognition rate, because the Arabic numbers are considered as genuine templates, if they are wrongly recognized by computer system, then the FRR increases. For example, number '5' is deemed as genuine template, if the computer system recognized it as 5, FRR is going to be zero, otherwise FRR will be increased. Finally, the equations that are used to estimate the accuracy are as (19) and (20):

$$Accuracy\% = 100\% - FRR\% \quad (19)$$

$$FRR\% = \frac{Total\ False\ Reject}{Total\ True\ attempt} \times 100\% \quad (20)$$

The proposed algorithm is summarized in figure 16 as follows:

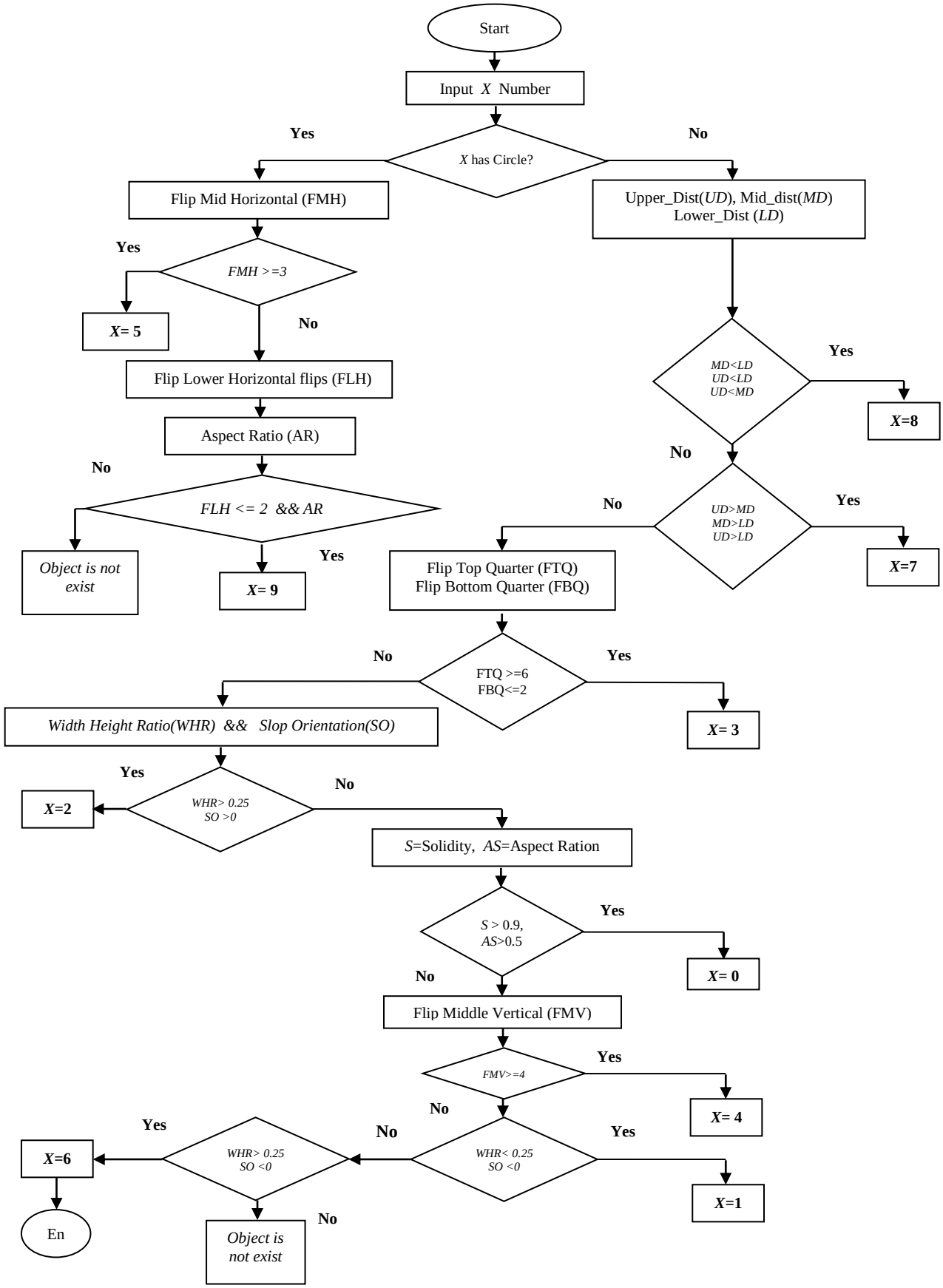


Fig. 16. The proposed algorithm for the number recognition

V. Result and Discussion

Two different results are reported as the outcomes of this research. The first result is the recognition error for each distinct Arabic number among the 226 sample set, which is attained by calculating the total output of a specific numeral type divided by the total input (queried) of the same numeral type that have been randomly collected in the dataset. For example, as shown in Table 1, numeral number 8 has been iterated 25 times with no error in the recognition output (FRR=0). For the remaining numeral types, Table 1 shows the details as dataset iteration times with their system output and also shows the successful accuracy.

It can be seen in table 1 that the highest False Reject Rate error (FRR) is 11.53%, which is specifically related to the Arabic number 6.

TABLE I. Shows the Character Number Types, Computer System Outputs, FRR and Successful Accuracy in Percentage

Numerical Type	Dataset Numerical Attempt	System Recognized Falsely:	FRR as in (20) %	Accuracy as in (19) %
0	19	0	0	100
1	46	1	2.17	97.83
2	15	1	6.66	93.34
3	24	1	4.16	95.84
4	20	1	5	95
5	18	0	0	100
6	26	3	11.53	88.47
7	18	0	0	100
8	25	0	0	100
9	15	0	0	100

In decreasing order, it can be seen that number 2 has 6.66 % error, number 3 has 4.16 and number 1 has FRR of 2.17%. These results are mainly due to the noise in the car plate images, while numbers zero, five, seven, eight, and nine have no error at all during testing. The second result that is reported in this research is obtained by calculating the recognition rate and FRR error for each car plate number, whether each one might consist of 4 or 5 or 6 numeral numbers. Figure 17 shows a bar chart which describes each attempt among the 40-image in *x* axes with their successful accuracy as in 100% in *y* axis of the chart.

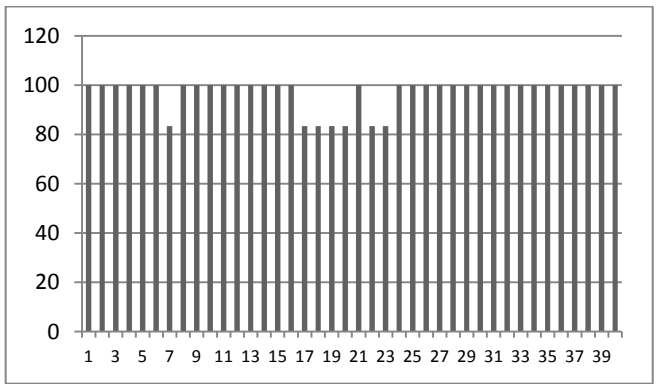


Fig. 17. Illustrates the 40 car plate numbers in X axis with their corresponding accuracies in Y axis

Here, the overall successful accuracy is 97%, which is the average recognition rate for 40 car plate numbers. It is clear in figure 17 that the following images: 7, 17, 18, 19, 20, 22 and 23 have accuracy 83% because one of the numbers is not recognized correctly due to presence of noise in the input image. This rate (83%) is calculated as follows:

$$100\% - \frac{1}{6} \times 100\% = 83\%$$

In this research method, there is no dataset training to be matched against it as matching operation. However, it works by extracting facts by using geometric feature extraction in order to be applied to the set of rules by using if-else statements as an expert system works.

VI. Summary

An effective algorithm for Arabic offline print written number recognition is proposed in this research. Several preprocessing operations are initially applied to the input image such as conversion to binary image, noise removal, morphological filtering, and segmentation before entering the data to the recognition system. The proposed approach is based on extraction of both local features such as computing number of flips of the only upper part of the object, as well as global geometric features such as computing the overall aspect ratio of the object, width, height and orientation of the object number. The features are further quantified into a set of facts or conditions that are used for classification based on a set of rules as an expert system. The experiment has been conducted on a random 226 numbers collected from 40 Iraqi car plate numbers. The output showed that the recognition error rate in terms of False Rejection Rate (FRR) is 3% or the overall successful accuracy is 97%. However, this algorithm is not robust against object translation and rotation.

Finally, improving and investigating the possibility of using the proposed algorithm for handwritten Arabic number recognition rather than print written is considered as a future work.

References

[1] A. Sabri Mahmoud and M. Sameh Awaida "Recognition Of Off-Line Handwritten Arabic (Indian) Numerals Using Multi-Scale Features And Support Vector Machines Vs. Hidden Markov Models," The Arabian Journal for Science and Engineering, Vol. 34, No. 2B, 2009.

[2] M. A. Abuzaraida, A. M. Zeki, "Feature Extraction Techniques of Online Handwriting Arabic Text Recognition," IEEE conference, 5th International Conference on Information and Communication Technology for the Muslim World, 2013.

[3] R. I. Zaghloul, Bader, D. M.K. Enas and F. AlRawashdeh "Recognition Of Hindi (Arabic) Handwritten Numerals. American Journal of Engineering and Applied Sciences," Vol.5, No. 2,pp. 132-135, 2012.

[4] S. Sonavane, A. Khade, V. B. Gaikwad, " Novel Approach for Localization of Indian Car Number Plate Recognition System using Support Vector Machine," International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 3, No.8, 2013.

[5] S. Sivanandan, A, Dhanait, Y. Dhepale and Y. Saiyyad " Automatic Vehicle Identification Using License Plate Recognition for Indian Vehicles," Emerging Trends in Computer Science and Information Technology -(ETCSIT) Proceedings published in International Journal of Computer Applications® (IJCA), 2012.

[6] A. T. Alqudah, H. R. Al-Zoubi and M. A. Al-Khassaweneh, "Shift and Scale Invariant Recognition of Printed Numerals," Abhath Al-Yarmouk: Basic Sci. & Eng. , Vol. 21, No.1, pp. 41-49, 2012.

[7] T. E. Milson and K. R. Rao, "A Statistical Model for Machine Print Recognition," IEEE Transactions on Systems, Man, and Cybernetics, SMC, Vol.6, No.10 (1976), 671-678.

[8] F. N. Said, R. A. Yacoub and C. Y. Suen, "Recognition of English and Arabic Numerals Using a Dynamic Number of Hidden Neurons," Proceedings of the Fifth International Conference on Document Analysis and Recognition, (1999) 237-240.

[9] G. Rajput, R. Horakeri, and S. Chandrakant "Printed and Handwritten Kannada Numeral Recognition Using Crack Codes and Fourier Descriptors Plate," International Journal of Computer Applications, Vol.1, No.1, pp.53-58, (2010).

[10] F. Al-Omari and O. Al-Jarrah, "Handwritten Indian Numerals Recognition System Using Probabilistic Neural Networks," Advanced Engineering Informatics, Vol.18, No.1,pp. 9-16, (2004).

[11] E. Kussul and T. Baidyk, "Improved Method of Handwritten digit Recognition Tested on MNIST Database," Image and Vision Computing, Vol. 22, No.12, pp.971-981, (2004).

[12] C. L. Liu, and H. Sako, "Class-Specific Feature Polynomial Classifier for Pattern Classification and its Application to Handwritten Numeral Recognition,". The Journal of the Pattern Recognition Society, Vol.39, No.4, pp.669-681, (2006).

[13] A. Goltsev, and D. Rachkovskij, "Combination of the Assembly Neural Network with a Perceptron for Recognition of Handwritten Digits Arranged in Numeral Strings," The Journal of the Pattern Recognition Society, Vol.38, No.3, pp.315-322, (2005).

[14] M. Kherallah, L. Haddad, A. Alimi and A. Mitiche " On-line Handwritten Digit Recognition Based on Trajectory and Velocity Modeling," Pattern Recognition Letters, Vol. 29, No. 5, pp.580-594, (2008).

[15] F. Lauer, C. Suen, and G. Bloch "A Trainable Feature Extractor for Handwritten Digit Recognition," The Journal of the Pattern Recognition Society, Vol.40, No.6, pp.1816-1824, 2007.

[16] H. Soltanzadeh and M. Rahmati, "Recognition of Persian handwritten digits using image profiles of multiple orientations," Pattern Recognition Letters, Vol.25, No.14, pp.1569-1576, (2004).

[17] A. D. Parkins and A. K. Nandi, "Genetic Programming Techniques for Hand Written Digit Recognition," The Journal of Signal Processing, vol. 84, No.12,pp.2345-2365, (2004).

[18] H. Al-Zoubi, and M. Al-Khassaweneh, "Offline Machine-Print Hindi Digit Recognition Using Translational Motion Estimation," Proceedings of IEEE International Conference on Computational Intelligence for Modeling, Control, and Automation, (CIMCA '08), Vienna, Austria, 10-12 December ,2789-2792, (2008).

[19] M. Haralick, Robert, and L. G. Shapiro "Computer and Robot Vision," Vo. I, Addison-Wesley, pp. 28-48, 1992.

[20] C. R. Gonzales and R. E. Woods: Digital Image Processing. 2nd edition, Englewood Cliffs, NJ: Prentice-Hull, 2002.

[21] R. Van den Boomgard and R. van Balen "Methods for Fast Morphological Image Transforms Using Bitmapped Images," Computer Vision, Graphics, and Image Processing: Graphical Models and Image Processing, Vol. 54, No. 3, pp. 254-258, May 1992.

[22] Jackson, Peter: Introduction To Expert Systems (3rd) Addison Wesley, p. 2,ISBN 978-0-201-87686-4, (1998).

[23] M. Ahmed, R. K. Ward "An expert system for general symbol recognition," Pattern Recognition, Vol.33, No.12, pp.1975-1988, 2000.

[24] M. Negnevitsky: Artificial intelligence: a guide to intelligent systems. Pearson Education, 2005.

[25] M. Sasikumar, S. Ramani, S. M. Raman, K. S. R. Anjaneyulu and R. Chandrasekar, "A Practical Introduction to Rule Based Expert Systems," New Delhi: Narosa Publishing House, 2007.

[26] T. Sharma, N. Tiwari, and D. Kelkar, "Study of Difference Between Forward and Backward Reasoning," Ujjain: IJETAE, 2012.

[27] H. M. Robert, and L. G. Shapiro "Computer and Robot Vision," Vol. I, Addison-Wesley, pp.28-48, 1992.

[28] T. K. Yung and A. Rosenfeld "Topological Algorithms for Digital Image Processing," Elsevier Science ,1996.

[29] F. L. Malallah, S. M. S. Ahmad, W. A. W. Adnan, O. A. Arigbabu,V. Iranmanesh and S. Yussof "Online Handwritten Signature Recognition by Length Normalization using Up-Sampling and Down-Sampling," International Journal of Cyber-Security and Digital Forensics (IJCSDF), Vol. 4, No.1, pp. 302-313, Dec. 2014.

[30] F. L. Malallah, S. M. S. Ahmad, W.A.W. Adnan, S. Yussof "Non-Invertible Online Signature Biometric Template Protection via Shuffling and Trigonometry Transformation," International Journal of Computer Applications, Vol .98, No.4.pp. 0975 – 8887. July, 2014.

Spatiotemporal Context Modelling in Pervasive Context-Aware Computing Environment: A Logic Perspective

Darine Ameyed
MMS Laboratory, Université du Québec, École de technologie supérieure
Montréal, Canada

Moeiz Miraoui
Al-Leith computer college, Umm Al-Qura University
Makkah, Saudi Arabia

Chakib Tadj
MMS Laboratory, Université du Québec, École de technologie supérieure
Montréal, Canada

***Abstract*—Pervasive context-aware computing, is one of the topics that received particular attention from researchers. The context, itself is an important notion explored in many works discussing its: acquisition, definition, modelling, reasoning and more. Given the permanent evolution of context-aware systems, context modeling is still a complex task, due to the lack of an adequate, dynamic, formal and relevant context representation. This paper discusses various context modeling approaches and previous logic-based works. It also proposes a preliminary formal spatiotemporal context modelling based on first order logic, derived from the structure of natural languages.**

Keywords—context modelling; logic; formal; pervasive system; context-aware system

I. Introduction

Context-aware computing has had significant attention in diverse areas such as human-computer interaction and mobile computing. The context awareness is also an important concept explored in pervasive systems and ambient intelligence. There are still questions around context: definition, modeling, and reasoning.

The notion of a context in itself is not new, and has been the subject of research in many areas. In linguistics for instance, the context was described as the surrounding text element of the language and can help to determine its meaning. While in artificial intelligence, the context definition is derived from two main approaches: the first is a so-called propositional logical approach offered by McCarthy [1] [2], and Guha [3] which defines the context as the circumstances that could determine the truth value of a term. This approach was formulated as a set of predicates that can process a context to make reasoning and logic deductions.

The second approach is called local reasoning (Local Models Semantics (LMS) / Multi-Background Systems (MCS)) [4] [5]. In the LMS / MCS approach, the context is determined by a known set of facts which perform a line of reasoning. The approach provides an incomplete environment description. Thus, the context is partial, inexact and approximate.

However, to design a reliable context-aware system, the context must be well represented and modeled in an appropriate form that allows sharing between different devices in a pervasive system. Using a richer model provides a higher level of abstraction to facilitate adaptation.

As indicated by Henricksen [6], there is usually, a significant difference between the input information and the one which is useful for applications. This difference may be overcome by various types of context information processing. Therefore, modelling is a crucial step in the context treatment. Indeed, modelling includes the analysis and design of contextual information comprised within the system, as an abstract representation at the data-structure level and at the semantics level.

Context-aware applications in pervasive computing environment can adjust their processing to the current context and thus can be easier to use and reliable. Developing systems that allow applications to be context-aware has been subject to extensive research.[7] [8] [9] [10]. One of the challenges in this respect, is developing a flexible and expressive context model. On one hand, there is a need for a uniform representation that can span a plethora of possible contexts. On the other hand, the model should be flexible enough to allow performing complex operations on the context.

The prupose of this paper, is to show that using a logic model based on spatiotemporal axes, yields a reliable way of dealing with contexts. In this model, first-order predicates are used to represent contexts, thus improving expressiveness and offering means to represent various kinds of contexts. The model supports operations on context like conjunction, disjunction, negation, and quantification. An important advantage of using a formal model is that one can clearly specify the efficiency and expressiveness of the model.

A lot of work has been done in the formal methods area of first order logic strength, expressiveness, and decidability. However, most researchers worked on current context and adaptation but not on prediction and anticipate adaptation. This paper develops our vision around context after having redefined in previous work: ``A Spatiotemporal Context Definition for Service Adaptation Prediction in a Pervasive Computing Environment`` [11], we propose actually our method to model contextual information. This paper offers a new model based on first-order logic and spatiotemporal axes.

The rest of the paper is structured as follows: Section II provides an overview of context modelling in the literature. Section III presents our proposed context model based on logic. We will introduce our modelling methodology, and we will show how our model help to get a context model with a high-level abstraction. At the end of section III, we will propose a scenario to demonstrate how we can model a context based in our logic model. Section IV concludes the paper with a discussion, our contributions and presents our future work.

II. Related Work

Context modelling is a fundamental step for the development of context-aware systems. The existence of well-designed context models will ease the construction of such systems. Context modelling consists of analysis and design of contextual information. This information is contained in the system as an abstract form at the data-structure level as well as the semantic level. Several modelling approaches have been proposed, studied and analyzed in the literature. This section, start by presenting a survey and comparison (Table.1) about different context modelling approaches in general, followed by another overview focused more on the work using a logic based model. In this part, this model is evaluated, contrasting its strength and weakness, leading to the motivation for a new model.

A. Overviews of context modelling approach

Strang et al.[12] surveyed the most relevant approaches for context modelling and compared them to some requirements of ubiquitous computing such as distributed composition, partial validation, quality information richness, incompleteness and ambiguity, formality level and applicability. They concluded that ontology makes the best context description compared to the surveyed methods because it provides a good information sharing with common semantics. However, this does not mean that the other approaches are unsuitable for the ubiquitous computing environment.

Bettini et al. [13] discussed the requirements that context modelling and reasoning techniques should meet. They have selected a set requirement for context models: heterogeneity and mobility, relationships and dependencies, timeliness, imperfection, reasoning, modelling formalisms usability and efficient context provisioning. They did not mention logic based context model: instead, they introduced hybrid approaches as an attempt to combine different formalisms and techniques to improve the identified requirements. Perera et al. [14] surveyed context awareness from an Internet of Things perspective. They discussed high-level context modelling techniques. Their focus was on the conceptual perspective of each modelling technique not on specific implementation. Their discussion was based on the six most popular context modelling methods: key-value, graphical, markup schemes, object-based, logic-based, and ontology-based modelling. In their conclusion, they mentioned that logic-based modelling provides much more expressiveness compared to the other models. However, lack of standardization reduces their reusability and applicability. Most importantly, they concluded that diversifying their modelling techniques is the best way to provide efficient results, which will lessen each other's weaknesses. Therefore, no modelling technique is ideal to be used in a stand alone manner.

TABLE I. Comparison of Context Modelling Approaches

Approach	Strength	weakness
Key-value	▪ Simple ▪ Flexible ▪ Easy to manage in a small system	▪ Model limited amount of data ▪ Depend in application ▪ No structure ▪ Not adaptive ▪ No standard processing tool available ▪ No validation support ▪ No relationship modelling ▪ Hard to extract information
Markup schemes	▪ Flexible ▪ Structured ▪ Available processing tools ▪ Useful as intermediate data organization format like network data transfer mode.	▪ Depends on application ▪ No standard ▪ Start be complexes in evolving ▪ Hard to extract information
Graphical	▪ Provide relationships modelling ▪ Easy to extract information ▪ Flexible implementation ▪ Useful for data archival and historic context store	▪ Complex to retrieve information ▪ Configuration is obligatory ▪ No standard ▪ Complex implementation ▪ Hard interoperability between different implementation
Object based	▪ Provide relationship modelling ▪ Available processing tools ▪ Easy integration ▪ Support data transformation over network	▪ Complex to retrieve information ▪ No standard
Ontology based	▪ Support semantic reasoning ▪ Provide an easier representation of context ▪ Advanced tools available ▪ Provide sharing model ▪ Supported by standardization	▪ Complex representation ▪ Complex to retrieve information
Logic based	▪ Generate high-level context based on low-level context ▪ Simple to use ▪ Simple to model	▪ Partial validation difficult to maintain ▪ Applicability can be complicated. ▪ No standard

	▪ Supports logical reasoning ▪ Processing tools available ▪ Can generate new knowledge ▪ Model event and action ▪ Define constrains and restrictions ▪ High level of formality	

B. Synthesis

The following comparison table (Table.1) summarises the review of context modeling approches.

Most of the previous work has focused on ontology-based context modelling, and less effort has been spent on logic-based context modelling. The following section, focuses on logic-based context modelling approaches.

C. Related work on context modelling approaches logic-based

A logic model provides a formal representation of contextual information. Using a reasoning process or an inference engine, a logic model can deduce new information based on existing rules in the system.

Among the first works using this approach, those by Carthy and Buvac [2] [15], introduced the context as a formal object. They defined simple axioms for events or phenomena with common sense and treated the context associated with a particular situation. They provided basic relationship ist (c, p), which means that the proposition p is true in the context C, defined by formulas such as:

C0: ist (context – of (Sherlock Holmes stories), Holmes is detective).

This model also uses the notion of inheritance [15].

Another early representative of this type of approach is the theory of situations introduced by Akram et al. [16]. This approach was inspired by the theory proposed by Barwise et al. [17]. They have tried to give theoretical semantics model of natural language in a formal logic system. Akram et al., have subsequently provided an extension to this model. They represented the facts related to a particular context with non-parametric expressions supported by the type of situation that matches the context.

A similar approach proposed by Gray and Salber [18] used the first order logic as a formal representation of context information and their relationships.

Another approach in this same category was used to develop a multimedia system [19]. In this system, location taken as a context aspect is expressed as a fact in a rules-based system. The system itself is implemented in Prolog.

Ranganathan et al. [20] proposed a context model based on first order predicate, in the ConChat project. Their context model describes context information properties and structure and the kinds of operations that can be performed on context, e.g. conjunction, disjunction, negation, and quantification. The predicate name is the type of context being described.

It is also possible to have relational operators inside predicates. The predicate form is not general and the meaning and number of the parameters depend on the context element. The context model didn't constrain the types of the value-spaces of the different arguments in the context predicate. So, predicate arguments can be randomly complex structures. Arguments of a context predicates can be functions that return some values. In the second time, the authors used rules to derive new contexts based on existing contexts.

Roman et al. [6] presented an experimental middleware infrastructure called Gaia (an Active Space System Software Infrastructure) where they used modeling techniques built on first order logic and Boolean algebra. This allowed them to easily write various rules to describe context information. They represented context through a predicate with an arity of 4, whose structure is borrowed from a simple clause in the English language of the form <subject><verb><object>. An atomic context predicate is defined as follows:

Context (<ContextType>, <Subject>, <Relater>, <Object>)

e.g. Context (location, Chris, entering, room 3231).

In some cases, one or more elements of a predicate may be empty. It is still possible to construct more complex contexts by performing first order logic operations such as context predicates using: quantification, implication, conjunction, disjunction, and negation.

Gu et al. [21] proposed a Service-Oriented Context-Aware Middleware (SOCAM) architecture for the building and rapid prototyping of context-aware services. In their model, contexts are represented as first-order predicate calculus. The basic model had the form of Predicate (subject, value). The context predicates structures are described in an ontology. The ontology is written in OWL as a collection of RDF triples, each statement being in the form (subject, predicate, object).

Other works followed the same approach [8],[22]. Nalepa & Bobek [7] proposed a new rule-based context reasoning platform tailored to the needs of intelligent distributed mobile computing devices. They made a comparison of existent context modelling approaches, and they took into consideration the following aspect of context modelling methods: formalization, simplicity, expressiveness, support for inference, uncertainty handling, and existing tools that support design. They also proposed an inference service that uses HeaRT inference engine to provide efficient on-line reasoning for mobile devices.

D. synthesis

Logic based model provides the ability to create complex expressions in first order logic and deduct a high-level context from the basic context (captured) using an approach based on rules. The model defines a base structure to present each object context atomically. Deduction approaches based on a logical modelling offer the most appropriate mechanisms to achieve abstraction information; it will be more specified later in section III.1.

In spite of the formal high level of logic, less effort has been spent on logic based context modelling and most previous

work on this topic has been centered on the ontology model. The previously proposed works on logic based context modelling suffer from two main weakness points:

1) The context predicates are not generic enough, and their components are not fixed and vary according to the predicate usage.
2) Predicates components do not cover all aspects of the context because they are not based on a clear and concise context definition. This limits their usage to some specific applications and negatively affects the expressiveness.

Therefore, the approach proposed in this paper, follows a logic model that solves these weaknesses. Based on the natural language and our context definition [11], focusing on spatiotemporal parameters and the contextual information usage which promotes proactive adaptation: current or anticipatory based on future context prediction. We have demonstrated that space and time are an important context information in many context-aware applications [11]. Most definition mentions a space as vital factor, e.g., the most frequently used by Dey [9].

As described in section II.2, many research works used the logical approach for its high level of formality, its abstraction benefits, effectiveness and its support for logical reasoning, except these works neglected the time aspect.

Given that contextual information is better defined with the spatiotemporal axes [11], the proposed model integrates these parameters. This will allow a better description of the space service context and thus a more expressive context reasoning and a more efficient adaptation.

III. Proposed Modelling Approach

The logic based models are usually used in context-aware systems for their strong formalism, allowing verification and validation of context models and their ability to automate inductive and deductive reasoning on contextual information. The proposed modeling approach relies on first order logic to model the contect context. The first order logic provides an expressive description of contextual information close to the real environment and natural language.

Firstly, a simple context is described using first-order logic-based formalism. Then, a more complicate complicated context is described with Boolean operators and existential and universal quantifiers (Section III.1). Secondly, we try to provide a simple reasoning logic model that provide a high-level representation of context which can be used as a basis for more advanced reasoning on the context, such as the context discovery or prediction. We believe that logic based models are very efficient tools for context reasoning and are adequate for general pervasive context-aware systems. (Section III.2).

A. Context formalism

1) The basic structure - The context predicate

The required context is the one in which, the service is more likely to be offered. If the current context satisfies this requirement, then the service will be offered.

Definition.1: The context is the set of entities with a spatiotemporal variation that affects the quality of the service, in a short or long term (current service vs anticipatory service) [11].

Definition.2: The state of a service space is the combination of the all the states of the entities existing in this space (including active services and contexts linked to those services).

A context can be reduced to an atomic form, derived from the structure of natural languages. For example, in a natural language people describe information with simple-clause sentence containing a subject and a verb:

Simple clause (<subject> <verb>)

Exp : Adam enters

The natural form sentence can also be used as follows: subject-verb-object

Sentence (<subject> <verb> <object>)

Exp : Adam enters in room

This sentence might be an observation in the context, which might influence the behaviour of a system and trigger an adaptation to offer a service:

Context (<user> <action> <localisation>)

However, the contextual information available is less useful unless we have a complement of information about the spatiotemporal qualifications. In a natural language sentence, time is implicit and given off by the tense of the sentence.

In a systematic description, we use parameters. Knowing that spatiotemporal information in the service space might lead to a more efficient adaptation [11], it becomes a requirement to add two parameters to the description: a time parameter and a location parameter.

This may take the following predicate form:

Context (<element> <state> <value> <times> <location>)

Example: describing the following information "Alex enters the room" in a service space, entails an emphasis the time and location parameters in that information

Context (<Adam> <presence> <active> <21 :00> <room 1>)

- Element: indicates the type of object (i.e. temperature, individual, printer, etc.).
- State: indicates the state of an element, an action, a functionality and is linked to the type of element it describes.
- Value: observation qualifying a state a functionality or an action (i.e. on, off).
- Time: observation time, the instant is when the element's state was observed.
- Location: the place where the observation happened.

To describe a complex context expression, requires use of Boolean, quantitative and existential operators, as will be detailed below.

2) Operation on context

Our goal is to have an accurate description of the physical world. With a pervasive system, we would also like to describe the service space. To ensure context-awareness with an accurate description yields an efficient system. As described above, the atomic form can be extended to describe all the elements of a context in a service space. It is possible to scale in complexity adapting the description accordingly by integrating Boolean operators and logical quantifiers to the predicates.

TABLE II. CONJUNCTION, DISJUNCTION AND NEGATION OPERATOR

Conjunction	And	∧
Disjunction	Or	∨
Negation	Not	¬

Conjunctions, disjunction and negation can also be performed for a complex description, as illustrated in table 2.

- Example:

Context (<lamp> <lighting> <on> <22:00> <room1>) ∧ Context (<User1> presence> <active> <22 :00> <room1>)

Describes user 1 as being in room 1 while the light is on, at 22h00.

¬ Context (<User2> <presence> <active> <22:00> <room1>)

Describes user 2 as not being in room 1 at 22h00.

Context (<lamp> <lighting> <on> <22:00> <room1>) ∧ Context (<User1> <presence> <active> <22:00> <room1>) ∨ Context (<User2> <presence> <active> <22:00> <room1>)

Describes light as being on in room 1 at 22h00 and either user 2 or user 1 are registering their presence there.

3) Quantification

An existential or universal quantification model allows us to represent even richer sets, table 3. A context might be quantified with respect to one of its parameters.

The existential quantifier indicates that the context is true, at least for one mentioned variable.

- Example

∃location Y Context (<user1> <presence> <active><22 :00> <Y>)

The user is present at least in one location

The universal quantifier shows that the context is true for all the occurrences of the mentioned variable.

- Example

∀user X Context (<X> <presence> <active><22 :00> <room1>)

To describe any user in the location designated by 'room 1'

TABLE III. QUANTIFICATION OPERATOR

Existential	Exists	∃
	Exists and is unique	∃!
Universal	For all	∀

4) Context Interpretation

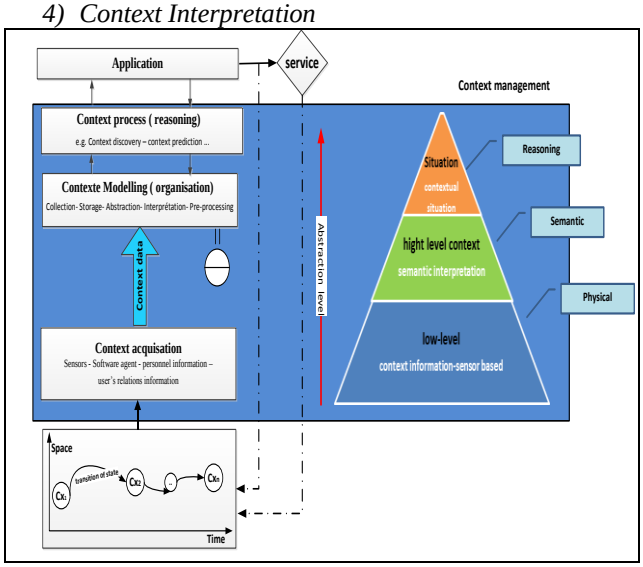


Fig. 1. Context awareness operations with different layers of abstraction

Low-level context consisting in information provided by physical sensor and acquired without further interpretation or analysis, can be meaningless, susceptible, superfluous, or uncertain, the limitation of low-level contextual cues: when modelling different service space interaction risks reducing the usefulness of context-aware applications. A way to mitigate this problem is the elicitation of higher-level context from raw and unrefined sensor values, named context reasoning and interpretation as show in Fig.1.

Using a formal approach for modeling, the context can be processed with logical reasoning methods (e.g.: rule-based, temporal logic). The context reasoning was useful to: (i) checking the consistency of context, and (ii) infering a high-level implicit context from a low-level explicit context.

The idea is to abstract from low-level context by creating a new model layer that generates a higher-level. This refers to a different work in literature discussed a contextual situation [9, 23]

In the context-aware system, situations are the semantic interpretation of low-level context, giving meaning to the application, making it more stable, and easier to define and maintain than basic contextual information [23]. Adaptations in context-aware applications are then caused by the change of situations. Operating at a high level of context abstraction to define a contextual situation make easier application implementation.

In literature, several approaches have been proposed to get abstraction and define situation. For example, [13] enumerates six different ways to describe the situation ``*in_meeting_now*`` based on:

- co-location of people and agenda information
- co-location of filled coffee cups in a room
- devices in the room
- weight sensors on the floor
- sounds and noises
- cameras: watching activity in the meeting room

In this paper, formal logic approach is used to model the context and acquire high-level contextual model concerning the situation.

Early approaches relied on formal logic to describe and represent these states. One of the first approaches: Situation Theory, was proposed by Barwise and Perry [17]. Situation Theory attempts to cover model-theoretic semantics of natural language in a formal logic system [24]. The situation inference affords a logical language for reasoning about action and adaptation.

Our approach based on formal logic provides a high level of abstraction and formality for specifying the context and contextual situation. It also establishes a logic link between context and situation and puts it under the causal connection. This is in agreement with our vision of the context and its use on adaptation or prediction [11].

Based on McCarthy's definition of a situation [1], who described a situation as a complete state of the universe at an instant of time. Therefore, in order to describe a service space situation we do not need to get the whole state of the universe but rather a system environment at this time; which in realty is the context, like a snapshot or instantiation of all context variables at some point of time in a space service as mentioned in definition.2 (section III.1).

The value of context entity parameters changes from situation to situation. To be able to deduce a situation and abstract a context into situation, the characteristic features of a context are used to get properties that are more stable over one situation. Situation encompass a complex context witch can be represented by a predicate and link structure. Situations are a complex context limited by time. The situation can be derived as:

S= (Ti, Te, Cs)

Where: (i) Ti is the starting time it is the first time context parameter associated to the specific situation; (ii) Te is the end time, it is the last time context parameter associated to the same situation; and (iii) Cs is the conjunction of all context entity associated to the situation.

This may take the following predicate form, that can be use it as a deduction rule:

∀time t ∈ [Ti,Te] Context (<element1> <state> <value> <t> <location>) ∧ Context (<element2> <state> <value> <t> <location>) ∧ Context (<element n> <state> <value> <t> <location>) → Cs

In high level:

Cs → S

B. Scenario Morning at work

Adam starts his day; it is a work day. He leaves to work and issues a vocal command to his car indicating his destination: the office. The computed commuting time is 30 minutes. Adam should be at the office at 09h00. On his today's schedule, he has a meeting planned for 10h00, where he is supposed to make a presentation for his team.

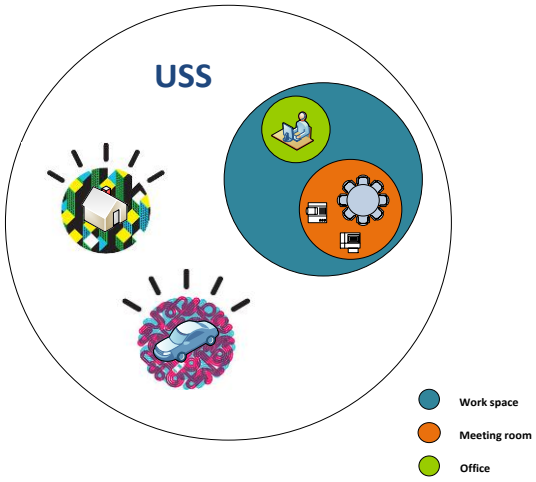


Fig. 2. USS services spaces

Reaching office by 09h00, Adam swipes his access card; the system authenticates him and opens the door. The blinds were already retracted, the temperature adjusted to ambient and the office computer, started. Adam starts working on his presentation. At 09h45, an audiovisual notification appears on the computer screen reminding Adam of his upcoming meeting in 15 minutes. According to the location, he needs 10 minutes to reach the meeting room. The desktop computer saves Adam's work and synchronizes with his laptop.

On the way to the meeting, the system issues commands to upgrade the service spaces:

- USS prepares the meeting room: launch the projector and the display screen, activate the audio system, adjust luminosity, temperature and deploys the blinds.
- USS manages Adam's office while he's away: turns off the desk lamp, locks his computer session, launch system upgrades on the computer and locks access.

Adam and his team are in the meeting room; the presentation is about to start. The system switches all phones to silent mode and locks access to the meeting room.

The meeting ends at 11h00. Adam goes back to his office. The system updates the office's context:

- Unlock the computer session.
- Activate the UV blinds.
- Switch the phone back to normal mode.

USS updates the state of the meeting room, according to its scheduled uses. This paper focus on the context modelling and use this scenario to define the context and the context situation.

In future work, we envision to use the same scenario for the context prediction approach and adaptation.

TABLE IV. DEVICES, STATES AND SERVICE SPACES

Service space	*Devices*	*Devices states*
Office	*Door*	*Locked / Unlocked*
	Blinds	*Open / Close*
	Light	*Switches on / Switches off*
	Pc	*Run / shut off / standby*
	Air conditioner	*Shut of / cool mode / warming mode*
	Smart phone	*Outdoors mode / indoors mode/ meeting mode*
Meeting room	*Door*	*Locked / Unlocked*
	Blinds	*Open / Close*
	Light	*Switches on / Switches off*
	Pc	*Run / shut off / standby*
	Lap-top	*Run / shut off / standby*
	Air conditioner	*Shut of / cool mode / warming mode*
	Screen	*Open / close*
	Projector	*Switches on / Swtch es off*
	Audio-system	*Switches on / Switches off*
	Smart phone	*Outdoors mode / indoors mode / meeting mode*
	Video conferencing system	*Switches on / Switches off*

To describe the situation in one of the service spaces scenario: the meeting room or the office, the model will be based on the ambiance (eg light, sound), the time, location of users (eg present, absent, co-present) and applications (type of application, run, off).

Our scenario's time: a morning in a working day

- Ti : initial time to context situation
- Te: end time to context situation

Based on ambiance – location – time (sample contextual-situation- office), various rules are formalized in first-order predicate based in our context model in order to deduce the space situation.

A few of these rules a few of these rules are as following:

1) Office Context Modeling

TABLE V. OFFICE CONTEXT SITUATION

Situation	*Ambient Cs information*
Cs-office→Work-time	▪ *Lighting (bright)* ▪ *Occupation (busy)* ▪ *Sound (noisy)*
Cs-office→At-rest	▪ *Lighting (gloomy)* ▪ *Occupation (empty)* ▪ *Sound (silent)*

∀time t Contexte (<lighting> <gloomy> <1> < t > <office>) ∧ Contexte (<sound> <silent> <true> <t> <office>) ∧ Contexte (<occupation> <empty > <user=0> <t> <office>) → Contexte (<adam-office> <at rest> <true> <t> <workspace>)

∀time t Contexte (<lighting> <bright> <1> < t > <office>) ∧ Contexte (<sound> <noisy> <1> <t> <office>) ∧ Contexte (<occupation> <busy > <user =1> <t> <office>) → Contexte (<adam-office-situation> <work time > <true> <t> <workspace>)

2) Meeting Room Context Modeling

TABLE VI. MEETING ROOM CONTEXT SITUATION

Situation	*Ambient Cs information*
Cs-room→Meeting	▪ *Lighting (bright-level2)* ▪ *Occupation (busy)* ▪ *Sound (noisy)* ▪ *Phone (meeting-mood)*
Cs-room→Presentation	▪ *Lighting (bright-bright-level1)* ▪ *Occupation (busy)* ▪ *Sound (noisy)* ▪ *Phone (meeting-mood)* ▪ *PowerPoint (run)*
Cs-room→Video-conference	▪ *Lighting (bright-bright-level1)* ▪ *Occupation (busy)* ▪ *Sound (noisy)* ▪ *Phone (meeting-mood)* ▪ *Video-conferencing-system (run)*
Cs-room→At-rest	▪ *Lighting (gloomy)* ▪ *Occupation (empty)* ▪ *Sound (silent)*

∀time t ∈ [Ti,Te]Contexte (<lighting> <bright> <level1> < t > <meeting-room>) ∧ Contexte (<software-app> <powerpoint> <on><t><meeting-room>) ∧ Contexte (<occupation> <busy > <user>1> <t> <meeting-room>) → Contexte (<room-situation> <presentation> <true> <t> <meeting-room1>)

∀time t ∈ [Ti,Te]Contexte (<lighting> <bright> <level2> < t > <meeting-room>) ∧ Contexte (<software-app> <powerpoint> <of><t><meeting-room>) ∧ Contexte (<occupation> <busy > <user>1> <t> <meeting-room>) → Contexte (<room-situation> <meeting> <true> <t> <meeting-room1>)

∀time t ∈ [Ti,Te]Contexte (<lighting> <bright> <level1> < t > <meeting-room>) ∧ Contexte (<video-conf-system> < 1> <on><t><meeting-room>) ∧ Contexte (<occupation> <busy > <user>1> <t> <meeting-room>) → Contexte (<room-situation> <video-conference > <true> <t> <meeting-room1>)

∀time t Contexte (<lighting> <gloomy> <level> < t > <meeting-room>) ∧ Contexte (<video-conf-system> < 1> <on><t><meeting-room>) ∧ Contexte (<occupation> <empty > <user=0> <t> <meeting-room>) → Contexte (<room-situation> <at-rest > <true> <t> <meeting-room1>)

IV. CONCLUSION AND FUTUR WORK

This paper, have presented a formal context model taking into account the spatiotemporal frame. In this model, context informations are presented as first-order predicate calculus. We showed how we can extend the basic model form to an

extended context and how we can use it to deduce various situations to provide high-level context information.

The proposed model follows our pervious reflection on the spatiotemporal contextual information and provides a formal method to introduce it in the context modelling. Compare to the other formal model, this proposal provides notables properties for context model: dynamic context easily understandable; natural language support, logic reasoning support, remaining faithful to a spatiotemporal framework.

Future work will concentrate mainly, on context reasoning and prediction method, and how to design better a context discovery engine, and formalizes it in a generic reusable model.

REFERENCES

[1] F. M. Brown, The Frame Problem in Artificial Intelligence: Proceedings of the 1987 Workshop: Morgan Kaufmann, 2014.

[2] J. McCarthy, "Notes on formalizing context," Stanford University, Stanford, CA, 1993.

[3] R. V. Guha, Contexts: a formalization and some applications vol. 101: Stanford University Stanford, CA, 1991.

[4] F. Giunchiglia, V. Maltese, and B. Dutta, "Domains and context: first steps towards managing diversity in knowledge," Web Semantics: Science, Services and Agents on the World Wide Web, vol. 12, pp. 53-63, 2012.

[5] C. Ghidini and F. Giunchiglia, "Local models semantics, or contextual reasoning= locality+ compatibility," Artificial intelligence, vol. 127, pp. 221-259, 2001.

[6] K. Henricksen, J. Indulska, and A. Rakotonirainy, "Modelling context information in pervasive computing systems," in Pervasive Computing, ed: Springer, 2002, pp. 167-180.

[7] G. J. Nalepa and S. Bobek, "Rule-based solution for context-aware reasoning on mobile devices," Computer Science and Information Systems, vol. 11, pp. 171-193, 2014.

[8] B. Hu, Z. Wang, and Q. Dong, "A modelling and reasoning approach using description logic for context-aware pervasive computing," in Emerging Research in Artificial Intelligence and Computational Intelligence, ed: Springer, 2012, pp. 155-165.

[9] A. K. Dey, "Understanding and using context," Personal and ubiquitous computing, vol. 5, pp. 4-7, 2001.

[10] A. Ranganathan, R. H. Campbell, A. Ravi, and A. Mahajan, "Conchat: A context-aware chat program," Pervasive Computing, IEEE, vol. 1, pp. 51-57, 2002.

[11] D. Ameyed, M. Miraoui, and C. Tadj, "A Spatiotemporal Context Definition for Service Adaptation Prediction in a Pervasive Computing Environment," arXiv preprint arXiv:1505.01071, 2015.

[12] D. Zhang, H. Huang, C.-F. Lai, X. Liang, Q. Zou, and M. Guo, "Survey on context-awareness in ubiquitous media," Multimedia tools and applications, vol. 67, pp. 179-211, 2013.

[13] C. Bettini, O. Brdiczka, K. Henricksen, J. Indulska, D. Nicklas, A. Ranganathan, et al., "A survey of context modelling and reasoning techniques," Pervasive and Mobile Computing, vol. 6, pp. 161-180, 2010.

[14] C. Perera, A. Zaslavsky, P. Christen, and D. Georgakopoulos, "Context aware computing for the internet of things: A survey," Communications Surveys & Tutorials, IEEE, vol. 16, pp. 414-454, 2014.

[15] J. McCarthy, "From here to human-level AI," Artificial Intelligence, vol. 171, pp. 1174-1182, 2007.

[16] V. Akman and M. Surav, "The use of situation theory in context modeling," Computational intelligence, vol. 13, pp. 427-438, 1997.

[17] J. E. Fenstad, P.-K. Halvorsen, T. Langholm, and J. van Benthem, Situations, language and logic vol. 34: Springer Science & Business Media, 2012.

[18] P. Gray and D. Salber, "Modelling and using sensed context information in the design of interactive applications," in Engineering for Human-Computer Interaction, ed: Springer, 2001, pp. 317-335.

[19] J. Bacon, J. Bates, and D. Halls, "Location-oriented multimedia," Personal Communications, IEEE, vol. 4, pp. 48-57, 1997.

[20] A. Ranganathan and R. H. Campbell, "An infrastructure for context-awareness based on first order logic," Personal and Ubiquitous Computing, vol. 7, pp. 353-364, 2003.

[21] T. Gu, H. K. Pung, and D. Q. Zhang, "A service‐oriented middleware for building context‐aware services," Journal of Network and computer applications, vol. 28, pp. 1-18, 2005.

[22] R. Kadouche, M. Mokhtari, S. Giroux, and B. Abdulrazak, "Semantic approach for modelling an assistive environment using description logic," in Proceedings of the 10th International Conference on Information Integration and Web-based Applications & Services, 2008, pp. 222-231.

[23] S. Dobson and J. Ye, "Using fibrations for situation identification," in Pervasive 2006 workshop proceedings, 2006, pp. 645-651.

[24] R. Reiter, Knowledge in action: logical foundations for specifying and implementing dynamical systems: MIT press, 2001.

Application of Data Warehouse in Real Life: State-of-the-art Survey from User Preferences' Perspective

Muhammad Bilal Shahid, Umber Sheikh, Basit Raza, Munam Ali Shah, Ahmad Kamran, Adeel Anjum
Department of Computer Science
COMSATS Institute of Information Technology
Islamabad, Pakistan

Qaisar Javaid
Department of Computer Science and Software Engineering
International Islamic University
Islamabad, Pakistan

***Abstract*—In recent years, due to increase in data complexity and manageability issues, data warehousing has attracted a great deal of interest in real life applications especially in business, finance, healthcare and industries. As the importance of retrieving the information from knowledge-base cannot be denied, data warehousing is all about making the information available for decision making. Data warehouse is accepted as the heart of the latest decision support systems. Due to the eagerness of data warehouse in real life, the need for the design and implementation of data warehouse in different applications is becoming crucial. Information from operational data sources are integrated by data warehousing into a central repository to start the process of analysis and mining of integrated information and primarily used in strategic decision making by means of online analytical processing techniques (OLAP). Despite the applications of data warehousing techniques in number of areas, there is no comprehensive literature review for it. This survey paper is an effort to present the applications of data warehouse in real life. It focuses to help the scholars knowing the analysis of data warehouse applications in number of domains. This survey provides applications, case studies and analysis of data warehouse used in various domains based on user preferences.**

Keywords*—*Data warehouse (DW); Data warehouse applications; Decision support systems; OLAP; Preference based

I. INTRODUCTION

Operational and transactional systems are the new generation systems which are different from 1970's decision support systems (DSS) [1]. In order to complete the life cycle, DSS needs the shadow of a Data Warehouse (DW). A DW pools the available data which is spread all over the organization, and makes a unify pool (like data structure) having the presence of similar and linked formats [2].

Data warehousing takes off in the 1980s as an answer to the very little or no availability of information propagated by online application systems, online applications were praised by a very limited domains of users, and integration was not there even [3]. Historical data kept by online applications are very little as they deposit their historical data for high performance in faster way. Thus organizations hold very little information as compared to data [3].

Inmon drafted that for building a DW most organizations starts with an architecture. "Inmon talks about DW that there is still a way long confusion as what it really is". Bill Inmon [3], [4 p.31], said that the description to a DW was and still is today. "A source of data that is subject-oriented, integrated, nonvolatile, and time-variant for the purpose of management's decision processes".

With the thirst and huge need for large blocks of information, DW gain much importance and became an essential strategy component for medium and large organizations. Timely and accurately decision making at management level becomes difficult due to the incapability of traditional databases to handle increasing demands of online information access, retrieval, maintenance and update efficiently which greatly impacts every industry [5]. So companies start seeking the solution for all their problems and adopt DW technology.

With sharp and harder competition, enterprises are targeting in availing fast and pinpoint information to have best decisions. Furthermore, with the thirst for huge chunks of information, enterprises' traditional DB (database) is off no use of smartly managing the increasing needs of online information update, access, maintenance, and retrieval. This lagging impressively effects the efficiently and effectively usage of internal data by the management to hold decision-making in time. As a result, to search for various ways and means to store, access, handle, and utilize the huge chunks of data in an effective manner, is the main concern of every business [5].

Organizations requires a database system for their daily decision making, with better adaptability, top flexibility, and best support. Considering the past decade, the educational (academia) side and the industry side, both have progressively plated different layouts to solve the problems and to present solution to craft an aforementioned system [5]. Adopting the data warehouse technology is one of the solutions to that. DW was defined by Inmon [3, 4] as, ''pooling data from multiple separate sources to construct a main DW". Proper data-analyzing tools can be used by different users to analyze and store required data.

Data Warehouse's purpose is to take large data from heterogeneous sources and furnish them in known formats that helps in understanding and for making smart decisions [6]. The Benefits linked to the DW applications include the region of time saving, with the availability of clean and handful of information, tough and exact decisions making in accordance with the improvement of processes related to business and to help achieving strategic business objectives [2, 4, 5, 6].

Realizing the need after researching literature and for further exploring on this research article, taking in account the importance of the applications of DW in real life and the shortfall of the factual research, we have all the concreate reason to explore the most applications of DW in real life. In this paper we discussed different applications of DW in real life along with available case studies. Its sections as follows; Section 2 presents DW technology. Section 3 presents the applications of data warehousing in different domains. Section 4 provides a tabular and descriptive view of different case studies under the umbrella of government and business categories. Section 5 provides a brief usage analysis of Data Warehouse applications. Finally, conclusion is presented in Section 6.

II. DATA WAREHOUSE TECHNOLOGY

Devlin and Murphy was the pioneer to present the concept of data warehousing [7]. Read-only database that is capable of storing historical datum for operating was suggested. It offers a variety of integration tools. Users can find and query what they want for supporting decision. Time-variant, non-volatile, integrated and subject oriented are the four key attributes of data warehouse defined by Inmon [8]. With the presence of different attributes, datum is encapsulated in "subject oriented" attribute, which is build and is combined in multiple angles. Talking about an example in a traditional system, a datum for point of sale (POS) might be not same as of other sale systems [4, 8]. The data are hidden separately as a one unit, irrespective of what the under used system is. "Subject oriented" entity tells about the datum that it is build and combined through different angles as said by different authors. Taking in account a traditional system, for example, "custom datum viewed from a POS for sure having different angles from other related sale systems (machines)". Whatever system is used, we have single topic from isolated custom data, by usage of DW [5, 8]. Consistency of data will not be present as it is being integrated, converted and/or extracted by different tools, thus getting an integrated data.

Any variation, in the form of result, can be very important, if the focus of system is on a "real-time" attribute, this includes in the characteristics of time variant. The need for related time and portions of time information is needed by the data stored in data warehouse for future querying. The massive past non-volatile data is held by data warehouse, by which we can perform analysis, prediction and discovery with the positivity of effectiveness, reliability and accuracy. Through modification, we ensure the perseverance of best quality, when data are uploaded in data warehouse. The Inmon's [8] definition of data warehouse has modified and/or redefined by many authors in recent span of time [9, 10, 11, 12, 13, 14]. The scope of data warehouse domain has broadened by different definitions, but is still align with Inmon's definition. According to the different definitions, DW could be summed as, "DW pools daily, both externally and internally "transaction-oriented" enterprise data, and then summed, divide in categories and hold (store) massive data from past (historical) for more computation, forecast, analysis, and discovery of data patterns". Obtained data are linked to non-modified, statistics, and stored in DW for longer period. Furthermore, for analyzing and making decisions they are integrated, time-oriented, and effectively used. We can find at least one chapter related to data warehouse in all major books of databases. As the existence of data warehouse exceeds over 20 years, we can get many useful resources of its design and implementation [15, 16].

A. Data warehouse architecture

Figure 1 shows a general view of data warehouse architecture acceptable across all the applications of data warehouse in real life. Every application of data warehousing include extraction of the informatics data from the key system with using as minor resources as it can, transformation of that data by applying a set of rules from source to the target and fetching (loading) the related data into a DW (called ETL process). Some of the areas DW architecture holds it importance are technical related design, data related design, and hardware and software related design [5, 6, 12].

Design domain of DW architecture widely grouped into enterprise DW design and data mart related design. The enterprise DW is the blend of those adoptive data marts [17]. A data mart is considered to be a tinier version linked to a DW but it aimed on specific subjects. Top-down along with bottom-up techniques linked with data design are followed by data marts [17, 18]. The general DW architectures include the presence of enterprise DW, along with "data marts", linked to the "distributed warehouses", and "operational related" data rooms with data marts, or any mixture to those [4, 17, 18, 19, 20].

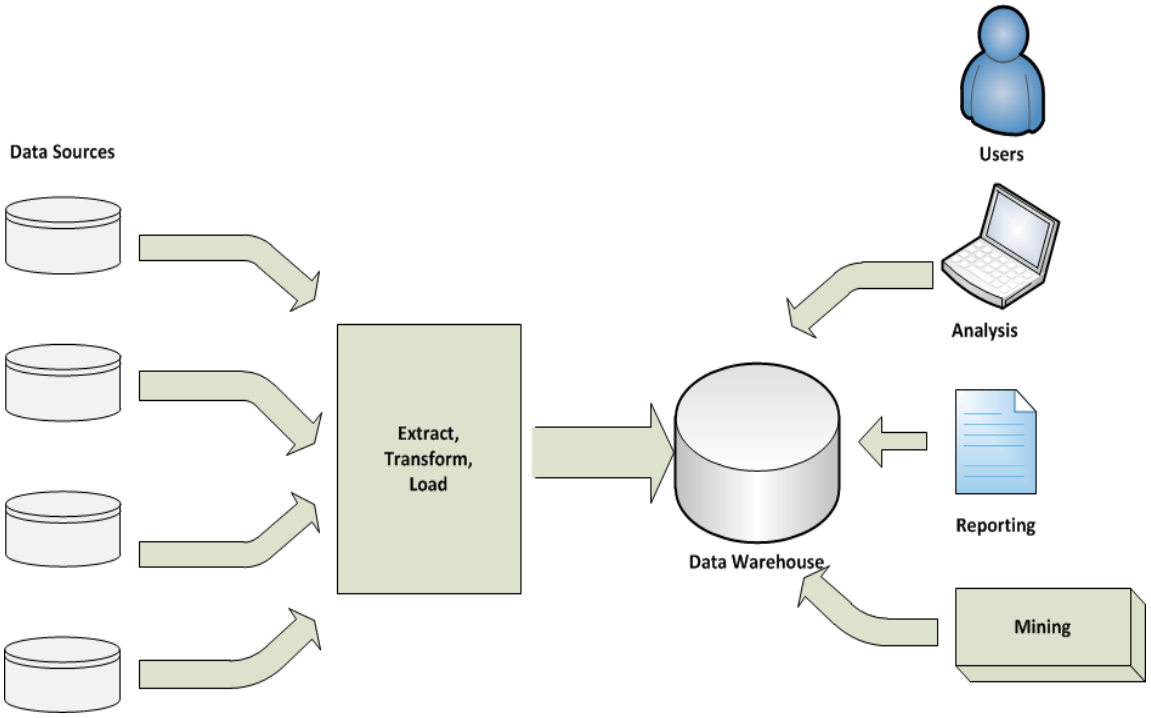


Fig. 1. Data Warehouse Architecture

Figure 1 deeply shows a standard DW architecture. There are many sayings on which architecture best suits the design and implementation. Authors [3, 4, 8, 11, 17] consider Inmon and Kimball as the top of every other, taking in account Sen and Sinha pushed 15 separate methodologies to DW architecture [20]. Figure 1 shows a color print of a general DW architecture. Data are propagated from "operational DBMS" and it is processed by the process called, "extraction, transformation and loading (ETL)" into the DW or data marts. The process or body of the ETL gives a unique data room for decision-making so we always have one unit for it. ETL is said to be the most difficult process of DW construction. Up-to-date and many powerful tools are available to assist this area, but along with artificial tools real human administration is important and for that we require front panels to assist human administrators. Once all the aforementioned processes are completed and the data gathers in DWs or data marts, then we came up with the tools called "online analytical processing (OLAP)". OLAP provides the data into graphical, and in multidimensional prints to help users to query, dig or mine and analyze the data [6, 20, 21].

State of the art research papers have also been published stating the overview, frameworks and up to date practices [22, 23]. Failures parts are also handled by many researchers [24]. The most important thing in making a DW is selecting the best architecture. Extraction from relational database, moving to Transformation, and at the end loading (ETL process), include in the data warehousing environment. It also includes Online Analytical Processing (OLAP) plus the client analysis tools [5, 23].

The process of data warehousing starts from propagation of data from main (original) format passed to a "dimensional data" region for storages purpose, it handles a huge amount of work, clock and money. Implementation and designing of a DW demands cost and is quite critical, for handling those critical tasks, tons of tools related to data extraction, data cleaning and load utilities are present to aide in. Data integration is considered to be the top and most useful part of the DW [1, 5, 6].

III. Applications of Data Warehouse in Real Life

Importance of DW cannot be denied due to its benefits because decisions at management level will no longer need to be taken on the limited and inaccurate data and it also helps the companies to avoid different challenges. So it becomes the need of every individual company to implement data warehouse.

It is estimated that by 2020 around 200% more devices will join the Internet and share data. DW strongly depends upon devices and inter linked data. The more interlinked devices are, the more powerful and useful DW. According to the forecast by many organization [25, 26] by 2016 around 6.4 billion connected peers will join the room globally, an increase of 30% from 2015. Cisco and other research agencies [25, 26] think that approximately 20 - 50 billion devices will be connected by 2020, (see Figure 2) [25, 26].

Other side of the picture is that cost will increase too. If we talk about spending on hardware, the applications related to consumer will hit to $546 billion by the end of 2016; apart from that the usage of connected items in the organization will be somewhere around $868 billion by the end of 2016 (refer to Figure 3) [25, 26].

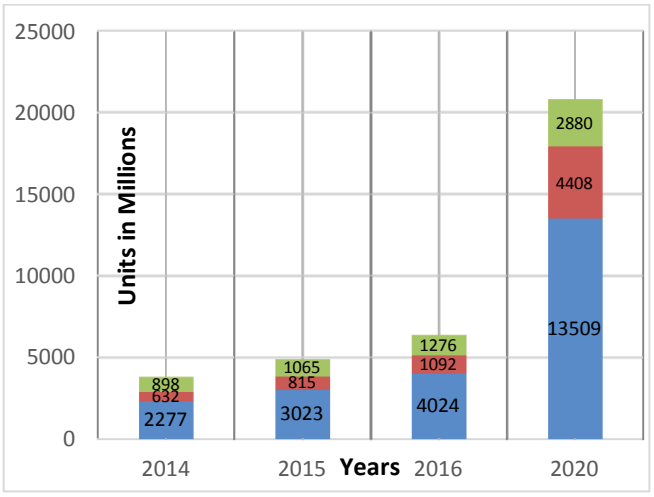


Fig. 2. Number of Units

Talking about relevance of DW, it is said that few of the application areas holds the presence and integration of data throughout the enterprise, furthermore a fast decisions on live and previous (historical) data, give specific information for those systems that are defined loosely. Figure 4 shows the cycle of real life applications of data warehouse in different fields and how they are interrelated according to user preference.

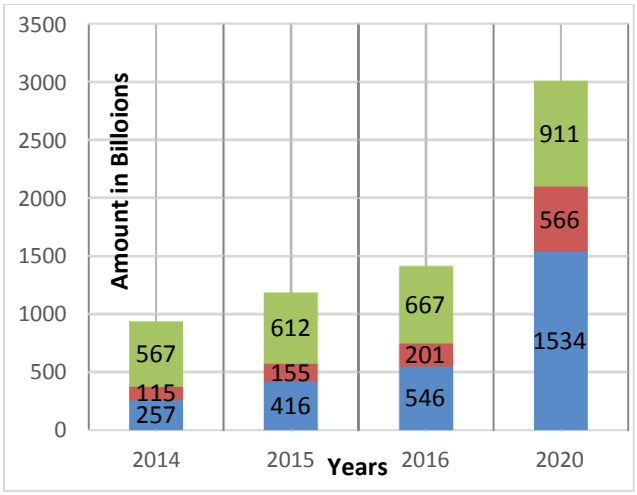


Fig. 3. Cost in Billions

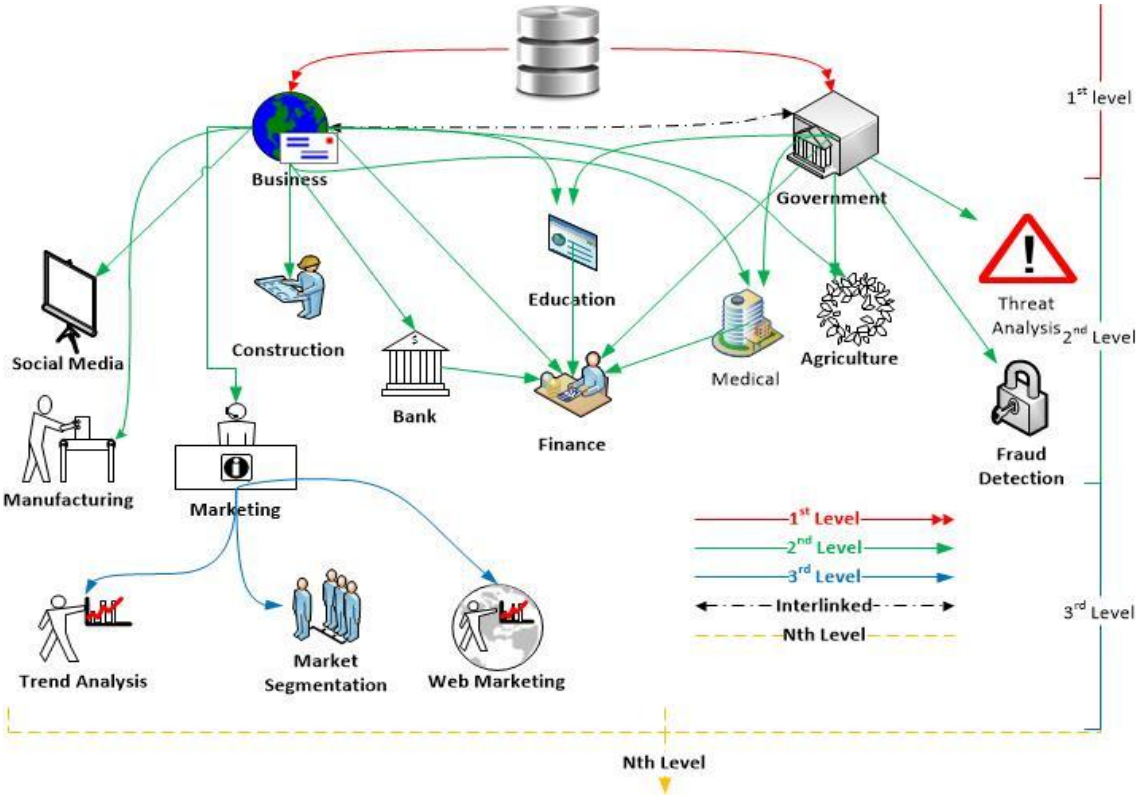


Fig. 4. Applications of Data Warehouse in Real Life

We have suggested a generic layout of interlinked applications of data warehouse (DW). As we can see that different levels are defined. These levels are associated with the hierarchy such that first level is the core component. The first level is always be a central DW (core system(s), hardware system(s)). Furthermore, 2nd level is associated with one of the world's top domains (Root level, business and Government). The reason behind selecting Business and Government as top of hierarchy is a handful of literature, and all other domains are encapsulated under them. With the presence of 2nd level all other sublevel gets populated. The 2nd level serves as the only pillar that supports all other domains. 2nd level is said to be a specific level. 3rd level domains are the more general than specific. The Nth level is the most general level that holds all minor to major domains. Figure 5 shows the flow diagram, which moves from specific to general.

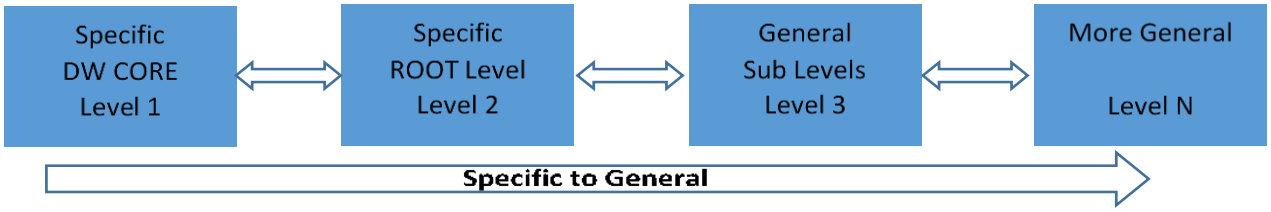


Fig. 5. Specific to General Flow of DW

A. Buisness

Improvement related to decision making and increasing organizational performances are the basic reasons to adopt DW in business [27]. Business holds a key location in applications of data warehouse. All other private and semi-private organizations come under its umbrella.

In DW, for easiness a single repository is used to store data, which is extracted from different databases. This data repository provides forecasting which helps the business personals and business managers. This complete cycle is used to help in identifying the requirements for business and to draft a plan for business [28]. Some of the major to minor fields effecting data warehousing in business are discussed further as shown in Figure 6.

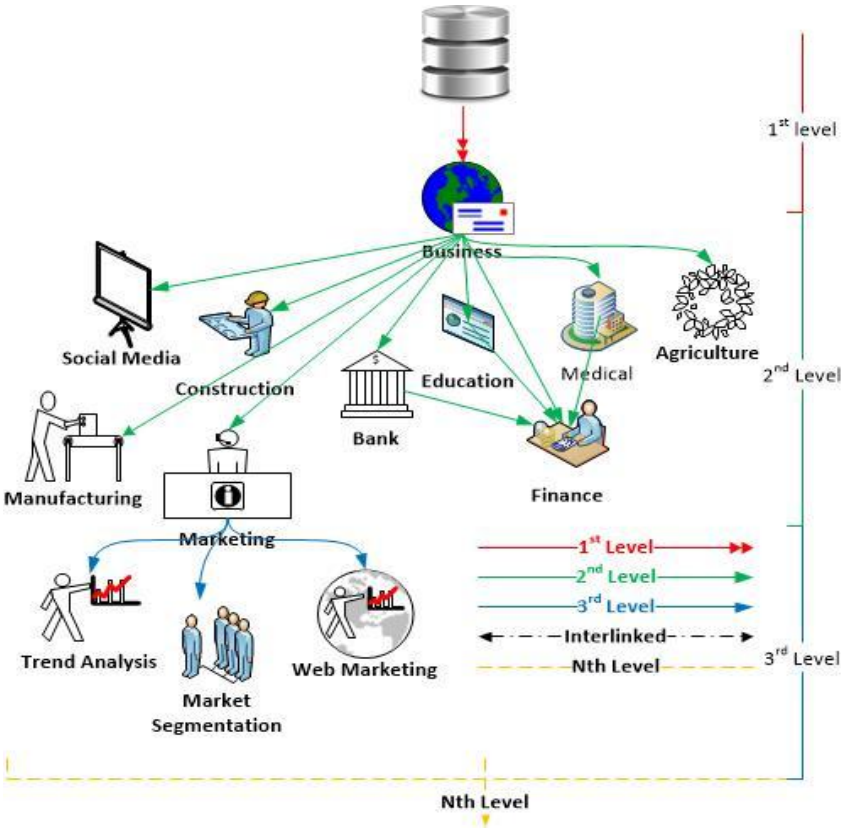


Fig. 6. Business (Application of Data Warehouse)

1) Social media websites

Social media is a great example of data warehousing. Social media industry is emerging and so is the need to implement DW in it. A number of features from Facebook, Twitter and other social media sites are also based on analyzing large data sets [29]. It gathers all data like groups, likes, friends, location mapping etc. and stores it in a single central repository. Although all this information is stored in separated databases but the most relevant and significant information is stored in a central aggregated database [28].

2) Construction (material based industries)

Data warehouse approach in construction industry seems to be efficient in decision making as it provides construction managers the complete internal and external knowledge about available data so that they can measure and monitor the construction performance.

Application of DW in construction industry clearly shows that construction bosses can smartly judge the stock remaining, inventory related trend linked to the materials, the amount and quantity of each material and also the price of all materials [30,

56]. It would also be helpful in reasonable resource allocation to fulfill the required services, maintenance and operation of the systems, allocation of financial budgets, effective managing of investment related long term plans and identification of potential risks [31].

3) Manufacturing Industry

DW plays a vital role in daily house to industrial hold things. Manufacturing industry includes product and process design, scheduling, planning, production, maintenance and huge investments in equipment, manpower and heavy machinery. In this scenario, decisions taken will have wide-ranging effects in terms of profitability and long-term strategic issues. Many industries are trying to convert themselves and many should adopt DW technology rather than traditional decision making so that a warehouse gathers, standardizes and stores data from various applications for improvement in processes and increasing its efficiency as analyzing the data in separate applications is time-consuming. At this stage, some transaction processing systems, which are updated timely, are often hired to propagate the routine business of manufacturing and construction companies [56, 57].

4) Marketing

Every business is not successful without proper marketing and marketing is not successful without knowing the latest trends and demands. Shown in Figure 7 is a general lay out of marketing and its sub domains. Relationship marketing is a new terminology linked with how different businesses handle their customers and the relationships in between that are assets for them and how they can be improved for long-term profitability. DW in marketing is used to examine the patterns of customer's behavior and use this customer information for implementing relationship marketing. They play a vital role in identifying and targeting the profitable customers [32].

Uses of Data Warehousing in marketing area shown in Figure 7 are further categorized as:

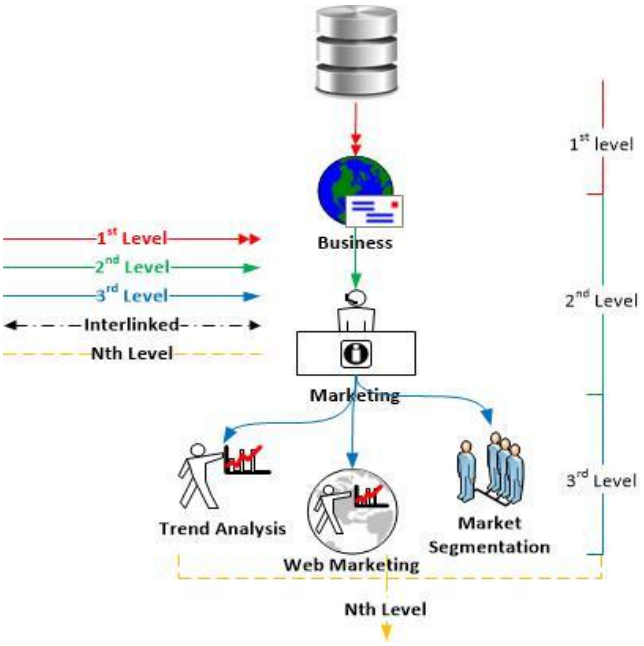


Fig. 7. Marketing (Application of Data Warehouse)

a) Trend Analysis

It is a technique that is used to predict future outcomes from historical results or information. Different medium to large scale enterprises are converting to this. In trend analysis, DW can be used to examine the behaviors of the customer by using historical records over consecutive months.

b) Web Marketing

Web is a hub of billions of devices and around 20 – 50 billion devices till 2020. It refers to a category of advertising that includes any marketing activity conducted online. Facebook, google, and many major to minor such like sites uses web marketing and are relying on latest updated data warehouse.

c) Market Segmentation

Behavior identification is the top most priority of any organization. Market segmentation is the identification of the customer's behavior and common characteristics related to the purchases made against that product of related company. Many organizations are focusing on integrating data warehouse to get best behavior analysis.

5) Banking

The banking industry is categorized as one of the highest information demanding industry in the business world. With the advancement in information technology sector, the role of business intelligence (BI) increases with great number in the process of banking operations [54]. The increased business speed and growing competition has shown the need of banking intelligence dramatically. Bank intelligence is the ability to gather, manage, and analyze a large amount of data on bank customers, products, operations, services, suppliers, partners and all the transactions. As data increases, it becomes difficult to collect, handle and transform it into useful knowledge and DW solves this problem. Many data warehouse flavors are designed for the support of banking industry.

6) Education

DW in education field is becoming popular day by day. Use of DW in educational field presents several potential benefits in making appropriate decisions and for evaluating data in time which is the basic target of DW process. DW provides an integrated and total view of an institute [33]. Most of the related departments use data warehouse as a source of information about faculty and students. DW helps the students in getting their results and notes from a web enabled database quickly through a student portal and last but not the least it helps in decision making by providing current and historical information of the institute.

On a large scale, a DW can integrate the information of different institutes into a single central repository for analysis and strategic decision making.

7) Finance

With the advancement in technology, especially IT industry has opened the doors to the new ways of handling business considering financial systems. Government and Business domain holds equal part in finance. Financial systems may include banks, post offices, insurance companies, income tax and all other tax departments etc. Implementation of data

warehouse in financial industry has several benefits e.g. it can maintain transparency in account opening and transactions. Similarly, government can take decisions against any financial crises. These systems are intelligent enough to spot the defaulters and may act according to the situation. As data warehousing is maintained in this scenario so efficient decision making process can easily be performed. These data warehouses in finance applications can also be used for the analyzation and to have forecasting of different aspects of business, stock and bond performance analysis [34, 58, 60].

B. Government

Amongst the two major sub-divisions of DW industry, government holds equal division. Government can use data warehousing technique in different fields e.g. for searching terrorist profile and threat assessments, in agriculture, in educational industry, in financing department, medical departments and for fraud detection. The telecommunication industry and Banking industry holds many issues related to user frauds. Figure 8 shows application of data warehouse in government departments.

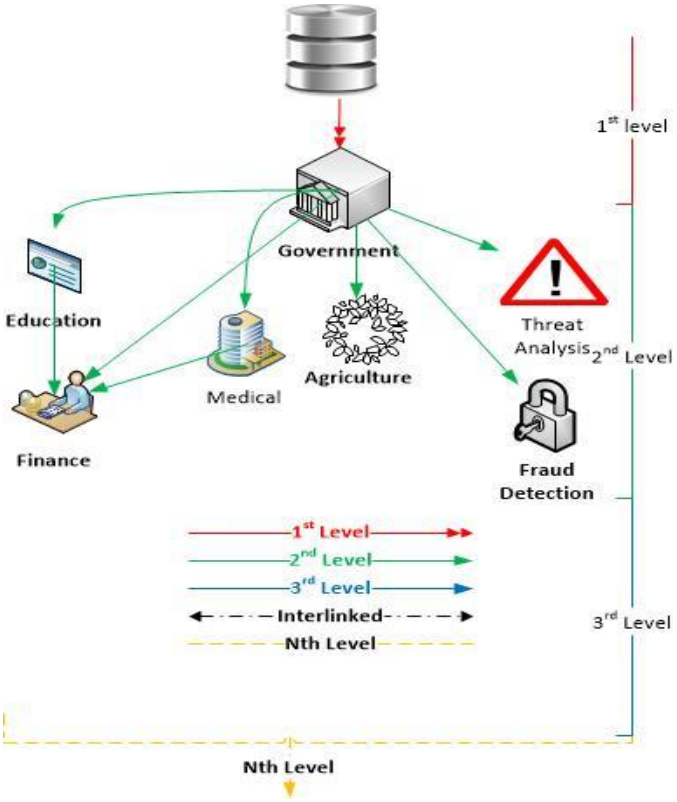


Fig. 8. Government (Application of Data Warehouse)

1) Medical

Medical sector is emerging as the highest DW implementer industry. In health-care, data quality and demand for quality medical services has become increasingly important [55, 59]. Due to the intricacy and variety of medical cum clinical data, the adoption of data warehouses by health care was slow as compared to other fields. Over the past few years it was reported that the usage of DW increased by the administrative and clinical areas. Data warehouses can help in improving the care of specific patients. These health-care institutions are adopting data warehousing for strategic decision making as a decision supporting tool. It provides the tools for acquiring medical data, for extracting the relevant information from that data and finally making this knowledge available to all the concerned persons. Administrative data in data warehouse can help in providing the information about skilled staff needed for a particular treatment and this information further used for the treatment scheduling and to help supporting medical personals in human resources area [36].

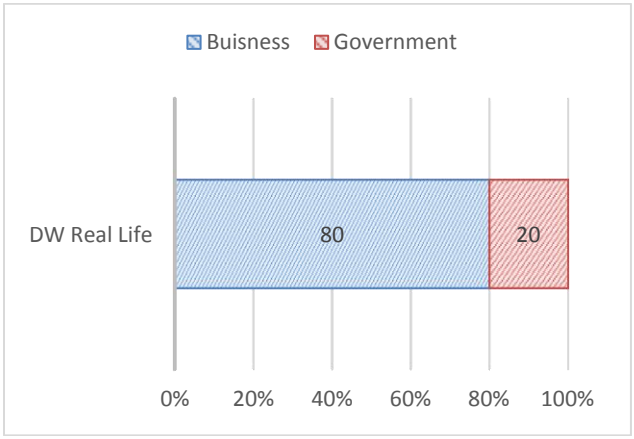


Fig. 9. Percentage wise Contribution of Data warehouse

2) Fraud and Threat detection

Governments are playing their part to detect any threat and fraud caused by ill-minded people. Unfortunately, almost no specific data warehouse implementation that is known is available. Data warehouse access to governments are there, but they need a data warehouse system that is linked with every corner so that threats and terrorists will be monitored.

IV. Case Studies

In this section few case studies are discussed. As discussed earlier data warehouse world is a blend of two parts i.e. business side and government side. Both sides have their own further divisions and any other increment will be added under them. A graphical view is presented in the Figure 9, which is related to the contribution made by business and government domains to DW. It is clearly observed by the survey that 80% of Business and 20% Government related organizations are contributing in the progress of data warehouse.

A. Business

DW in business is now emerging like a hurricane. Around 80% of data warehouse implementation is captured by business. Following are few case studies related to business implementation of data warehouse.

1) Finance

Financial services company (FSC) is considered to be the leading marketer of investment besides banking for products. They implemented DW named as VISION. The user of VISION consists of financial and marketing analysts, managers. It was developed with substantial business and technical goals that can gave a factual and precise picture of best customers of banks and also about most important products [27].

2) Medical

This case study is based on generation of evidence-based guidelines performed by University Health Network (Toronto) which clearly showed that it is authentic, influential and user-friendly to have a DW related to clinic for best strategic decision making. Without this IT support, it would not be imaginable to look for evidence-based medicine as it is difficult for clinicians to gather data for a specific disease [36].

3) Banking

Their research problem is based on the factors that banking industry should consider before and during the adoption of DW technology. Their results revealed the number of banks in Taiwan that adopted this technology and also the architectures that these banks implemented [5].

4) Manufacturing

Large Manufacturing Company (LMC) is making its way to top for production of home related appliances. LMC implemented data warehouse technology as there is a great need to improve the technical infrastructure of the company. Before this, data was scattered in different formats throughout the company and this makes normal and basic functioning difficult for business units. This warehouse provides support to marketing, manufacturing and logistic applications by providing data to dependent data marts [27].

B. Government

Data warehouse in government plays a vital and critical role. Around 20-35% of data warehouse industry is captured by government. Many developing countries are now transferring to the use of data warehouse. Few case studies related to government and usages of data warehouse are as follows.

1) Medical

In Utah and Idaho, Intermountain Healthcare implemented EDW. This healthcare system operates 22 hospitals, 179 clinics, physician offices. This case study is about venous thrombosis patients. Datasets consists of: records of Inpatients, columns of outpatient, financial data linked to or from patient's accounts, data from laboratories related to clinics for the process of imaging and surgery [35] etc. Their DW is updated each night that includes: Large Metadata Repository, Security and auditing infrastructure and Master Reference Data. By using latest information from data warehouse patients with high risk are identified and their reports were sent at every hospital or clinic [35].

2) Finance

Internal Revenue Service is the agency of U.S. that is responsible for tax collection and tax laws imposition. They implemented data warehouse CRIS as there is no way to recoup entity with convinced attribute and perform some analysis on these marked entities. This implemented DW consisted of five domains: business entity, tax returns entity, related to taxpayer transactions entity, peoples' income sources entity and tax payments details entity [27].

C. Tabular view of case studies

Table 1 is the tabular view of all aforementioned case studies.

TABLE I. TABULAR VIEW OF CASE STUDIES

Ref. No	Domain	Architecture	Methodology	Dataset	Method Description	Strength	Limitation
[35]	Healthcare	Enterprise Data Warehouse	Questionnaire	• 22 hospitals, • 179 clinics, • physician offices, home healthcare in Utah and Idaho	• A computer program was for monitoring. • Patient's identification according to score. • Update EDW. • Evaluation.	• Proposed framework can be reused easily for new applications.	• No enterprise database with daily updated patient lists.
[36]	Healthcare	Data Warehouse based approach (for integration of data sources) + Data mining techniques	From published clinical evidence i.e. books, magazines, journals, healthcare, protocols, clinical trials.	University Health Network (Toronto)	• Generation of treatment rules based on clinical evidence. • Data loading to DWH. • Trends identification by data mining techniques. • Rules examination and approval. • Judgments about recommendations and for improving patients care.	• Reliable. • Powerful. • User-friendly platform.	• External evidence-based knowledge is not enough but needs to be adjusted according to patient's health and preferences
[5]	Banking	Data Warehouse	Questionnaire	Banks of Taiwan i.e. from 50 banks and 30 valid responses with response rate 60%.	• Questionnaire with six sections. • Analysis about banks that adopted, in process of adopting or abandoned DWH technology. • Analysis about	• Identification of factors that can affect DWH adoption. • Facilitate implementation in global or overseas branches.	• Limited to domestic banks. • Approach is restricted to banking industry only. • Limited samples.

					architecture of DWH adopted by banks.		
[27]	Finance	VISION data warehouse	Interviewing of employees, examining documents and video tapes of key events.	Financial services company (FSC, US)	• In first phase, top revenue producing customers are identified. • Second phase provided, profitable information for all bank's customers and products.	• Gives more clear and accurate picture of most important customers and products.	• Limited to Critical Financial data • Limited data samples
[27]	Finance	Compliance Research IS (CRIS)	Interviewing of employees, examining documents and video tapes of key events.	Internal revenue service (IRS, US government	• A query processing front-end in CRIS for automatic weighting. • Business rules for facilitating queries.	• Improvement in accessing and generating the reports on taxpayers that was time-consuming without DWH. • Increase in revenue.	Not Defined
[27]	Manufact-uring	Data warehouse	Interviewing of employees, examining documents and video tapes of key events.	Large Manufacturing Company	• Transfer of data from 100 mainframes and 6 external data sources to DWH. • From DWH data is transferred to dependent data marts.	• Helpful in making better decisions and creating better information • Quality information access. • Performance and failure for all parts can be measured. • Reason of failure detection becomes easier.	• Limited data samples

TABLE II. COMPARISON OF DIFFERENT CROSS DOMAIN AREAS AFFECTING DATA WAREHOUSE

Ref. No	Domain	Areas of Usage	% age Used	Cross domain
[37,38,40,39,35,41,42,36,28,12]	Medical	Hospitals, Clinics, Physician offices	23.3%	Government/Business
[34,45,28,27]	Finance	Tax departments	6.2%	Government/Business
[28,5,46]	Banking	Baking industry all around the world	6.2%	Business
[27]	Manufacturing	Home appliances	1.9%	Business
[33]	Education	Schools, colleges, universities	3.8%	Government/Business
[28, 47, 48, 49, 50,32, 52]	Marketing	Customer relationship management, trend analysis and information system	16%	Business
[29]	Social Media	Facebook, Twitter, others.	6.2%	Business

[43,44,31]	Construction	Infrastructure management	8.7%	Business
[6]	Agriculture	Agricultural production department	3.8%	Government/Business
[37,38,40,39,35,41,42,36,28,12]	Fraud Detection	Airports, Crime Agencies	1.9%	Government
[34,45,28,27]	Threat Analysis	Airports, Crime Agencies	1.9%	Government
[28,5,46]	Others	Others	20%	All

V. Analysis and Results

In this section we will see the areas, cross domains and usage of data warehouse around the world and the graphical view of inter related data effecting data warehouse.

A. *Comparison of different cross domain areas affecting data warehouse*

Table 2 shows the comparison of different cross domain areas and their interlinked data.

B. *Graphical representation of Survey*

Following graph shows percentage captured by different areas in DW around the world. As we can see from the Figure 10, medical holds top position in using DW technology.

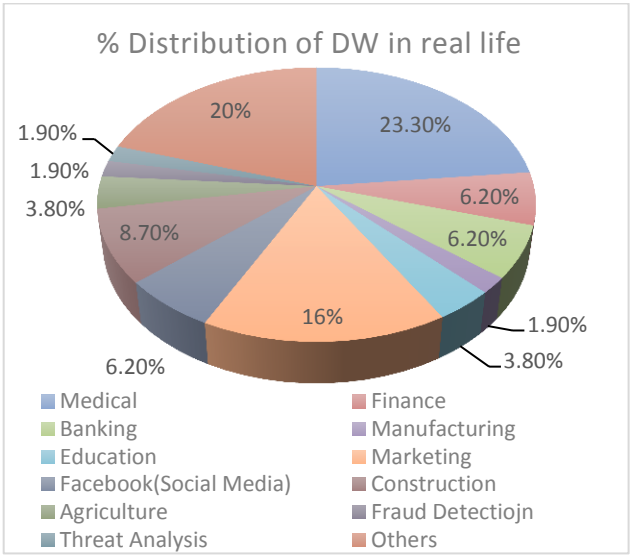


Fig. 10. Percentage Distribution of DW in Real Life

The Figure 11 shows the domain wise importance of DW, we can see clearly business domain holds top position. If we further drill down and look into specific business domain, we see from below the Figure 12 that banking and construction organizations are on top and competing each other with very less margin.

At the end if we take government domain we see that it holds a minor part in data warehouse. Fraud and threat detection are the only region effecting data warehouse through government as shown in the Figure 13.

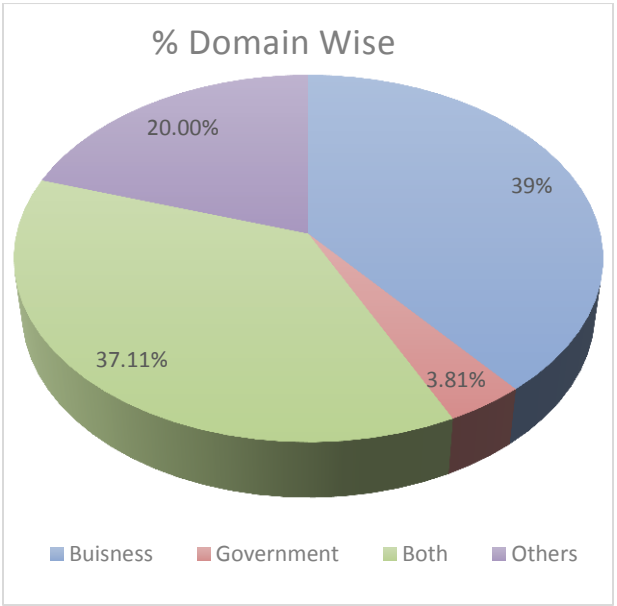


Fig. 11. Percentage Domain Wise

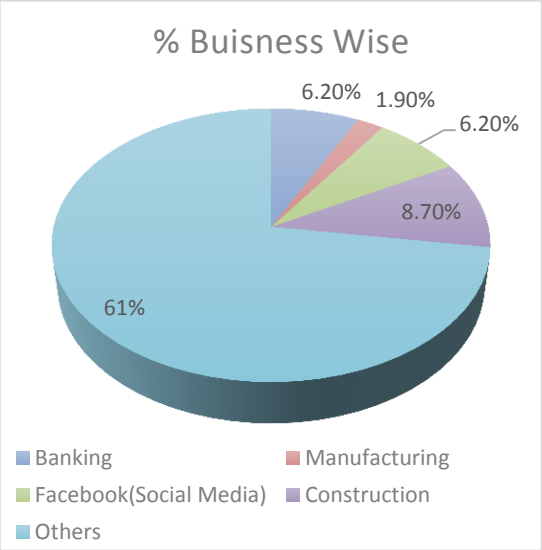


Fig. 12. Percentage business wise

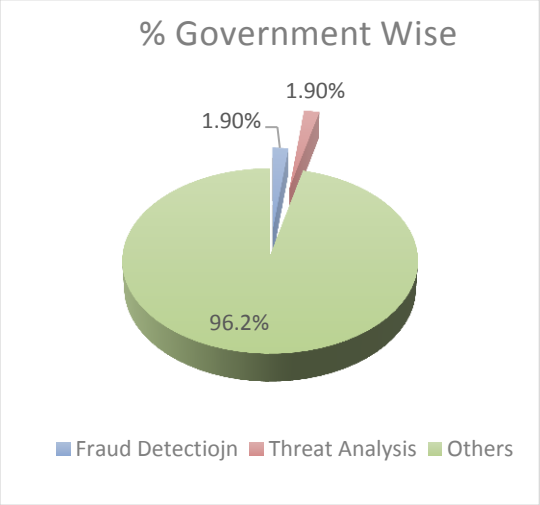


Fig. 13. Percentage government wise

VI. Discussion and Conclusion

This research survey describes the applications of data warehouse in various domains including government and non-government organizations. Our analysis is based on the literature review and case studies provided in this survey. The analysis of this study shows that the non-governmental organizations use data warehouse technology much more than the government organizations. The governments mostly use data warehouse for controlling the crime and fraud. Non-governmental organizations mostly use DW for data analysis, prediction and making decisions. Case studies are shown in the Table 1 that describe the importance of data warehouse in four domains; Healthcare, Banking, Finance and Manufacturing. The details of these case studies and their use of data warehouse have been discussed in the Section 4. The analysis of the Table 2 shows that data warehouse is being used in many application domains. The Figure 10 clearly depicts the areas that are using data warehouse. It shows that medical and marketing areas are using data warehouse much more than the other domains, whereas manufacturing, agriculture, education, and government sector are rarely using data warehouse. The areas such as social media, construction, and finance are moderately using data warehouse technologies. The Figure 12 shows business-wise comparison and the Figure 13 shows the government-wise comparison of data warehouse usage.

The analysis shows that data warehouse technology have been adopted in business as well as in government organizations for managing their huge data and for decision making. Still many organizations have not gone for the adoption of DW technology. Either they do not realize its importance or there may be difficulties in its adoption. The reasons for ignoring the importance of implementing DW technology have been discussed in literature that include quite large investment in terms of capital, more time utilization, looking for intangible benefits are difficult, the last but not the least problems holding with recent data management systems' infrastructure etc.

References

[1] T. Ariyachandra, H. J. Watson, "Key organizational factors in data warehouse architecture selection", Decision Support Systems 49 (2010) 200–212.

[2] T. R. Sahama, P. R. Croll, "A Data Warehouse Architecture for Clinical Data Warehousing", in Roddick, J. F. and Warren, J. R., Eds. Proceedings Australasian Workshop on Health Knowledge Management and Discovery (HKMD 2007) CRPIT, 68, pages pp. 227-232, Ballarat, Victoria.

[3] W.H. Inmon., "DW 2.0 Architecture for the Next Generation of Data Warehousing", DM Review, Apr 2006, Vol. 16 Issue 4, p.8-25.

[4] W.H. Inmon, "Building the Data Warehouse", Third Edition, York: John Wiley & Sons, 2002.

[5] Hwang, Hsin-Ginn, et al. "Critical factors influencing the adoption of data warehouse technology: a study of the banking industry in Taiwan." *Decision Support Systems* 37.1 (2004): 1-21.

[6] Nilakanta, Sree, Kevin Scheibe, and Anil Rai. "Dimensional issues in agricultural data warehouse designs." *Computers and electronics in agriculture* 60.2 (2008): 263-278.

[7] B.A. Devlin, P.T. Murphy, An architecture for a business and information system, IBM Systems Journal 27 (1) (1988) 60 – 80

[8] W.H. Inmon, Building the Data Warehouse, Wiley, New York, 1996.

[9] S.R. Gardner, Building the data warehouse, Communications of the ACM 41 (9) (1998) 52 – 60.

[10] J.V.D. Hoven, Data warehousing: bringing it all together, Information Systems Management (1998 Spring) 92 – 96.

[11] R. Kimball, The Data Warehouse Toolkit, Wiley, New York, 1996.

[12] R.M.T. Lu, K.A. Mazouz, A conceptual model of data warehousing for medical device manufacturers, Proc. of the 22nd Annual EMBS International Conference 2000 (July).

[13] D. Powell, To outsource or not to outsource? Netwo rking Management (1993) 56 – 59.

[14] Y. Yao, H. He, Data warehousing and the Internet's impact on ERP, IT Professional (2000 March) 37–41.

[15] Rob, P., Coronel, C., 2006. Database Systems: Design, Implementation, and Management. Course Technology.

[16] Sen, A., Sinha, A.P., 2005. A comparison of data warehousing methodologies. Commun. ACM 48 (3), 79–84

[17] Kimball, R., 2002. The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling. John Wiley & Sons, Inc.

[18] Alsquor, M., Matouk, K., Owoc, M. L., A survey of data warehouse architectures:preliminary results. Proceedings of the Federated Conference on Computer Science*and Information Systems*,Wroclaw,2012,Sivut 1121-1126.

[19] Hackney, D., 2002. Architectures and Approaches for Successful Data Warehouses, Oracle White Paper.

[20] CHAKIR, Aziza, Hicham MEDROMI, and Adil SAYOUTI. "Actions for data warehouse success." Editorial Preface 4.8 (2013).

[21] Chaudhuri, S., Dayal, U., 1997. An Overview of Data Warehousing and OLAP Technology. ACM SIGMOD Record 26 (1), 64–74.

[22] Thakur, Garima, and Anjana Gosain. "A Comprehensive Analysis of Materialized Views in a Data Warehouse Environment." IJACSA) International Journal of Advanced Computer Science and Applications 2.5 (2011).

[23] Watson, H.J., Haley, B.J., 1997. Data warehousing: a framework and survey of current practices. J. Data Warehousing 2 (1), 10–17.

[24] Watson, H.J., Gerard, J.G., Gonzalez, L.E., Haywood, M.E., Fenton, D., 1999. Data warehousing failures: case studies and findings. J. Data Warehousing 4 (1), 44–55

[25] www.informationweek.com/mobile/mobile-devices/gartner-21-billion-iot-devices-to-invade-by-2020/d/d-id/1323081

[26] www.gartner.com/newsroom/id/3165317

[27] Watson, Hugh J., Dale L. Goodhue, and Barbara H. Wixom. "The benefits of data warehousing: why some organizations realize exceptional payoffs." *Information & Management* 39.6 (2002): 491-502.

[28] Joseph, Madhuri V. "Significance of Data Warehousing and Data Mining in Business Applications." *International Journal of Soft Computing and Engineering (IJSCE) ISSN* (2013): 2231-2307.

[29] Thusoo, Ashish, et al. "Data warehousing and analytics infrastructure at facebook." *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*. ACM, 2010.

[30] Chowdhury, Rajdeep, et al. "Implementation of Central Dogma Based Cryptographic Algorithm in Data Warehouse Architecture for Performance Enhancement." International Journal of Advanced Computer Science & Applications 1.6: 29-34.

[31] Park, Taeil, and Hyoungkwan Kim. "A data warehouse-based decision support system for sewer infrastructure management." *Automation in Construction* 30 (2013): 37-49.

[32] Ryals, Lynette, and Adrian Payne. "Customer relationship management in financial services: towards information-enabled relationship marketing." Journal of strategic marketing 9.1 (2001): 3-27.

[33] Goyal, Monika, and Rajan Vohra. "Applications of data mining in higher education." International journal of computer science 9.2 (2012): 113.

[34] Bhedi, Vaibhav R., Shrinivas P. Deshpande, and Ujwal A. Lanjewar. "Data Warehouse Architecture for Financial Institutes to Become Robust Integrated Core Financial System using BUID." International Journal of Advanced Research in Computer and Communication Engineering 3.3 (2014): 2278-102.

[35] Evans, R. Scott, James F. Lloyd, and Lee A. Pierce. "Clinical use of an enterprise data warehouse." AMIA Annual Symposium Proceedings. Vol. 2012. American Medical Informatics Association, 2012.

[36] Stolba, Nevena, and A. Min Tjoa. "The relevance of data warehousing and data mining in the field of evidence-based medicine to support healthcare decision making." International Journal of Computer Systems Science and Engineering 3.3 (2006): 143-148.

[37] Schubart, Jane R., and Jonathan S. Einbinder. "Evaluation of a data warehouse in an academic health sciences center." International journal of medical informatics 60.3 (2000): 319-333.

[38] Liu, Baoyan, et al. "Data processing and analysis in real‐world traditional Chinese medicine clinical data: challenges and approaches." Statistics in medicine 31.7 (2012): 653-660.

[39] Leitheiser, Robert L. "Data quality in health care data warehouse environments." System Sciences, 2001. Proceedings of the 34th Annual Hawaii International Conference on. IEEE, 2001.

[40] Yoo, Sooyoung, et al. "Electronically implemented clinical indicators based on a data warehouse in a tertiary hospital: its clinical benefit and effectiveness." International journal of medical informatics 83.7 (2014): 507-516.

[41] Prokosch, Hans-Ulrich, and T. Ganslandt. "Perspectives for medical informatics." Methods Inf Med 48.1 (2009): 38-44.

[42] Adlassnig, Klaus-Peter, et al. "Fuzziness in healthcare-associated infection monitoring and surveillance." Norbert Wiener in the 21st Century (21CW), 2014 IEEE Conference on. IEEE, 2014.

[43] Chong, Heap Yih, Rosli Mohamad Zin, and Siong Choy Chong. "Employing data warehousing for contract administration: e-dispute resolution prototype." Journal of Construction Engineering and Management 139.6 (2012): 611-619.

[44] Chau, Kwok-Wing, et al. "Application of data warehouse and decision support system in construction management." Automation in construction 12.2 (2003): 213-224.

[45] Chen, Wenzhe. "The Application of Data Warehouse Technology in Modern Finance." 2015 International Conference on Advances in Mechanical Engineering and Industrial Informatics. Atlantis Press, 2015.

[46] Lin, Zhonglin, et al. "Banking intelligence: application of data warehouse in bank operations." Service Operations and Logistics, and Informatics, 2008. IEEE/SOLI 2008. IEEE International Conference on. Vol. 1. IEEE, 2008.

[47] Shaw, Michael J., et al. "Knowledge management and data mining for marketing." Decision support systems 31.1 (2001): 127-137.

[48] Ngai, Eric WT, Li Xiu, and Dorothy CK Chau. "Application of data mining techniques in customer relationship management: A literature review and classification." Expert systems with applications 36.2 (2009): 2592-2602.

[49] Nedeva, Veselina Ivanova. "ANALYSIS OF MARKETING INFORMATION SYSTEMS AND CONCEPTION OF AN INTEGRATED MARKETING INFORMATION SYSTEM." International Journal of Computing 3.2 (2014): 127-133.

[50] Payton, Fay, and Debra Zahay. "Why doesn't marketing use the corporate data warehouse? The role of trust and quality in adoption of data-warehousing technology for CRM applications." Journal of Business & Industrial Marketing 20.4/5 (2005): 237-244.

[51] Thomas, Davenport, et al. "Data to Knowledge to Results, Building an Analytic Capability." California Management Review 43.2 (2001).

[52] Cunningham, Colleen, Il-Yeol Song, and Peter P. Chen. "Data warehouse design to support customer relationship management analyses." Proceedings of the 7th ACM international workshop on Data warehousing and OLAP. ACM, 2004.

[53] Watson, Hugh J., Celia Fuller, and Thilini Ariyachandra. "Data warehouse governance: best practices at Blue Cross and Blue Shield of North Carolina." Decision Support Systems 38.3 (2004): 435-450.

[54] Sarkar, Anirban. "Data Warehouse Requirements Analysis Framework: Business-Object Based Approach." International Journal 3 (2012).

[55] Diana, Nova Eka, and Aan Kardiana. "Comprehensive Centralized-Data Warehouse for Managing Malaria Cases." International Journal of Advanced Computer Science & Applications 1.6: 40-46.

[56] N.L. Sarda, Temporal issues in data warehouse systems, Database Applications in Non-Traditional Environments ’99, TheProceedings of the 1999 International Symposium on Database Application in Non-traditional Environments (DANTE ’99), IEEE Computer Society, Los Alamitos, 1999, pp. 27–34

[57] J.-B. Yang, N.-J. Yau, Application of case-based reasoning in construction engineering and management, Proceedings of the Third Congress held in conjunction with A/E/C Systems 1996, Computing in Civil Engineering, American Society of Civil Engineers, New York, 1996, pp. 663–669.

[58] K.W. Chau, Y. Cao, M. Anson, J.P. Zhang, Application of data warehouse and decision support system in construction management, Automation in Construction 12 (2) (2002) 213–224.

[59] J. Vanegas, P. Chinowsky, Computing in Civil Engineering, American Society of Civil Engineers, New York, 1996.

[60] J. Dyche, e-Data Turning Data into Information with Data Warehousing, Addison-Wesley, Reading, 2000

Improving and Extending Indoor Connectivity Using Relay Nodes for 60 GHz Applications

Mohammad Alkhawatra
Department of Electronics and Communications Engineering
Al-Ahliyya Amman University
Amman, Jordan

*Nidal Qasem
Department of Electronics and Communications Engineering
Al-Ahliyya Amman University
Amman, Jordan

***Abstract*—a 60 GHz wireless system can provide very high data rates. However, it has a tremendous amount of both Free Space Path Loss (FSPL) and penetration loss. To mitigate these losses and extend the system range; we propose techniques for using relay nodes. The relay node has been positioned correctly in order to shorten the distance between a source and a destination, this gave a reduction in the FSPL value. In addition, the positioning of the relay node correctly gave an alternative Line of Sight (LoS) to overcome the penetration loss caused by human bodies. For the last challenge, the considerably short range of the wireless network in the 60 GHz band, the range has been extended by applying the multi-hop communication with the concept of relay nodes selection. The length of the room was doubled and still get the same losses as if there was no expansion. All three techniques were modeled inside 'Wireless InSite' by three scenarios. The first scenario was a conference room with no obstacles to focus on FSPL. In the second scenario, the same conference room was modeled but human bodies have been taken into consideration to check the penetration loss effect. The final scenario was the extended version of the first scenario to deal with the small range issue.**

Keywords—60 GHz; Indoor Wireless; Multi-hop; Relay; Relay Selection

I. INTRODUCTION

The millimeter wave technology has been known for many decades, and it has been deployed for military applications. With the advances of process technologies and low-cost integration solutions, this technology has started to gain a great deal of momentum from academia, industry, and standardization body [1].

The major quality of 60 GHz is the huge globally license-free spectrum between 57-66 GHz which will support very high data rate wireless applications [2]. One of the main challenges facing the 60 GHz technology is the heavy attenuation characteristics of the millimeter waves. As an example, a 60 GHz system has to deal with more than 20 dB greater FSPL than an equivalent 5 GHz system since the FSPL increases with the square of the carrier frequency [3]. Another challenge for using 60 GHz is the penetration loss which is also very high in the 60 GHz band. In a typical indoor environment, the LoS propagation path between two devices at 60 GHz may completely be blocked by surrounding objects and human bodies. When a 60 GHz link is blocked reflections from the surfaces can be exploited to sustain the link connectivity between the devices which will add more losses [4]. Short range is a huge challenge for 60 GHz system. For point-to-point indoor communication in order to get up to 10 m range, an antenna with high gain of 15 dBi or higher is required [5].

The effective interference levels for 60 GHz are lower than what for the congested 2–2.5 GHz and 5–5.8 GHz regions [6]. However, in some cases where dense 60 GHz wireless network existed, the interference level is considerable. So, interference mitigation techniques are needed [7]. Directional antenna proposed to overcome high values of FSPL in [8], but the signals with the proposed technique can be easily blocked by any obstacle. Beamforming or beam steering is proposed in [9] to overcome blockage in directional antennas and enhance their performance in 60 GHz system. However, the proposed scheme adds more overhead and complexity to the system. Many methods proposed to solve those challenges, but the easiest and most efficient method is the using of relay nodes [10]. The first study of using relay nodes with 60 GHz is provided in [4]. The paper shows that the value of FSPL can be reduced by more than 33% by using relay nodes. The proper positioning is provided in the paper. However, the simulation results based on device to device communication network. In [11] a relay selection scheme is proposed to replace a long direct path with several multi-hop paths to improve the network throughput. However, to the best of our knowledge, the small range mitigation by multi-hop communication isn't studied yet. By relaying signal from the source to the destination, the long bath between the source and the destination is then broken into short paths which in turn reduce the FSPL [12]. Indirect path via relay node can provide LoS in some cases where the direct link between the source and the destination is blocked. The main contributions of this paper include the following. (1) Finding the best position of relay node to reduce the FSPL between transmitter and receiver. (2) Finding the best position of relay node to reduce the penetration loss, human bodies, by providing LoS in case of blockage. (3) Extending range using same parameters by selecting relay nodes correctly in multi-hop communication. Interference mitigation mechanisms have been discussed in this paper.

This paper is divided into 6 sections. Section 1 gives the overview. Section 2 is the system model. Section 3 contains the positioning of relay nodes. Relay nodes selection and interference mitigation are in Section 4. Simulations setup and results are presented in Section 5. At the end conclusion is presented in Section 6.

II. System Model

A. FSPL Improvement Model

The chosen criterion to test improvement is the FSPL, because of its ease of being calculated and at the same time its importance in link budget [4]. In order to model the performance of the system with and without using of relay nodes, two cases need to be modeled. First case is the direct case, where the source can reach to the destination directly. The second case is the indirect case, where the source will reach to the destination through the relay. The Source (S) point is assumed to be fixed access point while the Destination (D) is assumed to be randomly placed in the area. The Relay Node (R) is assumed at the midpoint of the area where 60 GHz wireless system is going to be used. The suggested area for this study will take the shape of a circle with an appropriate radius value. So, the source can reach all points on the circle and the relay node will be at the center of the circle. The area has radius C. P is the link connecting source to destination which is the random variable, C is the link connecting source to relay node, and I is the link connecting relay node to destination, as shown in Fig. 1.

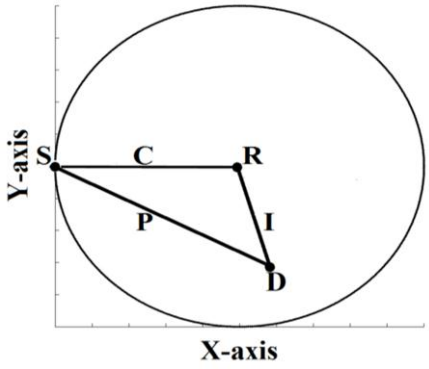


Fig. 1. Study area

Destination is randomly located in the circle, each location with different values on X and Y axis, the location of the destination is represented with the values (x, y). If the destination location considered to be uniformly distributed on the circle, then the Probability Density Function (PDF) of the location in the two dimensional plane as follows [13]:

$$f(x,y) = \begin{cases} \frac{1}{area} & for (x,y) \in area \\ 0 & otherwise \end{cases} \quad (1)$$

The factor that will decide if the link FSPL improved or not is the distance between the source and the destination according to the following formula [14]:

$$FSPL_{(dB)} = 32.4 + 20 \log d_{km} + 20 \log f_{MHz} \quad (2)$$

Where d is distance and f is frequency. So to get the distance, the conversion to polar coordination will be helpful. To get the PDF for r and θ from $f(x,y)$, this can be done using the Jacobian of the transformation matrix [15]:

$$J(x,y) = \frac{1}{r} \quad (3)$$

$$f(r,\theta) = \frac{f(x,y)}{|J(x,y)|} = \frac{r}{\pi C^2} \quad (4)$$

while θ is uniformly distributed over $0 - 2\pi$ then:

$$f(r) = \frac{2r}{C^2} \quad (5)$$

For the first case, where the link directly connects the source to the destination, the maximum link length is $P_{max} = 2C$. In the second case, there are two links: one that connects the source to the relay node which is C, this value is fixed because both of source and relay node are fixed. The other link connects the relay node to the destination, it's maximum length is $I_{max} = C$. So, longer one among the two links need to be compared with direct link in order to see whether the relay node made an improvement on the direct link or not. To find the maximum between two independent random variables, this can be done by multiplying their Cumulative Distribution Function (CDF) [16]. For first link, the CDF is $F(C) = 1$ because it is a constant value. For second link, the CDF is equal to the integral of $f(r)$, which is equal to $\frac{r^2}{C^2}$ and so:

$$F_{max}(r) = \frac{r^2}{C^2} \quad (6)$$

To inspect the reduction in the distance, expected value is a reasonable way:

$$E(r) = \int r f(r) dr \quad (7)$$

Since there is only one link in the first case then it's PDF equal $f(r)$, but for the second case we need the maximum PDF between the two links which is:

$$f_{max}(r) = \frac{dF_{max}(r)}{dr} = \frac{2r}{C^2} \quad (8)$$

Since $f(r)$ and $f_{max}(r)$ are equal, then $f(r)$ can be used as a PDF for both cases. By taking the extreme values of r in the two cases P_{max} and I_{max} then: (1) $E(r) = 2c * \frac{2*2C}{C^2} = 8$ for first case. (2) $E(r) = c * \frac{2*C}{C^2} = 2$ for second case. So, in the second case, where the relay node was fixed at the center, the distance can be reduced up to its one fourth of the no relay case. This reduction in distance could be translated to FSPL reduction by using (2). So, the reduction in FSPL could reach -12 dB at the extreme values of r.

B. Human Body at 60 GHz Model

In order to evaluate the performance of the system in the presence of humans a reliable characterization of propagation through human body at 60 GHz channel is needed. Indeed, the close proximity of antennas with the human body may result in significant changes in the input impedance, radiation patterns, antenna efficiency, and energy absorption of the signal induced in human bodies. Furthermore, the results might fluctuate due to differences and variations of the dielectric properties of biological tissues [17]. Typical values of the complex permittivity at 60 GHz for main human body tissues like skin, fat, muscles, and pure water at 20^{o} C are summarized in Table I [18].

TABLE II. HUMAN TISSUES COMPLEX PERMITTIVITY AT 60 GHZ

Human Tissues	ε_c
Skin	7.98 − j 10.91
Fat	2.51 − j 0.84
Muscles	12.85 − j 15.74
Pure Water (20 ° C)	11.9 − j19.5

Where:

$$\varepsilon_c = \varepsilon_r - j\frac{\sigma}{\omega\varepsilon_o} \quad (9)$$

ε_c is Complex permittivity, ε_r is relative permittivity, ω is angular velocity (rad/s), and ε_o is free space permittivity (8.85×10^{-12} F/m). The human body is simulated as a parallelepiped circumscribed with pure water cylinder model with the following dimensions: the lengths of sides of the basic rectangle is equal to 0.305 m, the height is equal to 1.7 m, and the thickness is equal to 0.305 m [19]. Based on (9), the conductivity of pure water at 20^o C is 65.06 S/m, the conductivity of skin is 36.4 S/m, the conductivity of fat is 2.802 S/m, and the conductivity of muscles is 52.51 S/m.

III. RELAY NODE POSITIONING

Calculations were performed in MATLAB [20] not only to prove the reduction of FSPL by using relay node, but also to find relay node location that gives maximum reduction. This was done on a room of 10 m x 10 m, the source is fixed at the top of the room with height of 3 m, the receivers are assumed to be at 1 m height and deployed all over the room with total number of 81 receivers, and the relay node fixed at the midpoint of the room with two different heights once at 1 m and once at 3 m, as shown in Fig. 2. Then a comparison between the direct link, which connects the source to a destination, with the corresponding link, that connects the source to the same destination but with relay node in between, have been calculated by (2). This procedure has been repeated at all destinations for two different heights and for different places of the relay nodes. Turned out there is no noticeable different in FSPL reduction between 1 m and 3 m height for relay node. So, the FSPL enhancement for both cases considered same. Midpoint position for relay node has achieved the best average of FSPL reduction with respect to the direct link case, as shown in Fig. 3.

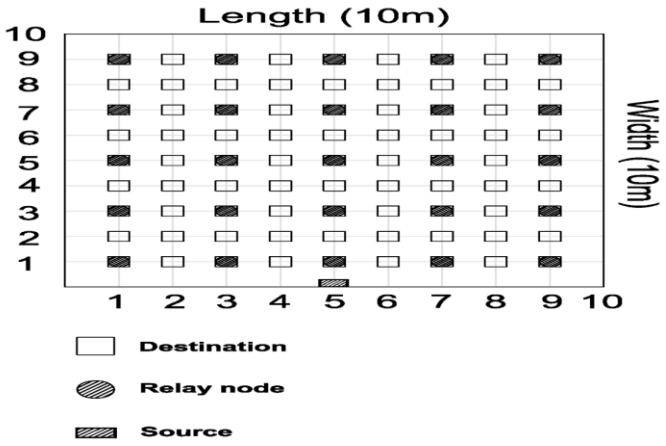


Fig. 2. Schematic of the procedure done via MATLAB

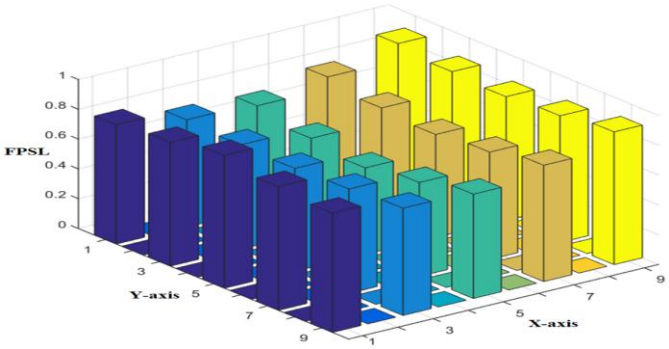


Fig. 3. Average percentage of FSPL reduction with Relay at 1m height

IV. RELAY NODES SELECTION AND INTERFERENCE MITIGATION

Coverage area is a big issue in 60 GHz wireless system, while the high frequency can cause high losses the distance can also cause very high losses since the FSPL increases with the square of the distance. Relay nodes can be used to solve this issue by the concept of multi-hop communication, where the relay that fits every hop's optimal distance need to be selected in order to increase the range of the system [21].

Relay nodes selection in this study is to select the relay node or group of relay nodes that will result in the smallest distance between any two communicating devices in the path between the source and the destination: source to relay node, relay node to another relay node, and relay node to destination. So, the path consists of source, destination, and relay nodes in between. All devices in the system are assumed to be capable of measuring the signal strength coming from other devices on their antenna elements. In order to make the comparison between distances that separate each two devices, device location should be known. In this paper, relay nodes are considered to be the reference devices where all communication going through them and they are assumed to be able to contact each other. Device location estimation in the system, while reference devices have information about their coordinates, can be done using the received signal strength which allows any two communicating devices to estimate the distance between them [22]. Each relay node will have a database that contains the signal strength, which can be translated to distance for all devices in its range [10], and distances between relay nodes are known because their coordinates are known. Since all relay nodes can contact each other, a replication of their databases can be available at each one of them. The Raspberry Pi, a full functioning computer, will be connected to the source. It will keep updating the databases received from the relay nodes and control the source. After the distances are available the next step is the comparison between them. This will be done by Raspberry Pi which will select the relay nodes that the message need to go through to get the best route from source to destination. To get the best route, the longest distance between two communicating devices in each path need to be compared together. So, the longest distance in path one will be compared with the longest distance in path two and path three and so on, then the path that have the smallest value of longest distance will be selected. Since the relay nodes are assumed to be fixed, it is easy to find

the followings: best paths between source and each relay node (p_i), where i is the relay node number, the longest distance between two communicating devices in each $p_i(dmax_i)$, and the number of hops (n) to reach from source to each relay node. Raspberry Pi supports many programming languages; the following code is written to work in C++ environment:

Algorithm 1 Relay Selection Algorithm

Input: each relay database, p_i,$dmax_i$
Output: best route from source to destination

```
1: mindmax = dmax_1;
2: bestp = p_1;
3: for (int i = 1; i < L; i++) // L= # of relays
       {
4:       if (dij > 0) // dij is the distance between relay i and
         destination j
             {
5:             if (dmax_i < dij)
                  {
6:                  dmax_i = dij;
                  }
7:           }
8:       if (dmax_i < dmax_1)
             {
9:             mindmax = dmax_i;
10:        bestp = p_i; // bestp is the best route from source to
           destination
             }
11:    } // end for
```

If there are a neighboring networks to the network in study and they are operating in the same channel, the Signal to Interference plus Noise Ratio (SINR) values are degraded. Since the data rate is proportional to the value of SINR this reduction may prevent the supporting of the required data rates [7]. The effective interference levels for 60 GHz band are relatively low. So, it could be neglected in some cases. But in some other cases where dense 60 GHz wireless network existed, the interference levels are considerable and need to be mitigated. Although an antenna array with many antenna elements is mainly used to maximize the very low level of power received at receivers in 60 GHz band by beamforming, it can also be used to mitigate interference [23]. One of the famous interference mitigation mechanisms is based on coordinator setup, which consists of coordinator and several devices within its transmission range [24]. A coordinator setup is defined as a wireless data communication system which allows a number of independent devices to communicate with each other, one device is required to be the coordinator of the system. The devices measure their signal and interference power levels over multiple fixed periods of time and report back to the coordinator, the coordinator then determine the schedule of transmission that will avoid transmission in the presence of interference. The coordination could be for single network or for multi networks together [25]. The following technique is more related to this study, since it is considered a relay application. The basic idea is that by reducing the transmitted power, by sending data over a portion of the transmitting power, this will reduce the chance for interference to happen. Since relay can reduce the distance between source and destination, and so the required power to transmit data will be reduced. Mitigation can be done by getting the value of SNR for the direct link between source and destination, and SNR for link with relay, the link with the higher SNR need to be selected. The transmitting power of the relay is assumed to be changeable [26]. Same Raspberry Pi used in this section could be used here to compare the values of SNR and then select the best link. If the best link was the one with relay, Raspberry Pi will make the relay to transmit at lower power. At the same time Raspberry Pi will keep SNR value acceptable to get the promised data rate.

V. Simulation Results

'Wireless InSite' is a simulation tool [27], which will be used to analyze the impact of relay nodes on 60 GHz wireless system performance in the indoor environment.

A. Setup

1) FSPL Simulation: A conference room scenario will be modeled to study the effects of relay node and to verify that the best position for relay node is at the midpoint. Dimensions of the room is 10 m x10 m, source is mounted at the top with height of 3 m, destinations are spread all over the room at 1 m of height with 1 m separating space between them with total of 81 receivers. Relay nodes are positioned at five different places with 1 m of height, as shown in Figs. 4 to 6. Used antennas have an omnidirectional radiation pattern with gain of 8.5 dBi [28]. Since the maximum transmitted power is limited to 10 dBm by taking the Radio Frequency (RF) safety issues into account [6], the input power to the source is limited to 5 dBm in this paper, since 5 dBm makes the received power level at destinations in the range of -55 dBm, which necessary to satisfy the required gigabits per second data rate [29]. Channel 2 as defined by Institute of Electrical and Electronics Engineers (IEEE) 802.11 ad with carrier frequency of 60.48 GHz and 2.16 GHz of bandwidth is chosen because it is completely covered in all countries [30]. The electric parameters of the materials used to build the room are presented in Table II [27].

TABLE III. Electrical Parameters of the Materials

The User Interface	Material	Relative Electrical Permittivity	Conductivity, (S/m)	Thickness, (m)
Ceiling &Floor	Concrete	7	0.015	0.3
Walls	Brick	4.44	0.001	0.125
Doors	Wood	5	0	0.03
Windows	Glass	2.4	0	0.003

2) Penetration Loss Simulation: In this subsection the effects of the midpoint relay node, best position for FSPL, on penetration loss will be presented for two different relay heights. Same as the previous conference room scenario, but with only five destinations and two relays at midpoint with height of 1 m and 3 m with taking the obstacles into consideration, as shown in Figs. 7 and 8. The obstacles are the human bodies in the area, human bodies are modeled into 'Wireless InSite' based on the human model presented in Section 3.

3) Short Range Simulation: The dimensions of the conference room after extending is 10 m x 20 m. The source mounted at the top of the room with height 3 m. Three receivers are fixed at the far end of the room at 1 m height. Relay nodes are positioned at 4 different places. One at the midpoint of original room, since it has been proved to be the best position, the other 3 relay nodes are fixed in the new extension of the room, each one of them far from the midpoint relay node by same distance. All relay nodes at height 3 m, as shown in Figs. 9 to 11.

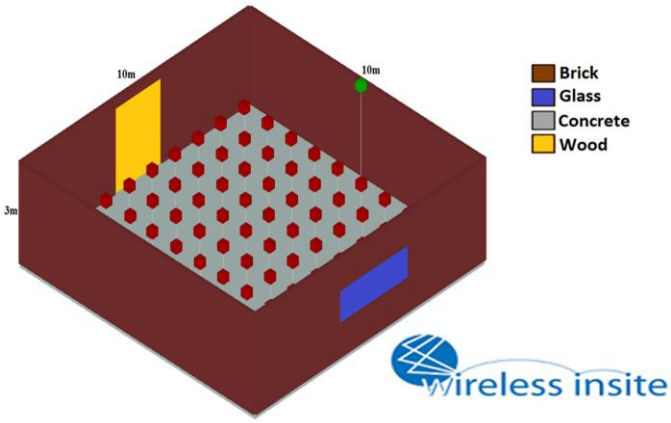


Fig. 4. 3D view of the room which green box is source and red is destination

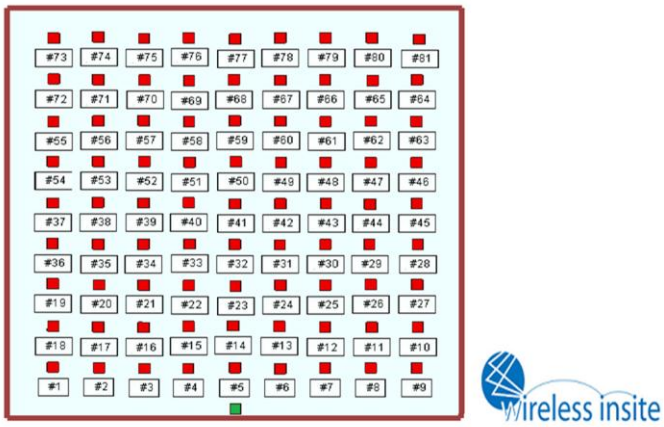


Fig. 5. 2D top view of the room which shows all destinations with the source

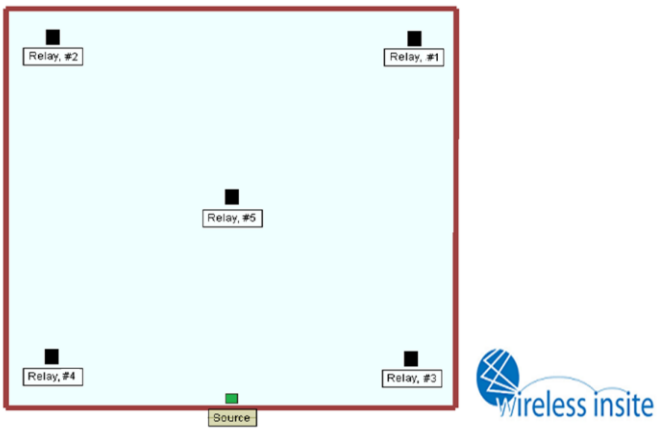


Fig. 6. 2D top view of the room which shows all relays with the source

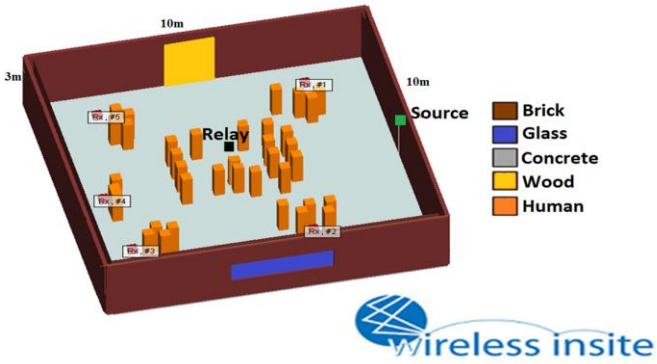


Fig. 7. 3D view of the room which green box is source, red is destination, and black is relay

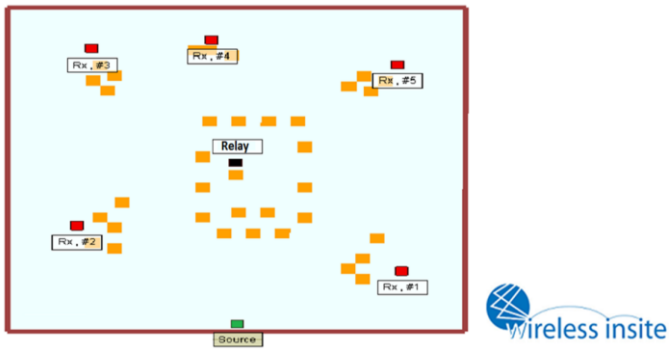


Fig. 8. 2D top view of the room which shows all destinations, humans, and source

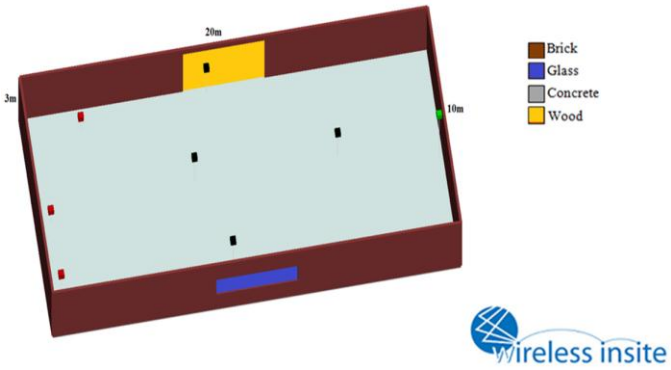


Fig. 9. 3D view of the room which green box is source, red is destination, and black is relay

Fig. 10. 2D top view of the room which shows all relays with the source.

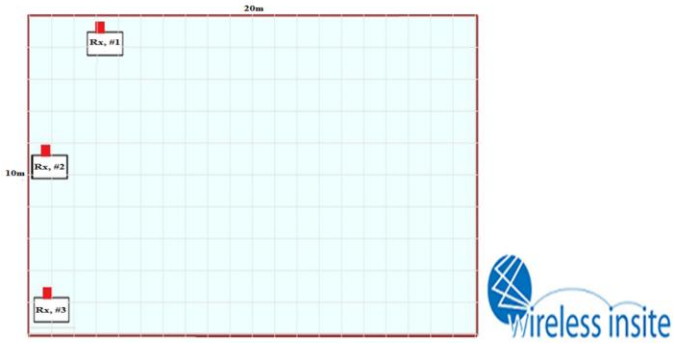


Fig. 11. 2D top view of the room which shows all destinations with the source

B. Results

For the first scenario the values of FSPL are keep going down while the relay node is getting closer to the midpoint of the room. Relay 1 and Relay 2 have same distance from the source, so they have same FSPL. Same thing with Relay 3 and Relay 4. The best FSPL values happened when the relay node was at midpoint which was the case for Relay 5, as shown in Figs. 12 and 13. For the second scenario, relay node at 3 m height can provide another path in the cases where the direct path is blocked. But for relay node at 1 m height the path can be blocked from the source by human bodies. This makes 3 m the best position to mitigate penetration loss, as shown in Fig. 14. So, the best location for the relay node in order to minimize the two types of losses is at the midpoint at the top of the room.

For the third scenario there are three hops: (1) From the source to Relay 1. (2) From Relay 1 to Relays 2, 3, and 4. (3) From Relays 2, 3, and 4 to the receivers. Since in first-hop FSPL is fixed and second-hop FSPL is fixed from Relay 1 to the other 3 relays. So, first and second hops have no effects in relay node selection procedure. This make the third-hop is the only hop that effects in relay node selection. So, the relay node in the third-hop which has the smallest FSPL to the destination will be selected, then the path will be like this: Source→Relay 1→ Selected relay from third-hop→Destination. The comparison shown in Figs. 15 and 16, Relay 4 will be selected if the destination is the Rx#1. This will make the path as the following: Source→Relay 1→Relay 4→Rx#1. Same procedure will be followed with Rx#2 and Rx#3, this will result in the selection of Relay 3 and Relay 2 respectively.

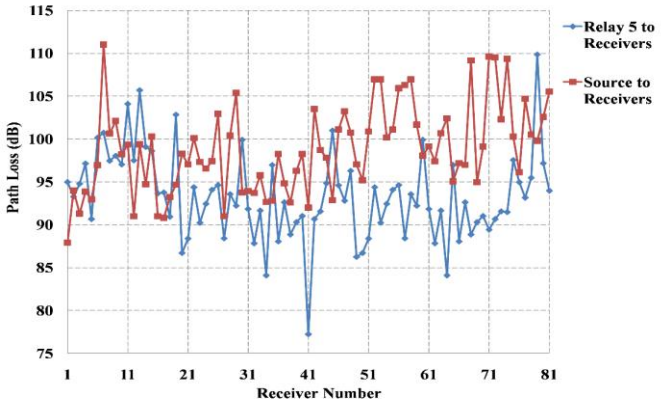


Fig. 12. FSPL comparison between the direct case and the Relay 5 case

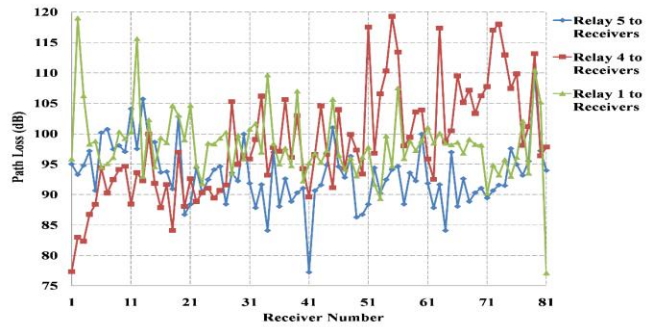


Fig. 13. FSPL comparison between from Relay 1, 4, and 5 to receivers

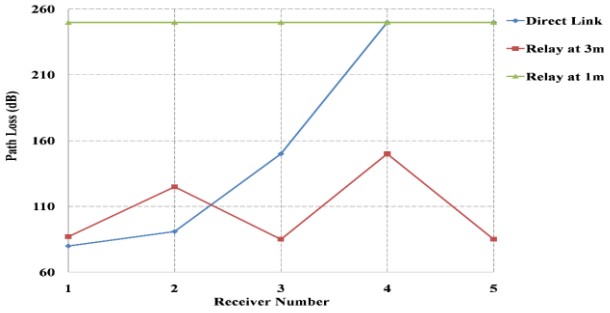


Fig. 14. Path loss, penetration loss and FSPL, at each receiver with and without relay

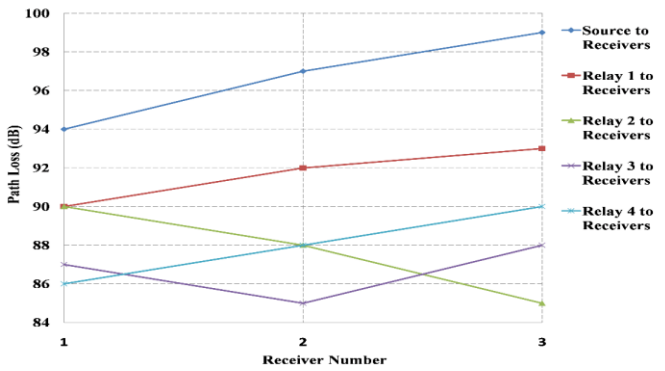


Fig. 15. FSPL for all paths from source and relays to receivers

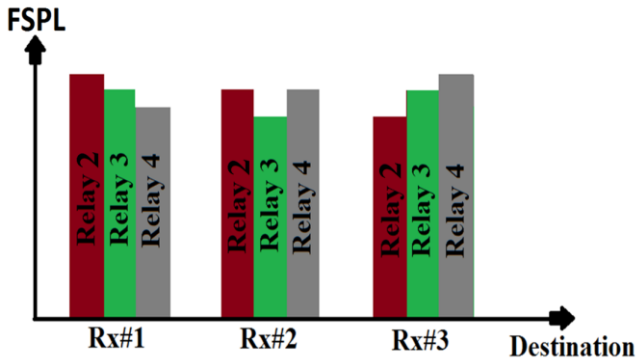


Fig. 16. FSPL comparison for third-hop paths

VI. Conclusions

In order to improve the performance in 60 GHz wireless networks, the relay nodes have been considered. Because of huge value of propagation loss, distance, and obstacles are the main metric for relay positioning. The results show that the best relay position for mitigating FSPL is the midpoint, the maximum reduction in FSPL is -12 dB. The best position for mitigating penetration loss is at the top of the room. So, in order to maximize the performance of the network and satisfy the promised high data rate the relay need to be positioned at the midpoint on the ceiling of the room. Another issue is the very short range of 60 GHz network, relay selection technique had been used to increase the coverage area. So, the room dimensions could be doubled and still get the same performance. Many interference mitigation mechanisms can be applied on this system to deal with interference at this communication band.

References

[1] Olver, A. D. "Millimetrewave systems-past, present and future." Radar and Signal Processing, IEE Proceedings F. Vol. 136. No. 1. IET, 1989.

[2] Cai, Lin X., et al. "REX: a randomized exclusive region based scheduling scheme for mmWave WPANs with directional antenna." Wireless Communications, IEEE Transactions on 9.1 (2010): 113-121.

[3] Zheng, Guanbo, et al. "A robust relay placement framework for 60GHz mmWave wireless personal area networks." Global Communications Conference (GLOBECOM), 2013 IEEE. IEEE, 2013.

[4] Genç, Zülküf, et al. "Improving 60 ghz indoor connectivity with relaying." Communications (ICC), 2010 IEEE International Conference on. IEEE, 2010.

[5] Liu, Duixian, and R. Sirdeshmukh. "A patch array antenna for 60 GHz package applications." Antennas and Propagation Society International Symposium, 2008. AP-S 2008. IEEE. IEEE, 2008.

[6] Yong, Su Khiong, and Chia-Chin Chong. "An overview of multigigabit wireless through millimeter wave technology: potentials and technical challenges." EURASIP Journal on Wireless Communications and Networking 2007.1 (2006): 1-10.

[7] Park, Minyoung, and Praveen Gopalakrishnan. "Analysis on spatial reuse and interference in 60-GHz wireless networks." Selected Areas in Communications, IEEE Journal on 27.8 (2009): 1443-1452.

[8] Williamson, M. R., G. E. Athanasiadou, and A. R. Nix. "Investigating the effects of antenna directivity on wireless indoor communication at 60 GHz." Personal, Indoor and Mobile Radio Communications, 1997. Waves of the Year 2000. PIMRC'97., The 8th IEEE International Symposium on. Vol. 2. IEEE, 1997.

[9] Li, Bin, et al. "Efficient beamforming training for 60-GHz millimeter-wave communications: a novel numerical optimization framework." Vehicular Technology, IEEE Transactions on 63.2 (2014): 703-717.

[10] Rehman, Waheed Ur, Tabinda Salam, and Xiaofeng Tao. "Relay selection schemes in millimeter-wave WPANs." Wireless Personal Multimedia Communications (WPMC), 2014 International Symposium on. IEEE, 2014.

[11] Qiao, Jian, et al. "Enabling multi-hop concurrent transmissions in 60 GHz wireless personal area networks." Wireless Communications, IEEE Transactions on 10.11 (2011): 3824-3833.

[12] Sivakumar, Vignesh Vellimalaip. Relay Positioning for Energy Efficiency and Improved Performance of Cooperative Wireless Networks. Diss. Auburn University, 2014.

[13] Miller, Scott, and Donald Childers. Probability and random processes: With applications to signal processing and communications. Academic Press, 2012.

[14] Aragon-Zavala, Alejandro. Antennas and propagation for wireless communication systems. John Wiley & Sons, 2008.

[15] Leon-Garcia, Alberto, and Alberto. Leon-Garcia. Probability, statistics, and random processes for electrical engineering. Pearson/Prentice Hall, 2008.

[16] David, Herbert Aron, and Haikady Navada Nagaraja. Order statistics. John Wiley & Sons, Inc., 1970.

[17] Chahat, Nacer, Maxim Zhadobov, and Ronan Sauleau. "Broadband tissue-equivalent phantom for BAN applications at millimeter waves." Microwave Theory and Techniques, IEEE Transactions on 60.7 (2012): 2259-2266.

[18] Gustafson, Carl, and Fredrik Tufvesson. "Characterization of 60 GHz shadowing by human bodies and simple phantoms." Antennas and Propagation (EUCAP), 2012 6th European Conference on. IEEE, 2012.

[19] Qasem, Nidal. Enhancing wireless communication system performance through modified indoor environments. Diss. © Nidal Qasem, 2014.

[20] MathWorks (2015). Retrieved fromhttp://www.mathworks.com/products/matlab/

[21] Song, Kan, Ran Cai, and Danpu Liu. "A fast relay selection algorithm over 60GHz mm-wave systems." Communication Technology (ICCT), 2013 15th IEEE International Conference on. IEEE, 2013.

[22] Patwari, Neal, et al. "Relative location estimation in wireless sensor networks." Signal Processing, IEEE Transactions on 51.8 (2003): 2137-2148.

[23] Tseng, Yi-Hsien, Eric Hsiao-kuang Wu, and Gen-Huey Chen. "Maximum traffic scheduling and capacity analysis for IEEE 802.15. 3 high data rate MAC protocol." Vehicular Technology Conference, 2003. VTC 2003-Fall. 2003 IEEE 58th. Vol. 3. IEEE, 2003.

[24] An, Xueli, et al. "Performance analysis of synchronization frame based interference mitigation in 60 GHz WPANs." IEEE Communications Letters 14.5 (2010): 471-473.

[25] Park, Hyunhee, et al. "Multi-hop-based opportunistic concurrent directional transmission in 60 GHz WPANs." Multimedia Tools and Applications 74.5 (2015): 1627-1644.

[26] Al Sukkar, Ghazi, Zaid A. Shafeeq, and Ahmad Al Amayreh. "Best relay selection in a multi-relay nodes system under the concept of

cognitive radio." Information and Communication Systems (ICICS), 2015 6th International Conference on. IEEE, 2015.

[27] Remcom. (2015, November 15). Retrieved fromhttp://www.remcom.com/electromagnetic-applications/,

[28] Peraso. Peraso PRS4000/PRS1025 60 GHz WiGig Chipset Product Brief. 2015.

[29] Verma, Lochan, Mohammad Fakharzadeh, and Sunghyun Choi. "WiFi on Steroids: 802.11 ac and 802.11 ad." Wireless Communications, IEEE 20.6 (2013): 30-35.

[30] Perahia, Eldad, and Michelle X. Gong. "Gigabit wireless LANs: an overview of IEEE 802.11 ac and 802.11 ad." ACM SIGMOBILE Mobile Computing and Communications Review 15.3 (2011): 23-33.

Localization and Monitoringo of Public Transport Services Based on Zigbee

Izet Jagodic
Faculty of Electrical Engineering
University of East Sarajevo
East Sarajevo, Bosnia and Herzegovina

Suad Kasapovic, Amir Hadzimehmedovic, Lejla Banjanovic-Mehmedovic
Faculty of Electrical Engineering,
University of Tuzla
Tuzla, Bosnia and Herzegovina

***Abstract*—Regular and systematic public transport is of great importance to all residents in any country, in the city and on commuter routes. In our environment, users of public transport can track the movement of vehicles with great difficulty, given that the current system does not meet the necessary criteria, and does not comply with the functioning of transport system. The aim of the final paper is to show the development of such a system using ZigBee and Arduino platforms. This paper shows an example of use the technologies mentioned above, their main advantages and disadvantages, with the emphasis on communication between the device and its smooth progress. In order to show the way in which the system could function, a simple mesh network was created, consisting of coordinator, routers for data distribution and end devices representing the vehicles. To view the results a web application was developed using open-source tool which is for display of the collected data on the movement of nodes in the network.**

Keywords—Wireless mesh network; Zigbee; Xbee; microcontroller; web development; integration

I. INTRODUCTION

Public transport plays a vital infrastructure link between individual cities and states, and even the major geographical regions. There are a large number of users of public transport and, therefore, we can say that its proper and smooth functioning is the imperative. However, the number of users varies and depends directly on the quality of services that the system provides. These services are conditioned by everyday occurrences such as traffic jams, accidents on the road, and some unexpected weather conditions. The causes are often unpredictable, and the delays caused by these phenomena prevent precise definition of the time of arrival of buses at specific locations, thereby creating difficulties for users, but also for the operators. The purpose of this paper is to show development of a system through which localization and monitoring of the vehicles of public transportation would be done using wireless mesh networks. Due to these factors, the establishment of a stable system of scheduling becomes extremely difficult task, both in urban and intercity lines. Since most bus station still does not have an intelligent system for localization and monitoring of vehicle users are given a fixed timetable [1]. On the other hand, administrators themselves must manually enter data related to the departure and arrival of vehicles, which is time consuming and subject to error during entering. This primarily refers to the main bus station, where such information is recorded for each vehicle that arrives at the station during the day. The work of the entire system can be improved by developing intelligent system of the vehicle with automatic collection of information that would be of great benefit for consumers, but also for transport operators and administrators [2]. Based on these various analyzes could be made with the aim of improving the quality of services. The current location of the vehicle, its delay, the number of vacancies or the temperature in the vehicle is information that is very important for users who still, based on it, may decide to use transport services. On the other hand, service providers can perform detailed analysis and to monitor the movement of vehicles from day to day, and that based to the results plan the development of the system, changes in the timetable, and increase or decrease of the number of vehicles for specific lines depending on the number of passengers.

The following section presents the characteristics and features of the wireless technologiese in mesh networks with advances and disadvantages and methodology of use proposed project. The third section provides explanation of use communications protocols, commands and registers for proposed system for localization and monitoring publlic transport based on Zigbee. Communication algorithm on the transmitting and receiving device is shown in the fourth part. The results of the analyses and diskussion of developed application for monitoring and localization objects are shown in the fifth part. At the end of the paper there is a conclusion.

II. FEATURES OF WIRELESS TECHNOLOGIES IN THE MESH NETWORKS AND METHODOLOGY

For the implementation of intelligent systems that could perform localization and monitoring of public transport services a variety of technologies and standards can be used. In the past, the attempts of identification and tracking of vehicles using CCTV (Closed-Circuit Television) technology have been reported, which was based on the recognition of images. The performance of such systems was extremely poor (the approximate value of the accuracy was 20%). Today, different wireless standards and technologies are used, as separate or in combination with any other wireless or wired technology. Some of them are GPS (Global Positioning System), GSM (Global System for Mobile Communication) and GPRS (General Packet Radio Service), RFID (Radio Frequency Identifier), WSN (Wireless Sensor Networks) and Zigbee [3]. GPS is often used in combination with GSM or GPRS technology. The GPS system is based on the exchange of

information between the GPS receiver and satellite. To determine the position of an object it has to be visible to at least three satellites, which also can be an issue, for example when you have an underground station. Systems used for civil purposes are not completely accurate and often make greater or smaller discrepancies when determining the coordinates, depending on the quality of equipment. The problems also occur because of the existence of the multipath, when the signal is reflected from various obstacles before it reaches the receiver. For example, the city in this work has already implemented the network to track the movement of buses using GPS systems for satellite tracking. Each vehicle has a built-in sophisticated GPS / GPRS device that sends data on the movement of vehicles on the server. But, The GPS usage may get delay in locating the moving bus and system may mislead them to different bus stop. Table I shows comparison of different wireless technologies for use in public transport.

TABLE I. COMPARISON OF WIRELESS TECHNOLOGIES FOR USE IN PUBLIC TRANSPORT

Technology	Accuracy	Infrastructure development	Interference	Price
GPS+GSM	**Good**	**Already available**	**Small**	**Great**
RFID	**Good**	**Implementation required**	**Great**	**Small**
WSN	**Good**	**Implementation required**	**Small**	**Small**

RFID technology is often implemented as a part of a wireless mesh network, in order to avoid wire connection of RFID reader to the host application. In this case, Bluetooth or Zigbee wireless standard is the most commonly used. However, it can also be implemented in combination with GPS and GPRS standards. RFID works on several different frequencies on which depends the range or distance at which the reader can communicate with shortcut. If you use low and medium frequency (125 kHz, 13.56 MHz) range is very small, 5cm-30cm with a maximum power of radiation. On the other hand, at high frequencies it is possible to communicate at a distance up to 10 m, but in this case it results in the drastic changes in the price and size of the equipment (reader). One of the technologies that can be used for wireless communication of RFID nodes is Bluetooth, a standard that is used for implementation of personal local networks of low range. The biggest drawback of this standard is high energy consumption; therefore it is not suitable when working with devices that can be powered using the battery. Also, the range is small and reaches a value of 10 m. The aim of this paper is to show that for the design of intelligent monitoring system, in addition to the above mentioned, we use Zigbee standard. It enables lower bit rate that is sufficient for the mentioned functionality, while it is making significant savings in terms of energy consumption in contrast to the for instance Bluetooth. Also, it is proposed to use the Arduino platform that opens the door for the implementation of various functions for communication between stations and vehicles or coordinator and vehicles, which further expands the range of functions of the system [4].

Based on set goals, as part of this work a simple wireless mesh network was implemented which relies on the Zigbee standard [5], [7]. The possibility of establishing a mesh network is of great importance, because the actual implementation of the entire system with 129 stations and 60 buses would require covering a wide area (city of Tuzla), and the use of redundant links is desirable so that in case of failure on one of the nodes it does not result in congestion or even system failure. The network described in this paper consists of a small number of nodes: coordinator, router, and two end devices [6]. The router is the bus station, and the end devices are buses. Communication between end devices and the coordinator is achieved through a router, although generally direct communication is possible between them. However, due to the specific development of communication protocol this form of communication does not give any results, so in determining the location of nodes it is desirable to avoid direct contact between the bus and the coordinator. Additional nodes are possible to include in the network, such as distribution routers for data transfer, while providing expansion of network range [8]. All collected information is sent to a central computer where processing and storing is done. In this regard, a special web application is developed using a simple graphical interface, which provides the user an insight into the movement of vehicles.

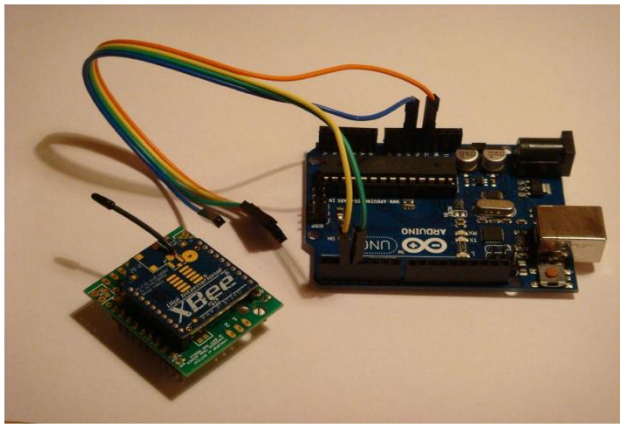


Fig. 1. Nod with Arduino and Xbee device

Node coordinator and bus consist of Arduino UNO platform and Xbee S2 module that communicate with one another using the UART protocol [9]. For the functionality of the station only an Xbee S2 module is needed. Coordinator communicates using two UART interfaces. One is hardware and is used for direct communication with the application on a computer. The second one is software, implemented using SoftwareSerial.h library and is used for communication between the Arduino and Xbee devices. The end device is made of the same components as the coordinator, but for communication it only uses software UART interface and exchanges information with the router. During the implementation and development of communication protocol, the hardware was mainly used for monitoring the device and printing errors. For powering coordinator constant and stable source is needed, so it is powered through the computer which is used for the processing of the data collected. Other nodes in the network can be powered using an external power source, such as batteries. The station is fixed, while buses are mobile and can move freely within the network. All Arduino devices are programmed using the Arduino IDE environment, where

the libraries for implementation a software serial port are used and a special library for communication of Arduino and Xbee devices. Setting of registries of Xbee modules, which work in API mode, is performed using X-CTU software.

III. Communication Protocol, Commands and Registries

Communication protocol, which was developed for this system, is based on the use of AT commands that give us insight into the value of appropriate registries of the Xbee devices. During the work the following controls were used: Serial Low (SL), Node Discovery Timeout (NT), Node Identifier (NI) and Node Discovery (ND). Within Xbee devices there are a large number of registries with fixed or variable values, but we will take a look at those that are essential to our system.

Each node in the network is uniquely defined by a 64-bit MAC address which is stored in the registries SH (Serial High) and SL (Serial Low) of the Xbee device [10]. What is characteristic for Xbee devices is that the value of SH registry is the same for all and is 0x0013A200, so it is actually SL value that uniquely identifies each device. These values are assigned during production and cannot be changed. Registries DH (Destination High) and DL (Destination Low) are used for storage of the 64-bit address of the destination node. Since all devices send data to the coordinator, these registries have the same value for all devices, and they are DH = 0x0, DL = 0x0. These values can be changed, so that by sending the appropriate AT commands we can change the final destination of the package. To send a broadcast message it is necessary to set the following value DH = 0x0, DL = 0xFFFF. ID (Extended PAN ID) is a registry that stores the value of PAN ID. All devices that have the same PAN ID can communicate with each other. If the value of this registry is 0, the coordinator will during the establishment of the network and initialization of variables choose an arbitrary, available value for the given registry. If routers and end devices have the pre-set value to 0 they can connect to any available PAN network. BH (Broadcast Hops) is a registry having a functionality of great importance to the implementation of our system. Namely, this registry enables us to limit the number of hops for transmission of broadcast messages. Given that the end devices often send ND command, which is the broadcast message, it is possible that the congestion in the network appears due to overloading. However, setting this registry value to 1 we restrict ND command to be forwarded only to the nearest node, thereby reducing network load. Since the devices do not need to send broadcast messages to all devices, set limits are not ruining the overall performance of the system. Registry value NJ (Node Join Time) determines the time within which the coordinator or router allows connection of new nodes. This value is set to the maximum value of 0xFF, which is the default value and allows the connection of new nodes at any time. This registry is important in situations where the end nodes move in the network and change their parents. PL (Power Level) and PM (Power Mode) are registries by which we determine the level of emitted power of devices. PL registry can have a value of: 0,1,2,3 and 4 which correspond to the value - 8dBm, - 4dBm, - 2dBm, 0 dBm and + 2dBm, respectively. Using this registry we limit the power of broadcasting of the end devices. Namely, let us look the scenario in which the vehicle is moving down the road, and the power level of radiation is set to default value. With a power of radiation of 2dBm the range of device can reach a value of 100m. This would mean that our bus could communicate with the stations that are not part of its route, and are, for example, on the part of the road where vehicles move in the opposite direction. In this case, information about the movement would not have a logical sense, and could create confusion in the system itself. Therefore, the value of this registry, in case of the end devices, is set to the lowest possible, so as to reduce the range in which the bus emits the ND command. PM is a registry that allows the boost mode. If enabled, boost mode increases the transmitting power for 2dB and improves sensitivity of reception for 1dB.

For all the devices values of the registry NI are predefined. This registry is used for the storage of the string that could be used to identify the device. For the value of the string we can use characters from the ASCII table, provided that the string cannot start with an empty space and it is not allowed to use the comma. For the purposes of communication protocol, devices on the network are assigned names based on their functionality. Value of the NI registry starts with one of four possible letters: C for coordinator, R for distribution router, S for the station and B for the bus. For all devices it is characteristic that an initial letter is followed by three digits that uniquely identify the device (e.g. B305, B241, S902, R301, and C141). NT (Node Discovery timeout) is registry closely related to the ND command. Specifically, this value is the timeout value for ND command or the period of time within which the device will wait for a response. The value of NT registry is included in the ND command which is sent to other devices. In addition to these, it is important to mention the MY registry which stores the network address of the device, and CH (Operating Channel) in which contains the value of the number of the channel in which the device operates. There are registries for setting up encryption and encryption key (EE, EC, NK, KY) which are used in case we want to secure the information that is sent within the network. However, for this paper, which has the presentation character, the mentioned registries are not used.

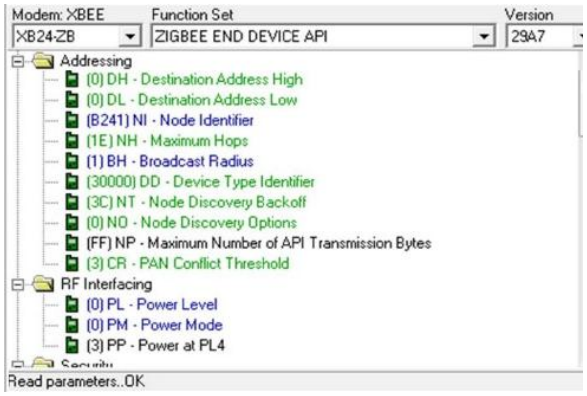


Fig. 2. X-CTU, setting of the parameters of the Xbee device

IV. Communication Algorithm on the Transmitting and Receiving Device

Communication between the devices takes place in the following manner. Once connected to a power source, the end device (bus) first performs initialization of variables where, among other things, a string of 20 bytes is initialized and used to send information to the coordinator. The first four bytes have a value of 5, and are used for checking on the receiving side, while others are dynamically filled. Bus first sends SL and NI command and writes responses in the specified string to positions 4-11. These functions are called within the setup() function, so they are executed only once during the initialization of the system. Bus then sends the NT command, and stores the resulting value for the timeout for further use. After that, the device enters the loop function in which, at appropriate intervals, it sends ND command.

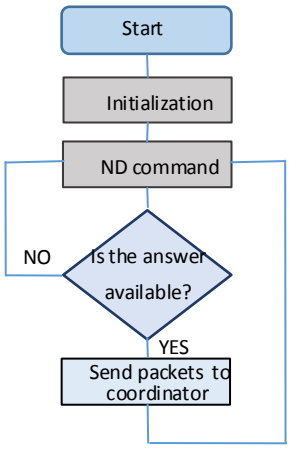


Fig. 3. Algorithm of the end device - bus

The bus is a mobile node, and moves within the network traveling from station to station. In his range in this case we can find another buses, stations and routers. We will assume that the coordinator is at the location where it is not directly accessible to mobile buses.

Each router or stations that are found on the route of the bus are potential parents of a given node. When some of these devices are in range of the bus, the bus will get a response to the transmitted ND command. The response contains a range of information such as network and MAC address, value of the NI registry, network address of parents, as well as some others. The total length of the response depends on the value of NI registry. However, since these values are pre-defined our response contains 23 bytes and characters of the NI string and SL are in positions 6-14. The idea is that the value of said device is written to the remainder of the string to be sent to the coordinator. However, sending of a string is possible only when the byte in which the value of the first letter written is equal to 83. It is the hexadecimal value for the letter 'S'. The simple comparison is done using if() loop. If this byte contains another value (82 for R, 67 for C) the package will not be sent, and the device returns to the initial position from which it again sends the ND command. From this we see that the behavior of node is defined even if there is direct communication with the coordinator. There are many reasons why the stations are chosen to carry the router functionality [11], [12], and buses carry the functionality of the end devices. On the ND command of the end device other end device never answers, so when we would have two buses close to each other they would not mutually communicate. On the other hand, the station must be able to communicate with each bus that passes.

End devices can be in sleep mode, enabling energy savings and extension of battery life for instance. The station, on the other hand, cannot be in sleep mode because it performs the functionality as a router. The final destination of all devices is the coordinator. Its role is that after receiving the package forwards the received information to the base. At the beginning of each package the coordinator adds two characters '?' which are used to determine the beginning of the string that contains the necessary information. The hexadecimal value for the mentioned character is '3F'.

Forwarding is done using the script that is written in the Python programming language. Due to the potentially large number of devices that can send data to the coordinator and that a collision is possible, the script will write the appropriate data into the database only when the hexadecimal value of the first two bytes within a string are equal to '3F'.

Otherwise, the program flow leaves the loop in which the registration is done, and we come back to the state of listening to the serial port and waiting for reception of information from the coordinator.

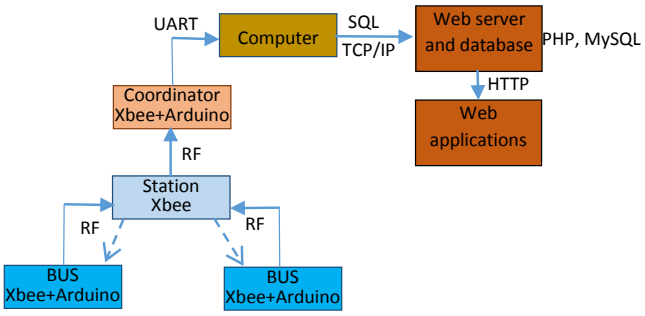


Fig. 4. Network elements

The database is implemented using MySQL (My Structured Query Language), an open-source management system for relational databases. Within the base there is a series of tables containing data related to buses, stations and lines, which are used within a web application that is designed for end users of the system.

To establish a connection to the base, and allow the storage to full extent it is necessary to enter the parameters for the connection. After that it is necessary to enter the number of the serial port on which it is expected to receive data from the coordinator. Web application is developed using PHP, a scripting programming language that is one of the most commonly used for programming dynamic web pages. Information from the database is retrieved by sending appropriate requests. For certain functionalities of the web page Java scripts are used. The frame of the web page is written using HTML (HyperText Markup Language), and for the design of elements CSS (Cascading Style Sheets) is used. Below we will describe in detail the functionality of this application.

V. Results and Testing Developed Application for Monitoring and Localization

All information about the vehicle's current location in the network is stored in a database. Data from the database is directly accessible to the administrator, but not to end users. Therefore a web application is developed to be an interface through which the user can obtain the desired information. When a user opens the home page he or she can see a number of basic elements. The first is certainly a scheme that shows all the city's bus lines. On the scheme all stations are indicated, and bus lines are painted different colors for clarity.

The user has the option to use Google Maps. The original coordinates and zoom are defined in a way that it shows the entire city. The user can, if necessary, move the map, or zoom in to make it easier to display the desired location. Under this scheme and map there are two elements. The right one is to display a list of city lines, with their initial and final destination. The left element allows the user to select the desired station or line.

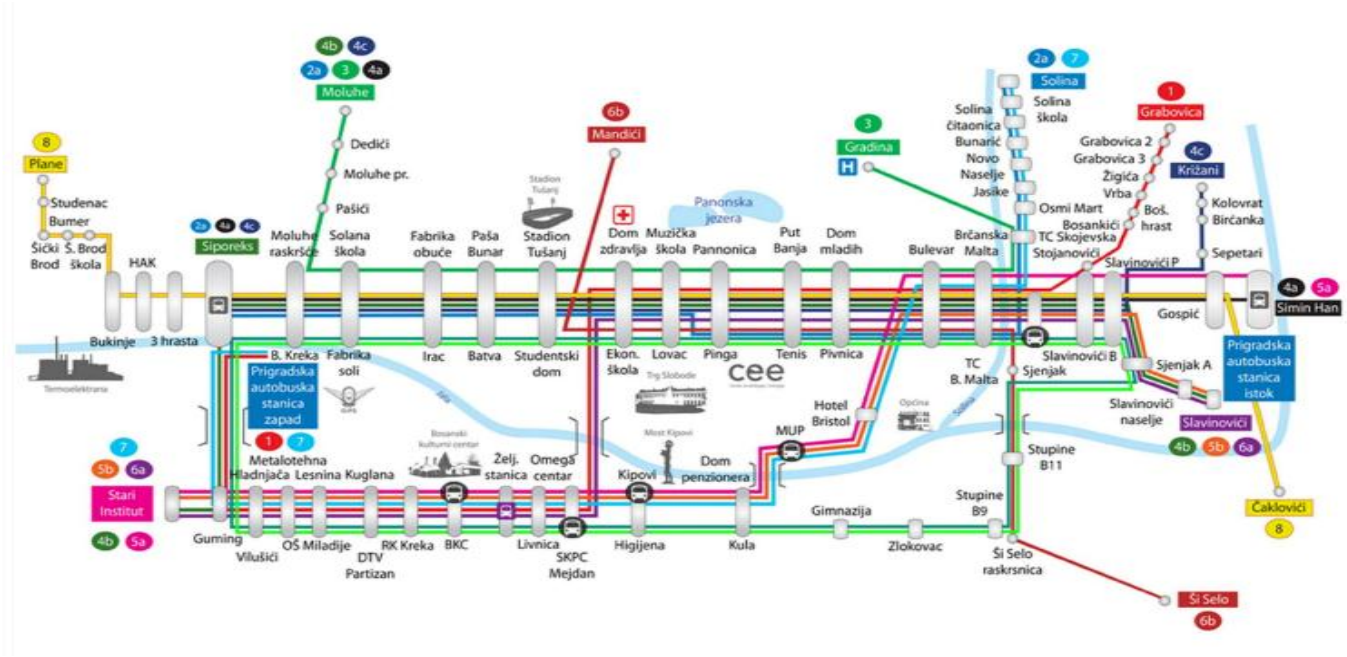


Fig. 5. Scheme of stations and city's bus lines and locations of the vehicles on the map

Fig. 6. The beginning of the search and the offered options in case of search based on bus lines

Fig. 7. Search results for the line '5a' and search results for the station 'Tenis'

If the user selects the option 'Line', in the next field a list with the options opens. Let us assume that the user has selected the line '5a. To the right is a table of the results where the user can read where the vehicles on selected line were registered, the time and date of registration, which is the next station, and a rough estimate of the arrival of the vehicle at the next station.

On the left side the user can see which line is selected and current time, and pressing one of the buttons offered user can use their functionalities. 'Location of the vehicle' in the scheme small icons will display to indicate the station where the buses of the line '5a' were last registered. Icons blink, so that the user can relatively easy see them.

'Main lines' in the menu gives the user access to a list of the main bus lines at any time. If the user on the home page decides to search based on the name of the station, then the table with results will show last location, time and date of registration on last location for all the buses which move towards the given station and which are within the range of three stations from the given station. The next figure shows the results in case where user has chosen the station 'Tenis.' What is important to mention is that it is possible that a result table is empty, depending on whether the previous stations previously registered vehicles. Regardless of whether the search is conducted based on the number of the line or the station name, the user always has available option to see the results on Google Maps. In this case, the map shows markers that mark the approximate location of the station.

VI. CONCLUSION

The paper describes the architecture of such a network on the example of the part of the city, its advantages and disadvantages that affect the performance of the system. It was found that part of the needs of such a system can be met by using Zigbee wireless standard that is extremely suitable for large networks where the transfer of packages of small length is done. Consequently, the Xbee RF modules were used to implement Zigbee protocol, together with the Arduino microcontroller platform. Zigbee has proven to be a good solution to the problem because it is relatively easy to implement, and communication can be achieved in areas where there is no 3G, 4G or other wireless technologies of long range. This system will be an efficient and low cost one due to the use of Zigbee. Thus this paper gives a complete passenger friendly system. In the case of development of network for the entire city, it would be necessary to conduct additional studies that would determine which locations are suitable for setting up nodes and what is the total number of required nodes. Such a system would easily be able to integrate with mesh networks, and in addition to the location data of vehicles it could also collect other information such as: number of passengers, external and internal temperature, the level of CO_2, the level of brightness, as well as any other information that can be collected using a variety of analog and digital sensors.

BIBLIOGRAPHY

[1] Alireza Faed, ”An Intelligent Customer Complaint Management System with Application to the Transport and Logistics Industry”, Springer International Publishing, 2013.

[2] Yunhao Liu, Zheng Yang, “Location, Localization, and Localizability: Location-awareness Technology for Wireless Networks”, Springer-Verlag New York, 2011.

[3] Misra, S., Misra, S. C., & Woungang, I., “Guide to wireless mesh networks”, London: Springer-Verlag, 2009.

[4] Wheat, D., “Arduino Internals. Technology in action”, Apress, 2011.

[5] Aswin Sayeeraman, P.S.Ramesh,. “Zigbee and GSM Based Secure Vehicle Parking Management and Reservation System”, Journal of Theoretical and Applied Information Technology, 31st March 2012., Vol. 37 no.2, pp 199 – 203.

[6] Xiaoya Hu; Wei Xiong; Wei Li; Li Ke "Application scenarios of wireless sensor networks for urban transportation: A survey", Control Conference (CCC), 2015 34th Chinese, pp. 7688 – 7691.

[7] “Wireless Mesh Network Concepts and Best Practices Guide”, Schneider Electric, Revision C, 2010.

[8] Farahani, S., “Zigbee Wireless Networks and Transceivers”, Elsevier Ltd., 2009.

[9] A.Mellis, D., Banzi, M., Cuartielles, D., & Igoe, T.” Arduino: An Open Electronics Prototyping Platform”, CHI, 2007, San Jose, USA .

[10] M.Ortiz, A., Royo, F., Olivares, T., & Orozco-Barbosa, L. ,“Intelligent Route Discovery for Zigbee Mesh Networks”, The 12th IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks, Tuscany, Italy, june 20-23, 2011 .

[11] He Feng , Li Lulu Yin , Heng Huang Xia, “Bus Monitoring System Based on Zigbee and GPRS”, 5-6 March 2012, International Conference on Computer Distributed Control and Intelligent Environmental Monitoring, pp. 178 – 181.

[12] G.Lavanya, Preethy. W, Shameem.A, Sushmitha.R, “Passenger Bus Alert System for Easy Navigation of Blind”, International Conference on Circuits, Power and Computing Technologies, 20-21 March 2013, 798 – 802.

An approach of inertia compensation based on electromagnetic induction in brake test

Xiaowen Li
College of Computer Science and Technology
Chongqing University of
Posts and Telecommunications
Chongqing, China

Han Que
College of Computer Science and Technology
Chongqing University of
Posts and Telecommunications
Chongqing, China

***Abstract*—This paper briefly introduced the operational principle of the brake test bench, and points out the shortcomings when controlling the current of brake test, which means the reference measuring data is instantaneous. Aimed at this deficiency, a current control model based on electromagnetic induction and DC voltage is proposed. On the principle of electromagnetic induction, continuous data and automatic processes are realized. It significantly minimized errors owing to instantaneous data, and maximized the accuracy of the brake test.**

***Keywords*—Brake test; Electromagnetic induction; DC transformer**

I. Introduction

Vehicle brake design is one of the most important processes in the vehicle design. In order to detect the comprehensive performance of brake, thousands of brake tests are required. The actual test is generally divided into the road test and simulation test. The simulation test basically based on vehicle brake test bench. However, the road test is impossible to operate during vehicle design stage. Thus, brake test simulation on brake test bench is the best method in this situation. The principle of this method is simulating the road test on the brake test bench as much as possible.

In the simulated road test experiments of brake test for brake performance, due to the mechanical inertia flywheel group could not concisely achieve the rotational inertia which test system required, typically the industry introduce motor into the test bench. In order to meet the principles of simulation tests, the current of the motor could be controlled specifically when it participating in the experiments to compensate the energy the mechanical inertia required. However, due to the complexity of the brake performance, the precise relationship between the motor driving current and the time is difficult to obtain. The normal method is discretization. The entire braking time is discretized into quite a few tiny time periods. Then according to the instantaneous speed and instantaneous torque which observed in previous period, devising the driving current value of the current time. This process successively operated until the completion of brake test. Actually, the driving current value devised by this discretion method possess a certain error comparing to theory value. The error would cause unnecessary trouble in the brake test. Through the establishment of the current control and the DC transformer model based on electromagnetic induction and electromagnetic induction method, the error is able to be eliminated. Consequently, improvement of the brake test bench experiments is realized.

II. Components of test bench system

1) Basic components of brake test bench: Brake test bench generally consists of spindle with flywheel group, motor which driving the spindle rotation, basement, assist devices applying on the brake, measurement system and control system, etc. Shown in Figure 1.

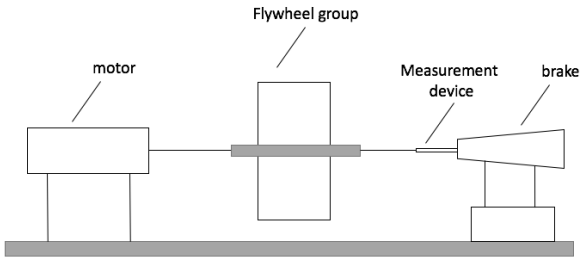


Fig. 1: Basic components of brake test bench

As shown, when the brake test start operating, the motor drives the flywheel group rotating to reach the set rotational speed which corresponding to the speed of vehicle (in the simulation, assuming the angular velocity of the spindle is always constant with the angular velocity of the wheels). Once it reached the set speed, then disconnecting the power supply. Meanwhile, apply braking for the motor. When the termination condition is satisfied, one brake test is completed. Comparing to brake test bench, the vehicle wheels in road test would absorb load when applying brake. Thus, the energy which the 'load' posses when it moving with the vehicle (ignoring rotation energy of the wheel itself) should be equally converted to the energy of flywheel group, spindle and other devices in brake test bench when they are rotating. This energy which corresponding to the rotational inertia (short for inertia below) is called equivalent rotational inertia. In order to simulate the road test with mechanical inertia accurately, we introduce the motor into the brake simulation test and controlling the current of motor with regular rules to compensate the energy which mechanical inertia require. Consequently, the principles of simulation test are satisfied.

Basic assumptions:

1. Assuming that flywheel group is strictly rigid.
2. Assuming that the braking torque of road test provided entirely by the brake.
3. Assuming that the brake torque from motor and brake is applying on the spindle completely.
4. Assuming that brake torque of spindle is entirely provided by the brake
5. Assuming that the angular velocity of spindle in brake test bench is constant with the angular velocity of the wheel in road test.
6. Assuming that the measured data is accurate and reliable.
7. Assuming that when the measurement is performed, the time interval is stable after discretization.

In the experimental test, the entire simulation processes are shown as below:

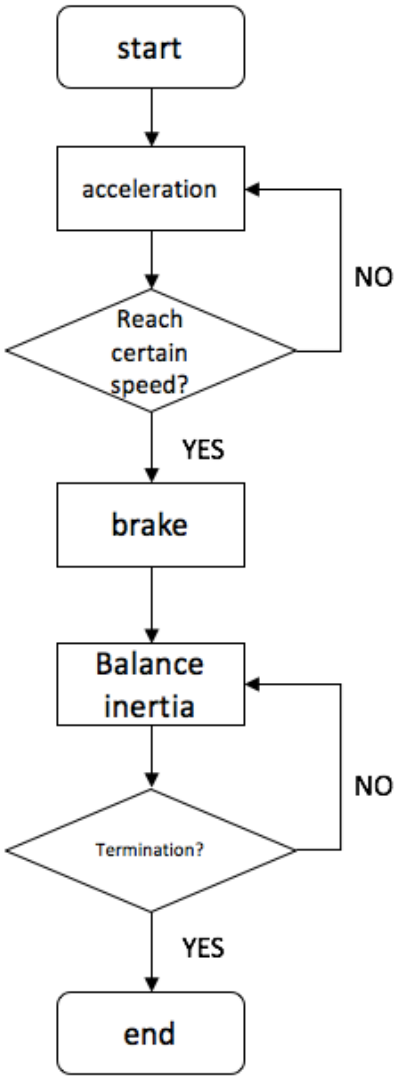


Fig. 2: Simulation process

In a typical test bench experiment, the motor driving current control is determined by measuring the discrete instantaneous speed and instantaneous torque. Therefore, the errors are larger in actual situation. According to the thought of continuous discrete control, we introduce the current control and the DC transformer model based on electromagnetic induction.

2) Establish the current control model based on electromagnetic induction and DC transformer: On the basis of brake test bench system, a miniature DC generator is connected and fixed on the spindle. Accordingly, the generator coil inside the generator have the same rotational speed with the spindle. Thus, when the spindle is rotating, the DC generator will produce induced electromotive force by electromagnetic induction.
We introduce induced electromotive force into the circuit 1. By changing the current I_1, the inductance L in circuit 1 generate induced electromotive force. The circuit 2 generate induced electromotive force through transformer inductance L. We add a DC transformer in circuit 2 and connect the power supply to provide additional energy output. By adjusting the magnification of the voltage, making the voltage applied to both ends of the dynamic system in the brake test bench. Meanwhile, the energy of this voltage provided by the electric motor is equal to the energy consumed by heat plus the energy required for spindle rotation. Thus, we can make the discrete data which in original model continuous and reduce random errors generated by discretization.
The main structure and a circuit diagram of the model shown in Figure 3:

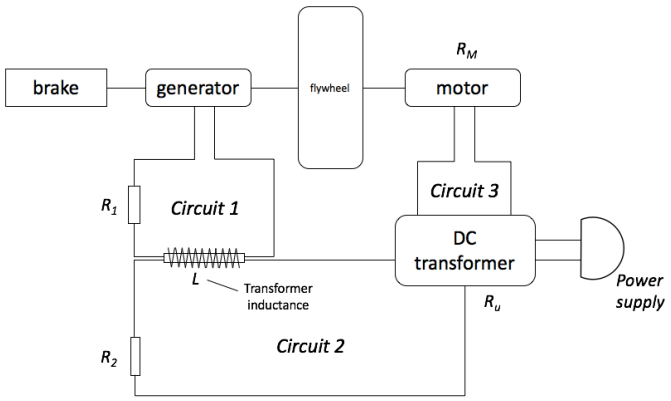


Fig. 3: Circuit diagram of the model

3) DC generator: The following context described the DC generator.
Figure 4 is a simple DC generator model. The *N* and *S* are a pair of fixed poles. They are either permanent magnet or electromagnet. There is a rotatable iron cylinder between the magnetic poles which known as the armature core. On the surface of the core fixed the armature coil '*abcd*' which consist of insulating conductors. On both ends of the coil connected to 2 mutually insulated arc-shaped copper respectively. The arc-shaped copper is called commutator segments. The combination of them is called a commutator. The commutator brushes A and B are placed on the commutator stationary and sliding contact with the commutator. The coil '*abcd*' connected to external circuit through the commutator and brushes. The armature core, armature coils and commutator together integrally

called an armature. The armature rotating with prime mover and transform the mechanical energy into electrical energy and supplied to the electric load which connected to the brushes.

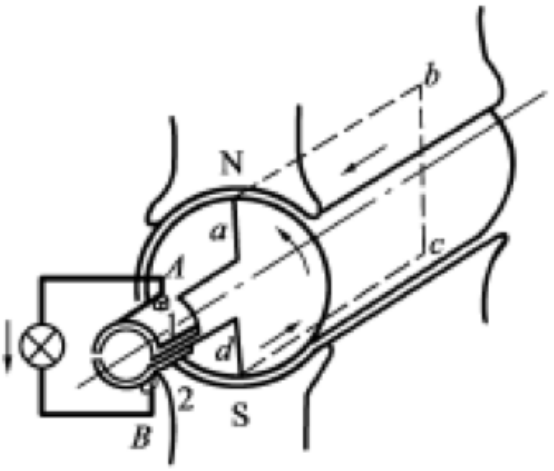


Fig. 4: DC generator model

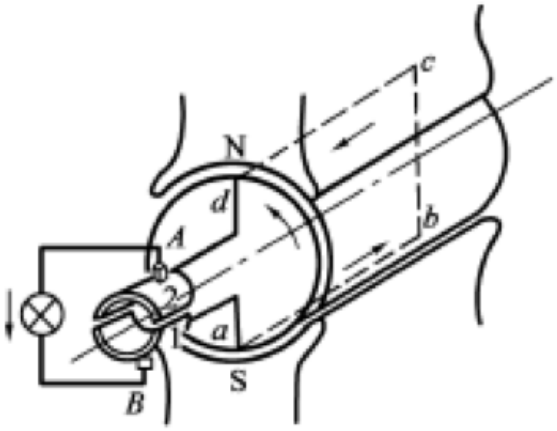


Fig. 5: DC generator model

In figure 4, when the armature counterclockwise rotating, the coil edges *ab* and *cd* cut magnetic lines and generate induced electromotive force. The direction of induced electromotive force shown in figure 4 is from *dcba*. The brush *A* is positive and brush *B* is negative. The direction of current flow in the external circuit is start from positive *A* to the negative *B* through circuit load which could be judged according to the right-handed rule.

When the armature rotated with 180 degrees, as shown in figure 5, the induced electromotive force direction of the coil changed to *abcd*. The brush *A* originally contact with the commutator 1 consequently changed to contact with the commutator 2. The brush *B* originally contact with the commutator 2 consequently changed to contact with the commutator 1. Thus, the brush *A* still positive and brush *B* is still negative.

The analysis shown above turns out that when a certain edge of the coil rotating from the range of magnetic poles to the range of the adjacent opposite magnetic poles, the direction of induced electromotive force changed for one time. However, in terms of the external circuit, the direction of the induced electromotive force between the brushes is constant and the value of induced electromotive force ranges from zero to the maximum.

Although we can resolve induced electromotive force and current in a constant direction through this kind of single coil DC generator, but its value is pulsating. This is the basic principle of DC generator.

Actually, the value of induced electromotive force generated by the DC generator is not only pulsating but volatile. Thus, the electromotive force is not acceptable with actual requirements. To reduce the fluctuation of the electromotive force, we can appropriately increase the number of coils and the armature segments. As an example, winding two coils on the armature, namely there is one coil edge every 90 degrees. Thus, half-wave induced electromotive force is changed to 90 degrees and the pulsating has been significantly reduced. The coil number in practical applied DC generator is generally multiple and the number of magnetic pole pairs is also more than one pair. Thus, the fluctuation of induced electromotive force is rather slight (experiments and analysis show that when the number of conductors in each magnetic pole is more than 8, the fluctuation of induced electromotive force is less than 1%). Thus, we considered it as constant DC electromotive force.

4) DC adapter: Firstly, transforming direct current boost voltage of direct current into alternating through electronic components and then change the voltage through the transformer. This process is used for inverter voltage. The devices used for converting DC power boosting called an inverter.

Transformer works based on electromagnetic induction principle. Transformer owns two sets of coils, primary coils and secondary coils. The primary coils are outside of the secondary coils. When the alternating current flow through the primary coils, the transformer cores generate alternating magnetic field and the secondary coils generate induced electromotive force. The ratio of transformer coil turns is equal to the voltage ratio. As an example: the primary coil is 500 turns and the secondary coil is 250 turns. If the voltage of primary coils applied on 220V (alternating current) then the voltage of secondary coils is 110V. Transformer can either boost voltage or step-down voltage. If the turns of primary coils are less than the turns of secondary coils, namely it is a boost transformer which promoting to a high voltage.

Generally, the DC voltage transformer is used for the conversion of voltage and obtain energy from a certain power supply. Then use oscillators turning the current to alternating current and use the transformer to boost or drop voltage. Finally, reduced the current to the direct current through a rectifier circuit. The feedback loop control circuit is necessary if high criterial required in obtaining relatively stable output voltage.

III. Model solution

First, we establish a mathematical model which the motor driving current depends on observables. The model is based on the rigid physics theory and other physics theories. According to the force analysis of any rigid objects and relevant rigid physics conclusions, the torque of rigid objects is proportional to the product of its rotational inertia and angular acceleration, namely:

$$M \propto J\beta$$

When the torque, rotational inertia and angular acceleration are all SI units, the proportional coefficient is 1. We can resolve the relationship of rigid torque, rotational inertia and angular acceleration, namely:

$$M = J\beta$$

Based on the conditions we assumed, the angular velocity of the spindle in the brake bench test is completely constant with the angular velocity of the wheel in road test. And the equivalent rotational inertia J_0 and the mechanical inertia J_m are both constant value.
We assume that the wheel speed is uniformly changed when the car start braking in the road test. Namely, the wheel angular velocity decrease linearly and angular acceleration remain constant. Thus:

$$\beta = \frac{\Delta\omega}{\Delta t}$$

Due to the torque is provided by the brake completely in road test. The torque of brake denoted as M_b is a constant value according to the relationship of torque, rotational inertia and angular acceleration, which is:

$$M_b = J_r\beta$$

The time curve of torque, rotational inertia and angular acceleration are shown as below:

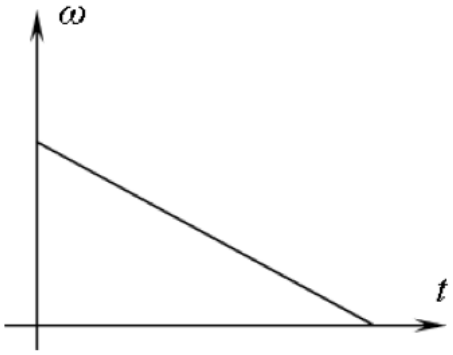


Fig. 6: Time curve of angular velocity

In the simulation test, the flywheel angular velocity decreases linearly and angular acceleration remains constant. It is the same with the assumptions in the road test. Due to the mechanical inertia is less than the equivalent rotational inertia, the angular acceleration of the flywheel in simulation test is greater than the angular acceleration of wheel in road test. Therefore, the driving current is provided to compensate the torque gap caused by the insufficient rotational inertia. Thus:

$$M_c = M_b - M_m = (J_r - J_m) \cdot \beta = (J_r - J_m) \cdot \frac{\Delta\omega}{\Delta t}$$

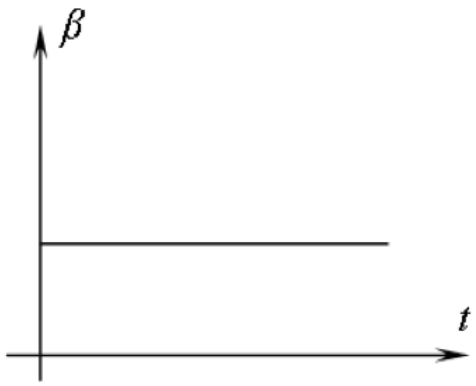


Fig. 7: Time curve of angular acceleration

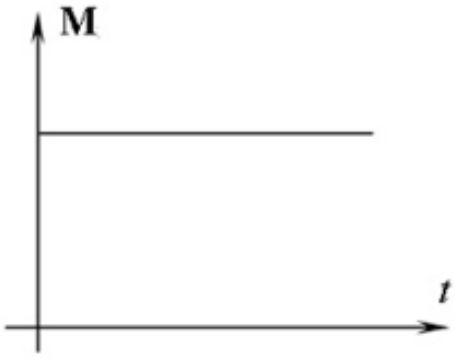


Fig. 8: Time curve of torque

According to the motor parameters, the motor driving current is proportional to its torque and we denoted the proportional coefficient as θ:

$$I = \theta \cdot M_m$$

Therefore, according to the conclusions above, we established the motor drive current mathematical model depending on the observables, namely:

$$I = \theta \cdot (J_r - J_m) \cdot \frac{\Delta\omega}{\Delta t}$$

In Figure 3, suppose that the equivalent magnetic induction density of the coil in the generator is *B*. The equivalent number of coil turns is *n*. The magnetic flux area of the coil is *S*. The rotational angular velocity of coil is ω and consistent with the spindle rotational speed. According to Faraday's law of electromagnetic induction, induced electromotive force generated by the DC generator is:

$$E_1 = 2nBS\frac{d\varphi}{dt} = 2nBS\omega$$

Introducing the current generated by DC generator into the circuit 1 when the coil is rotating, in which the self-inductance coefficient of the inductance is L and the resistance of the

closed circuit is R_1. According to Ohm law, the induced current in circuit 1 is:

$$I_1 = \frac{2nBS\omega}{R_1}$$

Then transfer the changing current from circuit 1 into the circuit 2 through the transformer inductance. Namely the circuit 2 generate the induced electromotive force by the changing current from circuit 1. In this situation, the transformer inductance efficiency is η. The inductance inductance coefficient in circuit 2 is L. The resistance in closed loop circuit 2 is (R_2 + R_u). Meanwhile, the induced electromotive force in circuit 2 equal to the product of current change rate and self inductance in circuit 1. Therefore, the induced electromotive force in circuit 2 is:

$$E_2 = \eta L \cdot \frac{dI_1}{dt} = \frac{2\eta nLBS}{R_1} \cdot \frac{d\omega}{dt}$$

The relationship between the spindle angular acceleration and angular velocity is:

$$d\omega = \beta \cdot dt$$

Therefore, the induced electromotive force in circuit 2 can be written as:

$$E_2 = \frac{2\eta nLBS}{R_1} \cdot \beta$$

As shown in figure 3, Suppose that the resistance connected to the DC transformer in circuit 2 is R_u. According to the principle of voltage dividing, the voltage transporting to the DC voltage transformer is:

$$E_u = E_2 \cdot \frac{R_u}{R_2 + R_u} = \frac{2\eta nLBSR_u}{R_1(R_2 + R_u)} \cdot \beta$$

In circuit 3, we denoted the current which the motor required for providing torque as I_3, according to the the relationship among the current, inertia compensation and the rotational angular velocity, the value of I_3 can be obtained:

$$I_3 = k \cdot (J_r - J_m) \cdot \frac{d\omega}{dt} = k \cdot J_c \cdot \beta$$

Let motor internal resistance be R_M. The coefficient proportion of energy for self-consumption to the total energy in the entire circuit 3 is ξ. Then we can solve the electromotive force which circuit 3 required, which denoted as E_3. These are the equations:

$$\int E_3 I_3 dt = \frac{1}{\xi} \int I_3^2 R_M dt$$

then:

$$E_3 = \frac{k \cdot J_c \cdot \beta \cdot R_M}{\xi}$$

Simultaneous equations:

$$E_u = \frac{2\eta nLBSR_u}{R_1(R_2 + R_u)} \cdot \beta$$

DC transformer voltage ratio is obtained as follow:

$$\lambda = \frac{E_3}{E_u} = \frac{k \cdot J_c \cdot R_M \cdot R_1(R_2 + R_u)}{2\eta\xi nLBSR_u}$$

When the test bench experiments come to 'Start brake' step shown in the flowchart and disconnect the power switch which spin up the flywheel, then connect the motor with the entire circuit which designed as figure 3(including power supply energy for DC transformer). The rest processes of brake test would be completed automatically.

IV. Error Analysis

According to the computer simulation model above, we can start the error analysis based on the data and obtain the time curve of energy relative error, shown in figure 9:

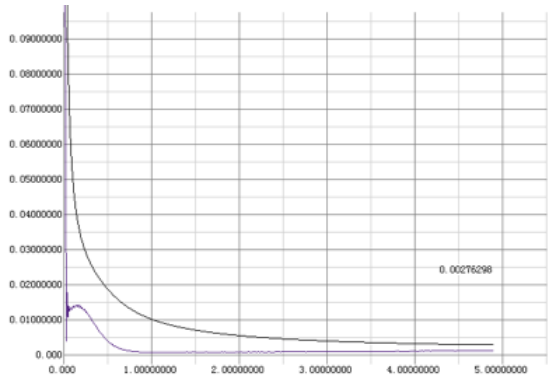


Fig. 9: Error analysis

Based on energy relative error curve, the error can be controlled within 1.5% in the beginning of the experiment. The error in this part is mainly systematic error and error due to weakly mutual inductance generated between the circuit 1 and the circuit 2. After 0.9s, the error of the entire system is only 0.1%. The average energy relative error of the entire process is 0.276%. Apparently, the error with this model is rather slight.

V. Conclusion

The model is strongly supported by software. Especially the support from physics, mathematics and relevant theory, and computer control. Due to the motor drive current is generated automatically based on electromagnetic through physical theory and the motion state of the entire test bench system. Therefore, the result is more precise than the PID intelligent control technology in industry which has been widely used

currently. This is the advantage of this model.
However, the model needs to be supported by precise hardware, as well as the various hardware parameters. For instance, the various parameters of the motor must be very precise. In summary, the error of the model mainly from system hardware errors, which is insufficient for this model.

REFERENCES

[1] H. Kopka and P. W. Daly, *A Guide to LaTeX*, 3rd ed. Harlow, England: Addison-Wesley, 1999.

[2] Gezhi Zhu and Xiaolan Zhang, *Motor and Drag*. Chongqing University Press,Sep 1999.

[3] Renguang Wang, Zhaodu Liu, Yuefeng Ma, Zhiquan QI, *Improved design of brake test bench*. Agricultural Machinery, Vol.37, No.6, May 2006.

[4] Jianjun Chen, *Brake test bench mechanical inertia electrical analog control method*. Shaoxing, Zhejiang Province, Shaoxing University Institute, April 2007

[5] Shubao Wei, Changsheng Xu, *Inertial load brake test bench design and virtual prototype simulation analysis*. Wuhan University of Technology, Vol.30, No.6, pp. 988, Dec 2006.

[6] Feng Xie, Jiwang Fang, Juguang Lin, *Single-ended inertia brake performance test bench development*. CLC: TH16; U467. Technology and Equipment, Nov 2008.

[7] Bo Liang, Yuren Li, *Fuzzy self-tuning PID application of inertia simulation in the brake test bench*. Electronic Measurement Technology, Vol. 31, No.10, October 2008

[8] Haishu Du, Zhi Yang, Rongsheng Qiu, *Neural intelligent PID control algorithm*. CLC: TP273.2, Document code: A, Jan 1999.

[9] Shiqin You, *Flywheel rotational inertia calculation*. Huainan Technical College, Volume 2 (Total 5), Apr 2002

[10] Qiang Fei, Wuyun Zhao, Fei Dai, *Design of vehicle brake speed control system based on LabVIEW [J]*. agricultural equipment and vehicle engineering, 2013, 51 (11): 6-10.

[11] Qiang Song, Huihui Zhang, Yiming Feng, *Research on grey prediction control of brake test bench[J]*. [1] Earth, 2014 (5).

[12] Tonghua Niu, Hu Ni, *Development of Automotive Hydraulic brake test bench [J]*. Quality of goods and construction and development, 2015 (1)

A Frequency Based Hierarchical Fast Search Block Matching Algorithm for Fast Video Communication

Nijad Al-Najdawi
Al-Balqa Applied University
Al Salt, Jordan

Sara Tedmori
Princess Sumaya University for Technology
Amman, Jordan

Omar A. Alzubi
Al-Balqa Applied University
Al Salt, Jordan

Osama Dorgham
Al-Balqa Applied University
Al Salt, Jordan

Jafar A. Alzubi
Al-Balqa Applied University
Al Salt, Jordan

***Abstract*—Numerous fast-search block motion estimation algorithms have been developed to circumvent the high computational cost required by the full-search algorithm. These techniques however often converge to a local minimum, which makes them subject to noise and matching errors. Hence, many spatial domain block matching algorithms have been developed in literature. These algorithms exploit the high correlation that exists between pixels inside each frame block. However, with the block transformed frequencies, block matching can be used to test the similarities between a subset of selected frequencies that correctly identify each block uniquely; therefore fewer comparisons are performed resulting in a considerable reduction in complexity. In this work, a two-level hierarchical fast search motion estimation algorithm is proposed in the frequency domain. This algorithm incorporates a novel search pattern at the top level of the hierarchy. The proposed hierarchical method for motion estimation not only produces consistent motion vectors within each large object, but also accurately estimates the motion of small objects with a substantial reduction in complexity when compared to other benchmark algorithms.**

Keywords*—Video coding; Frequency domain; Motion estimation; Hierarchical search; Block matching; Communication.

I. Introduction

A moving video frame (image) is captured by taking a rectangular snapshot of the natural signal at periodic time intervals. Playing back the series of frames produces the appearance of motion. A higher temporal sampling rate (frame rate) gives a smoother playback, but requires more samples to be captured and stored. Most video coding methods utilize both temporal and spatial redundancy to compress video data [1]. In the temporal domain, there is usually a high correlation between frames captured at around the same time. Temporally adjacent frames are often highly correlated, especially if the temporal sampling rate is high. In the spatial domain, there is usually a high correlation between pixels (samples) that are close to each other. Thus, the values of neighbouring samples are often very similar [3]. In video compression, intra frame and inter frame coding are applied in order to reduce the number of bits needed to represent a video. In intra-frame coding, each frame is coded without any reference to other frames. This process involves transforming the block into the frequency domain, where the resulting coefficients are quantized and encoded. A better compression may be achieved with inter-frame coding which exploits the temporal redundancy. In inter-frame coding, motion estimation and compensation (two vital processes within video coding) have become powerful techniques to eliminate the temporal redundancy due to high correlation between consecutive frames. Successive video frames may contain the same objects. Motion estimation is the process that describes the transformation from one image to another through examining the movement of objects in an image sequence to try to obtain vectors representing the estimated motion. Motion compensation uses the knowledge of object motion obtained to achieve data compression [4]. In a video scene, motion can be a complex combination of translation and rotation. Such motion is complicated to estimate and requires huge amount of processing. However, translational motion is simply estimated and has been used successfully for motion compensated coding. Most of the motion estimation algorithms make the following assumptions: objects move in translation in a plane that is parallel to the camera plane, i.e., the effects of camera zoom, and object rotations are not considered. Illumination is spatially and temporally uniform, and occlusion of one object by another, and uncovered background are neglected [5]. Several motion estimation approaches have been proposed, two of which are the pel-recursive algorithms (PRAs) and the block-matching algorithms (BMAs). In general, BMAs are more suitable for a simple hardware realization because of their regularity and simplicity. They estimate motion on the basis of rectangular blocks and produce one motion vector for each block. These algorithms assume that all the pels within a block have the same motion activity. PRAs involve more computational complexity and less regularity, so they are difficult to realize in hardware [3].

In a typical BMA, each frame is divided into blocks, each of which consists of luminance and chrominance blocks. Usually, for coding efficiency, motion estimation is performed only on the luminance block. Each luminance block in the present frame is matched against candidate blocks in a search area on the reference frame. These candidate blocks are just the displaced versions of original block. The best matched i.e., lowest distortion, candidate block is found and its displacement

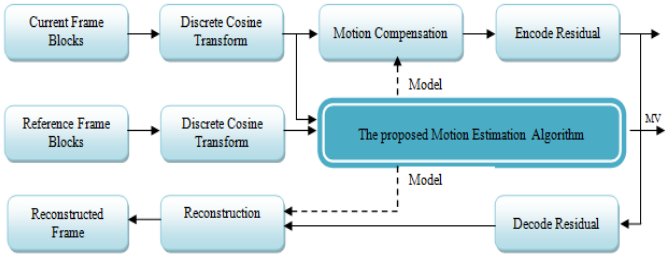


Fig. 1: Motion Estimation and Compensation in Block Diagram.

(motion vector) is recorded. In a typical inter-frame coder, the input frame is subtracted from the prediction of the reference frame. Consequently the motion vector and the resulting error can be transmitted instead of the original luminance block; thus inter-frame redundancy is removed and data compression is achieved. At receiver end, the decoder builds the frame difference signal from the received data and adds it to the reconstructed reference frames. The summation gives an exact replica of the current frame. The better the prediction the smaller the error signal and hence the required transmission bit rate is reduced [4]. Although the full-search motion estimation algorithm yields the best results, its intensive computation process limits its practical application. However, there is a trade-off between the complexity of the algorithm and the quality of the predicted frame. With this trade-off in consideration, many fast search motion estimation algorithms have been developed in literature. The fast search motion estimation algorithms can be classified mainly into two broad categories: spatial domain and frequency domain. The term spatial domain refers to the video frame plane itself, and approaches in this category are based on direct matching of pixels in successive video frames [2]. In the spatial domain, high correlation exists between pixels inside each frame block; therefore, the general block matching usually require measuring the similarities between every pair of pixels inside each block. Frequency domain motion estimation algorithms can be used to test the similarities between groups of frequencies which form a subset of the total frequencies in each block; therefore fewer comparisons can be considered for this task with a massive reduction in block matching calculations. Transforming a video frame into the frequency domain is a vital step that has to be performed in intra-frame coding. In this research, a new low complexity fast search motion estimation algorithm is proposed, as shown in Figure 1. The algorithm uses the intra-coded frequency domain transformed frame in order to perform the proposed block matching technique. Section 2 provides an up to date literature review of both spatial and frequency domain motion estimation algorithms. Section 3 introduces the spatial-frequency transformation process. In section 4, the proposed matching technique is described. Section 5 provides the experimental results. Finally, section 6 concludes this research.

II. LITERATURE REVIEW

Many sub-optimal spatial domain motion estimation algorithms have been proposed in literature such as the well-known: Cross-Search, Spiral-Search, Three-Steps-Search, Two-Dimensional-Logarithmic-Search, Binary-Search, Four-Step-Search, Orthogonal-Search, and Diamond-Search algorithms. These algorithms are called sub-optimal because although they are computationally more efficient than the Full Search, they do not result in a quality that is as good as that of the Full Search algorithm [3]. A more recent variant of fast search motion estimation approaches may be found in [7][8][9][10][11][12][13][14]. The extensive variety of algorithms available for block-based motion estimation makes it difficult to choose between them. The choice depends on different criteria, such as: complexity, implementation, matching performance, rate-distortion performance, and scalability [15]. Motion estimation algorithms quality and performance has been a popular research area and different results have been obtained by different researchers. According to Kuhn et al. [16], the Three Step Search gives the best results. The five step diamond search performs well, but suffers in some cases from a too small search range of pixels. The hierarchical search algorithm depicted results which were not as good when compared with other algorithms. On the other hand, alternate pixel sub-sampling depicts very similar results as the original full search algorithm, where no extreme case of performance degradation occurs. According to Ghanbari [3] with regards to speed, the Two-Dimensional-Logarithmic algorithm outperforms the rest of the algorithms at the cost of quality. The Three-Steps-Search achieves a marginal improvement in terms of quality but has a high computational complexity in comparison with the Two-Dimensional-Logarithmic algorithm. The Four-Steps-Search algorithm outperforms the Three-Steps-Search algorithm in terms of complexity; however, its quality does not approach that of Full-Search as the hierarchical algorithms do. Although the complexity of the hierarchical algorithms is worst than some of other fast search algorithms, they outperform any other algorithm in terms of quality and they almost have the same quality as the Full-Search algorithms, with a significant reduction in complexity. Motion estimation in the frequency domain has been investigated by fewer researchers. Argyriou and Vlachos [17] proposed a motion estimation scheme for broadcast-quality digital video applications. The proposed scheme is based on the principle of gradient correlation in the frequency domain. The scheme involves the quad-tree decomposition of a frame. Quad-tree decompositions are obtained by using the motion compensated prediction error to control the partition of a parent block to four children quadrants. The partition criterion is applied iteratively until a target number of motion vectors or a target level of motion compensated prediction error is achieved or, until no more than a single motion component can be identified. Erdem et. al, [18] in their work model the discontinuous motion estimation problem in the frequency domain where the motion parameters are estimated using a harmonic retrieval approach. In the proposed work, the vertical and horizontal components of the motion are independently estimated from the locations of the peaks and they are paired to obtain the motion vectors using a specific procedure. L.Lucchese et al., [19] in their work introduced an alternative for 3-D motion estimation based on the Fourier transform of the 3-D intensity function described by the registered time-sequences of range and intensity data. The proposed system can lead to an unsupervised method for 3-D rigid motion estimation. This method has several advantages since it uses the total available information and not sets of features. Briassouli and Ahuja [20] in their work

analysed a video containing multiple objects in rotational and translational motion through a combination of spatial and frequency domain representations. It is argued that the combined analysis can take advantage of the strengths of both representations. Initial estimates of constant, as well as time-varying, translation and rotation velocities are obtained from frequency analysis. Improved motion estimates and motion segmentation for the case of translation are achieved by integrating spatial and Fourier domain information. For combined rotational and translational motions, the frequency representation is used for motion estimation, but only spatial information can be used to separate and extract the independently moving objects. The proposed algorithms are tested on synthetic and real videos. Tzimiropoulos et al., [21] proposed a frequency domain approach for the detection of symmetries in real images is presented. The framework is based on recent state-of-the-art research where motion estimation techniques are employed to sequentially determine all the associated parameters. In particular, the researchers introduce several modifications regarding the order of symmetry estimation and the detection of the axes of possible bilateral symmetry. Preliminary results demonstrate the efficiency of their approach. Pingault and Pellerin [22] describe a method to test motion transparency phenomena in image sequences based on an image sequence analysis in the frequency domain. It is mainly composed of a Stochastic-Expectation-Maximisation algorithm which provides a new statistical model for this problem. Young and Kingsbur [23] proposed a frequency-domain algorithm for motion estimation based on overlapped transforms of the image data. This method is developed as an alternative to block matching methods. The complex lapped transform is first defined by extending the lapped orthogonal transform to have complex basis functions. The complex lapped transform basis functions decay smoothly to zero at their end points, and overlap by 2:1 when a data sequence is transformed. A method for estimating cross-correlation functions in the complex lapped transform domain is developed. Block matching is subject to noise, therefore, researchers have attempted to use a predictor-corrector type estimator such as the Kalman Filter in order to enhance the motion vectors predictions and measurements and to obtain a better performance. The Kalman filter addresses the general problem of estimating the state of a discrete-time controlled process that is governed by the linear stochastic difference [24]. Various researches has been conducted in this field to incorporate Kalman filtering with block matching algorithms for the purpose of obtaining better motion vectors estimates such as the work in [25][26][27][28][29][30]. Although Hierarchical motion estimation algorithms (usually combines several block matching algorithms at different levels) are widely used in the spatial domain for their accuracy at extra complexity, those algorithms have not yet been investigated in the frequency domain. In this work, the authors propose a frequency based two-level hierarchical motion estimation algorithm that incorporates a novel searching method at the top-level of the hierarchy, with a matching criterion that reduces the complexity of the proposed method. The next section discuses the spatial-frequency transformation method used in this research.

III. TRANSFORMATION FROM SPATIAL TO FREQUENCY DOMAIN

Video frames enclose high spatial and temporal correlation between adjacent pixels and consecutive frames respectively. Video compression involves reducing the spatio-temporal redundancy using intra-frame and inter-frame coding methods, in order to reduce the required number of bits that represent a video. The former process involves, transforming the block into the frequency domain, and quantizing the transformed coefficients in order to achieve compression. In the latter, further compression may be achieved by exploiting the temporal redundancy using motion estimation and compensation algorithms. In intra-frame coding the transformation process is used in order to represent the image data in another form, by switching from the spatial to the frequency domain or vice versa. The choice of transformation technique is governed by a number of criteria. However, regardless of the chosen transformation method, data in the transform domain should be separated into components with minimal inter-dependence. Moreover, any transformation method should be reversible and computationally tractable with low memory requirement and a low number of arithmetic operations [5]. Many transforms have been proposed for video coding, and the most popular transforms can be classified into two categories: block-based and frame-based transformations [31]. Although frame-based transformations are more suitable for images and give better decorrelation results, block-based methods are widely used in video coding and are more appropriate for this research, for the reason that motion estimation algorithms are based on block matching criteria which are based on matching portion of the frequency block in this research. The Discrete Cosine Transform (DCT) is chosen as the transformation method due to its accuracy and low complexity; DCT operates on B, a block of N N samples (pixels) and creates Z, an N N block of coefficients. A discrete cosine transform (DCT) expresses a sequence of data points in terms of a sum of cosine functions oscillating at different frequencies. The DCT is valuable for various applications in science and engineering. The use of cosine rather than sine functions is important in image and video applications as the sine functions lead to complex numbers and unnecessary complex computation. Specifically, a DCT is a Fourier-related transform that only uses real numbers. The most common variant of discrete cosine transform is the type-II DCT, which is often called ”the DCT”; its inverse, the type-III DCT, is correspondingly often called ”the inverse DCT” or ”the IDCT”. The action of the DCT (and its inverse, the IDCT) can be described in terms of a transform matrix W (see eq 1). The DCT of an N N sample block is given by: . And the inverse DCT (IDCT) is given by: , where B is a matrix of samples, Z is a matrix of coefficients, and are represented as in eq.1, and eq.2 respectively,

$$B_{ij} = \sum_{x=0}^{N-1} \sum_{y=0}^{N-1} C_x C_y Z_{xy} \cos \frac{(2j+1)y\pi}{2N} \cos \frac{(2i+1)x\pi}{2N} \tag{1}$$

$$Z_{xy} = C_x C_y \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} B_{ij} \cos \frac{(2j+1)y\pi}{2N} \cos \frac{(2i+1)x\pi}{2N} \tag{2}$$

W is an $N \times N$ transform coefficients matrix, the elements of W are defined based on eq.3:

$$W_{ij} = C_i \cos\left[\frac{(2j+1)i\pi}{2N}\right] \quad (3)$$

where $c_i = \sqrt{\frac{1}{N}}$ for $c_i = 0$ and $c_i = \sqrt{\frac{2}{N}}$ for $c_i \geq 0$;

The DCT Transformation matrix coefficients are image independent; they are always fixed for the same block size, and hence can be pre-computed and stored separately. The output of a two-dimensional DCT is a set of N N coefficients representing the image block data in the DCT domain and these coefficients can be considered as weights of a set of standard basis patterns [5]. The basis patterns for an 88 DCTs are composed of combinations of horizontal and vertical cosine functions. Any image block may be reconstructed by combining all N N basis patterns, with each basis multiplied by the appropriate weight. The result of the DCT transformation for a block in the spatial domain is a set of frequencies that are arranged in a zigzag ascending order. The frequency located at is the lowest frequency (highest wavelength) and is called the DC value. This value represents the general style of the block and is considered the most important frequency amongst all the other frequencies in the block. The rest of the frequencies range from low to high in a zigzag pattern and are called the AC values. The AC values contain the details of the block which ranges from general to fine details, as we progress forward in the zigzag order. For the purpose of this research, video frames are intra-coded using 4x4 and 8x8 DCT transformation block sizes at different levels of the proposed hierarchy. Further, selected frequencies are used in the block matching algorithm in order to obtain the best match as will be illustrated in the next sections.

IV. MATCHING CRITERION

The matching criterion has a huge impact on the performance of the algorithm. When comparing algorithms, different criteria should be investigated such as the well-known Mean Square Error (MSE), the Mean Absolute Difference (MAD), and the Sum of Absolute Difference (SAD). In addition to those standard criteria, other specific criteria were introduced by researchers such as: the Reduced Bit Mean Average Difference, the Min/Max-Error, and the Different Pixel Count [5]. The nonnegative matching error function (Sum of absolute differences as shown in eq.4) is normally defined over all the positions to be searched.

$$D_{m,n} = \sum_{x=0}^{N-1} \sum_{y=0}^{N-1} |\mu - \gamma| \quad (4)$$

where $\mu = f_t(r+x, s+y)$ is the current frame reference block of its upper left pixel at the coordinate (r, s) and its lower right pixel at coordinate $(r+x, s+y)$, $\gamma = f_{t-1}(r+m+x, s+n+y)$ is a candidate block in the previous frame, and $-W \leq m, n \leq W$ (W is the window size). The matching criterion has an enormous impact on the performance of the algorithm, therefore, reducing the number of required computations negatively affects the matching results when applied in the spatial domain since the pixels are highly correlated and it is impossible to differentiate between the significance of pixels inside a given block. However, reducing the number of required computations is possible in the frequency domain because the frequencies are highly de-correlated, making it possible to categorize frequencies based on their significance inside each transformed block. A simple method of block matching algorithm is the Full-Search Algorithm, where $D_{m,n}$ is computed for all $(2W+1)^2$ positions of candidate blocks in the search window. This results in $(2W+1)^2 \times N^2$ subtractions, $(2W+1)^2 \times (2N^2-1)$ additions and $(2W+1)^2$ comparisons for each reference block. However, with fast search motion estimation algorithms, the SAD criterion shown in Eq. (1) requires N^2 computations of subtractions with absolute values and N^2 additions for each candidate block at each search position. The absence of multiplications makes this criterion computationally more attractive for real-time implementation. In this work the SAD criterion is used but with fewer numbers of computations. This approach requires $N^2/4$ computations of subtractions with absolute values, and $N^2/4$ additions for each candidate block at each search position. As stated earlier, the frequency coefficients produced by the DCT represent the basis functions to the source image, where the basis function increases as we move in a Zig-Zag pattern from the top-left to the bottom-right corners of the block. As shown in Fig. 2, the highest left, and the lower right corners' coefficients contain the lowest and the highest vertical and horizontal frequencies respectively. In this research, the first quarter of the transformed block is used in the matching criterion. Frequencies in this quarter consist of a combination of low and reasonably high frequencies, representing the most important characteristics of the block. As will be shown later, information in this portion of the block is sufficient to distinguish the desired block from amongst the rest of the neighbouring blocks that can be assumed as candidate locations for the search operation.

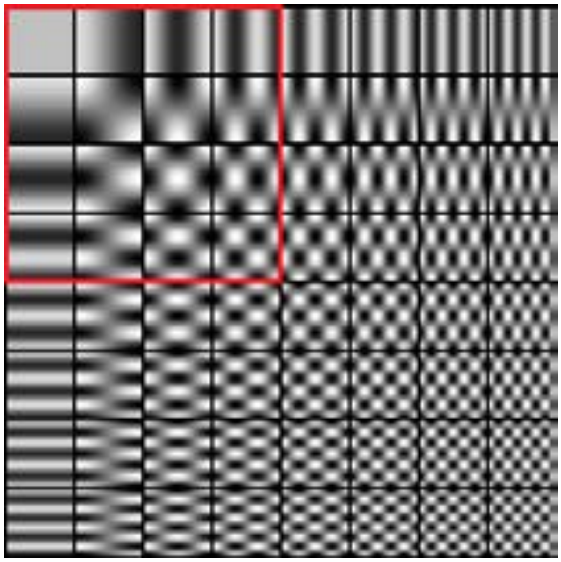


Fig. 2: Basis functions of an 8x8 DCT block, the top left quarter of the block is used in the matching criterion in the work, as frequencies in this quarter consists of a combination of low and reasonably high frequencies that represents the most important characteristics of the block.

A. THE PROPOSED HIERARCHICAL SEARCH MOTION ESTIMATION ALGORITHM

Hierarchical block matching techniques attempt to merge the advantages of large blocks with those of small blocks. The reliability of motion vectors is influenced by the selected block sizes. Larger blocks are more likely to track actual motion than smaller ones and thus are less likely to converge on local minima. Although such motion vectors are reliable, the quality of matches of large blocks is not as good as that of small blocks. Hierarchical block matching algorithms exploit the motion tracking capabilities of small blocks and use their motion vectors as starting points for searches for larger blocks. Normally, three level hierarchical searches are widely used in the spatial domain, where initially large blocks are matched and the resulting motion vector provides a starting point for a search for a smaller matching block. In this research a two-level hierarchy is used in the frequency domain, where a new search pattern is applied at the top of the hierarchy. The following summarises the steps of the proposed algorithm. This hierarchy is applied on both the previous and the current video frames:

- Step 1: the lowest level (level-1) consists of the video frame at its full resolution. This step involves sub-sampling level-1 by a factor of 2 in vertical and horizontal directions to produce level-2.
- Step 2: In this step, the frames at different levels (level-1) an (level-2) are transformed into the frequency domain using the two dimensional discrete cosine transform with different block sizes (4×4 block size at level-2 and 8×8 at level-1). The search starts from the highest level (level-2) using block sizes, where the new proposed cross-diamond search pattern (described in the next section) is used to get a coarse motion vector that will be passed to level-1 (lowest level).
- Step 3: In this step, the Enhanced Three-Step-Search algorithm (described in section 3.3) is used on level-1 utilizing 8×8 block sizes, to get the final motion vector that will be added to the previous image to get the next predicted image frame.

B. THE PROPOSED CROSS-DIAMOND SEARCH PATTERN

The steps of the proposed algorithm are applied on the two hierarchies (current frame and previous frames hierarchies) and the search pattern is applied between corresponding levels of the hierarchies. The steps of the algorithm can be summarized as follows (Figure-2 illustrates the proposed method):

- Step 1: This step involves setting the window size to $2^N + 1$ where N is the number of levels in the hierarchical search (i.e., $N = 2$ in the proposed algorithm), and setting the step size to the standard 2^N (i.e., step size= 4).
- Step 2: Starting at the center point location around the obtained coarse motion vector, this step involves searching the four points forming a diamond shape pattern. The best match will be passed to step-3 as the new center of search.

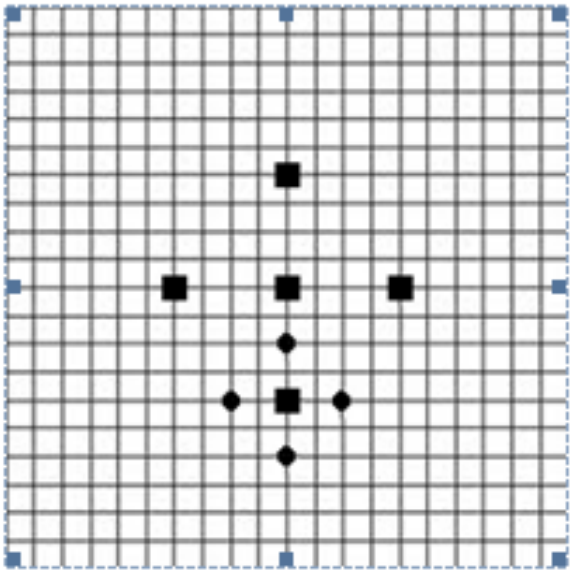


Fig. 3: The first step of the proposed algorithm involves four locations to be searched around the canter forming a diamond shape pattern, the second step involves additional four locations to be searched around the best match point obtained from the first step with step size reduced to the half.

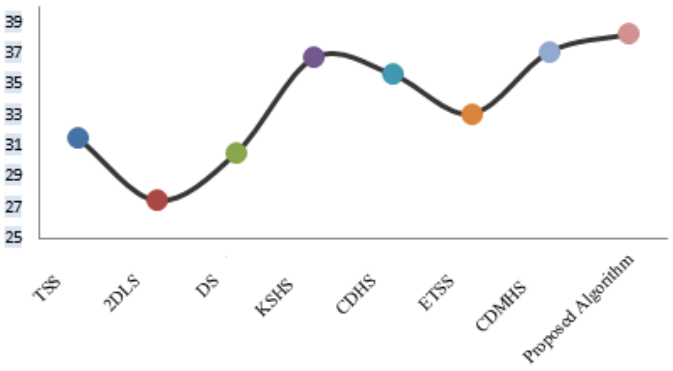


Fig. 4: Visual representation of PSNR values presented in Table-1

- Step 3: This step involves setting the step size to $N/2$, searching the four neighboring points around the new center obtained from step-2, and forming the diamond shape (see Fig. 3). If the step size >1, then the step size is set to $N/2$ and step-2 is repeated; otherwise, the best match point that is found is passed to level-1 of the hierarchy. The best matched block will be used to obtain the resulting motion vector and will be passed to the lower level (level-1), where it will be used as the centre point of search for the ETSS algorithm.

C. THE ENHANCED THREE-STEP-SEARCH ALGORITHM (ETSS)

The obtained motion vector in the previous hierarchy is used as the centre point of the TSS algorithm (using 8×8 block sizes). The TSS starts with 9 points to be checked (that

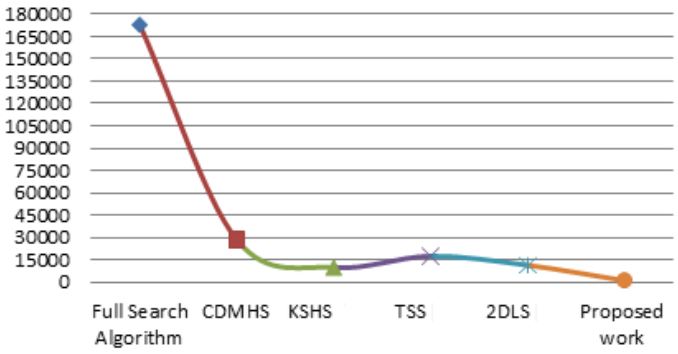


Fig. 5: Visual representation of the complexity of the proposed work compared to the rest of the standard and benchmark algorithms.

form a rectangular shape). The TSS is described as follows and is based on the following conditions:

- Condition 1: If the best match is the centre of the search window, the algorithm stops, and the same motion vector (obtained from the previous hierarchy) is considered as a final motion vector for the current block.
- Condition 2: If the best match is one of the eight rectangular neighbouring points, then the benchmark Three-Step-Search algorithm is performed based on the following criteria:
  - Search the location around the best match, and set the step size to $S = 2^{N-1}$.
  - Search the eight locations $+/-S$ surrounding the location centre.
  - Reduce the step size to $S = S/2$ and then go back to step 2.
  - Terminate when $S = 1$

The number of comparisons required to find the best match is $8N + 1$ for a search area of $+/-2N - 1$ pixels in N-Step Search algorithms. Since $N = 3$, the required computations are 25.

V. EXPERIMENTAL RESULTS AND DISCUSSIONS

In this research, 13 standard Quarter Common Intermediate File (QCIF) and Common Intermediate File (CIF) video sequences of different motion contents are used to compare the performance of different algorithms. These video sequences are categorized into three classes; Class A, Class B, and Class C, with increasing motion complexity. This means that the video sequences in Class A have slow motion activities, those in Class B have medium motion activities and those in Class C have high or complex motion activities. The video sequences of Silent, Claire, Mother and Daughter belong to the Class A category. The video sequences of News, Suzie, Miss America, and Hall monitor are of moderate motion thus categorised as class B. Finally, the video sequences of Foreman, Carphone, Salesman, Flower, Coastgard, and Akiyo sequence which have fast object translation with high motion activity belong to Class C. More than 650 video frames of standard test video sequences with different formats were used in the experiments. These comprise of the first 50 frames from each of the 13 test video sequences listed in Table-2. The results are evaluated subjectively and objectively. The PSNR (Peak signal to Noise Ratio) is used to objectively evaluate the system performance, where $PSNR = 10\log_{10}(L^2/MSE)$ is measured in decibel units (db units), where L is the range of pixel values (when the luminance component is only used $L = 255$), and $MSE = 1/N\sum_{i=1}^{N}(x_i - y_i)^2$ is the Mean Square Error, where N is the number of the pixels per frame, and x_i, y_i are the pixels within the original and predicted frames, respectively. A standard measurement states that, if the PSNR result is larger than 30db, then the difference between the original image and the resulted processed image will not be recognized through the human visual system. The higher the PSNR, the better quality it represents. Using the original and reconstructed frames, Table-2, illustrates the PSNR values for the proposed algorithm and compares the results with those of other benchmark and standard algorithms. The results in Table-2 show that, using the standard set of test videos, the proposed algorithm outperforms the standard Three-Step-Search [32] with 17% average enhancement, Two-Dimensional-Logarithmic-Search [33] with 28.6% average enhancement, and the Diamond Search algorithm [34] with 20.2% average enhancement. The average PSNR shows an enhancement of 16db units in some particular cases, signifying an enormous enhancement of quality.

In addition to the above standard algorithms, the enhanced Three-Step-Search algorithm [35], the Kalman simplified hierarchical search algorithm [36], and the Cross-Diamond Modified Hierarchical Search Algorithm [37] are chosen as the state-of-the-art benchmarks in the field of hierarchical search algorithms. When compared with the proposed work, using the same set of test videos, the average PSNR results show that the current proposed algorithm outperforms the work in [35][36], and [37] with 13.49%, 4% and 3% average enhancement respectively. The results of the PSNR values of the proposed work can be improved if the Kalman filter is applied as a stochastic predictor/ corrector estimator. Unfortunately, this will add to the complexity of the proposed work. Even without the use of an additional set of filters, the proposed algorithm has results comparable to those of the full search algorithm. Fig.5, visually illustrates the significant quality enhancement of the proposed work when compared with the rest of the benchmark and standard algorithms. In addition to the objective evaluation, a subjective evaluation of the proposed work can be seen in Fig.6-9, which illustrate visual representations of the reconstructed frames resulted from the proposed HS algorithm when applied to the set of standard videos with different class categories. The reconstructed frames in Fig.6 belong to class A with low motion complexity. The reconstructed frames in Fig.7 belong to class B with moderate motion complexity. Finally, the reconstructed frames in Fig. 8 and Fig. 9 belong to class C with high motion complexity.

The complexity of the proposed algorithm is evaluated and compared against some of the benchmark searching methods. Table-2 shows that the proposed algorithm outperforms the Full Search with a lower number of operations per block. Compared the FSA, the proposed algorithm requires 0.67% of the total number of additions, 0.7% of the total absolute differences, and 13.7% of the total number of comparisons. This can be summarized with a total Number of Operations

TABLE I: PSNR values of the proposed work, compared to standard and benchmark algorithms.

Video Sequence	TSS [32]	2DLS[33]	DS [34]	KSHS [36]	ETSS [35]	CDMHS [37]	Proposed Algorithm
Akiyo	32.74	30.74	31.23	34.92	33.31	34.89	35.61
Carphone	29.43	26.73	28.32	33.32	30.12	33.61	35.23
Claire	33.87	25.23	32.78	36.14	33.94	36.47	38.49
Coastguard	30.61	28.45	31.82	44.83	33.81	45.56	46.98
Flower	31.72	28.12	30.14	34.22	32.34	34.24	34.47
Foreman	29.26	25.80	27.96	33.79	30.04	33.91	35.71
Hall Monitor	34.81	30.11	32.19	41.81	36.81	41.73	42.34
Miss America	32.19	27.72	31.90	36.60	32.38	36.85	37.29
Mother and Daughter	31.20	28.98	28.02	38.36	34.62	38.78	39.84
News	29.50	26.29	30.74	36.31	31.71	37.64	39.21
Salesman	31.7	24.22	32.58	34.32	34.23	35.47	36.43
Silent	28.60	25.27	28.35	34.78	31.44	35.29	36.64
Suzie	33.54	29.17	30.49	37.79	35.32	37.95	38.83
Average PSNR	**31.47**	**27.45**	**30.50**	**36.71**	**33.08**	**37.11**	**38.24**

Fig. 6: Samples of reconstructed frames from Class A with Slow motion activities video sequences.:
a Reconstructed Silent video, frame number 43.
b Reconstructed Claire video, frame number 31.
c Reconstructed Mother and Daughter video, frame number 16.

Fig. 7: Samples of reconstructed frames from Class B with moderate motion activities video sequences:
a Reconstructed Suzie video, frame number 49.
b Reconstructed News video, frame number 25.
c Reconstructed Hall video, frame number 49.

Per Block (NOPB), where the proposed algorithm requires only 0.69% of the total NOPB required by the FSA (i.e., 99.31% reduction in complexity). When compared to the rest of the algorithms, the algorithm require 4.1% of the total complexity required by the Cross-Diamond Modified Hierarchical Search algorithm [37], 12% compared to Kalman Simplified HSA [36], 6.8% compared to the TSS algorithm [32], and 10.4% compared to the 2DLS algorithms [33]. This enormous reduction in complexity is due the substantial reduction in the total number of operations required in the proposed matching criterion. Fig. 4 visually illustrates the significant complexity reduction of the proposed work compared to the rest of the benchmark and standard algorithms.

VI. CONCLUSION

Digital videos require a large amount of bandwidth for transmission or storage. Therefore, researchers have attempted to develop algorithms that compress video data whilst maintaining the highest quality possible. Motion estimation with block matching algorithms has proven to be effective in the reduction of video bit-rates while preserving the good quality. Block matching algorithms involve searching for block movements between consecutive video frames in the spatial domain. Hence, various fast searching algorithms have been investigated, each aiming at reducing the number of comparisons. However in the spatial domain, the high correlation that exists between pixels inside each frame block forces testing the similarities between every pair of pixels inside each block. In

TABLE II: The Total Number of Operations per Block (NOPB) required by the proposed algorithm, compared to the benchmark algorithms.

Algorithm	Addition	Multiplication	Absolute Difference	Comparisons	NOPB
Full Search Algorithm	$(2w+1)^2 \times (2N^2-1)$ =114975 $w=7, N=16$	Null 0	$(2w+1)^2 \times (N^2)$ =57600 $w=7, N=16$	$(2w+1)^2$ =255 $w=7$	**=172800**
Cross-Diamond MHS [37]	$\frac{\text{level(3)}}{(2w+1)^2 \times (2N^2-1)}$ $w=3, N=4$ $\frac{\text{level(2)}}{8 \times (2N^2-1), N=8}$ =21421 $\frac{\text{level(1)}}{23 \times (2N^2-1), N=16}$	Null 0	$\frac{\text{level(3)}}{(2w+1)^2 \times (N^2)}$ $w=3, N=4$ $\frac{\text{level(2)}}{8 \times (N^2), N=8}$ =7184 $\frac{\text{level(1)}}{23 \times (N^2), N=8}$	$\frac{\text{level(3)}}{(2w+1)^2}$ $w=3$ $\frac{\text{level(2)}}{=8}$ =80 $\frac{\text{level(1)}}{=23}$	**=28685**
Kalman Simplified HS [36]	$\frac{\text{level(3)}}{(2w+1)^2 \times (2N^2-1)}$ $w=3, N=4$ $\frac{\text{level(2)}}{8 \times (2N^2-1), N=8}$ =6633 $\frac{\text{level(1)}}{8 \times (2N^2-1), N=16}$ $\frac{\text{Kalman Filter}}{10}$	10 10	$\frac{\text{level(3)}}{(2w+1)^2 \times (N^2)}$ $w=3, N=4$ $\frac{\text{level(2)}}{8 \times (N^2), N=8}$ =3323 $\frac{\text{level(1)}}{8 \times (N^2), N=16}$ $\frac{\text{Kalman Filter}}{15}$	$\frac{\text{level(3)}}{(2w+1)^2}$ $w=3$ $\frac{\text{level(2)}}{=8}$ =65 $\frac{\text{level(1)}}{=8}$	**=10016**
TSS [32]	$P \times (2N^2-1)$ =11753 $P=23, N=16$	Null 0	$P \times (N^2)$ =5888 $P=23, N=16$	$P=23$ =23	**=17664**
2DLS [33]	$P \times (2N^2-1)$ =7665 $P=15, N=16$	Null 0	$P \times (N^2)$ =3840 $P=15, N=16$	$P=15$ =15	**=11520**
Proposed Work	$\frac{\text{level(2)}}{8 \times ((N^2/2)-1), N=4}$ =769 $\frac{\text{level(1)}}{23 \times ((N^2/2)-1), N=8}$	Null 0	$\frac{\text{level(2)}}{8 \times (N^2/4), N=4}$ =400 $\frac{\text{level(1)}}{23 \times (N^2/4), N=8}$	$\frac{\text{level(2)}}{=8}$ =31 $\frac{\text{level(1)}}{=23}$	**=1200**

a b c

Fig. 8: Samples of reconstructed frames from Class C with fast motion activities video sequences with complex activities:
a Reconstructed Carphone video, frame number 46.
b Reconstructed Coastguard video, frame number 10.
c Reconstructed Salesman video, frame number 26.

a b c

Fig. 9: Samples of reconstructed frames from Class C with fast motion activities video sequences with complex activities:
a Reconstructed Flower guarden video, frame number 37.
b Reconstructed Foreman video, frame number 46.
c Reconstructed Akiyo video, frame number 37.

this work, video frames are intra-coded and transformed into the frequency domain, where block matching can be applied to test the similarities between a subset of selected frequencies that correctly identifies each block distinctively, this yield to a fewer number of required comparisons that reduces the algorithms complexity. In this work a two-level hierarchical fast search motion estimation algorithm is proposed in the frequency domain that incorporates a novel search pattern at the top level of the hierarchy. In terms of quality and matching performance, the proposed algorithm outperforms the other benchmark algorithms with an enormous reduction in complexity.

REFERENCES

[1] O. Alzubi, *An Empirical Study of Irregular AG Block Turbo Codes over Fading Channels*, Research Journal of Applied Sciences, Engineering and Technology, 11(12), (2015) 1329-1335.

[2] O. Alzubi, *Performance Evaluation of AG Block Turbo Codes over Fading Channels Using BPSK*, ICEMIS '15 Proceedings of the The

International Conference on Engineering (MIS 2015), Istanbul, Turkey, (2015) 36:1-36:6.

[3] M. Ghanbari, *Video Coding: An Introduction to Standard Codecs*, Institution of Electrical Engineers, (1999).

[4] I. Richardson, *Video Codec Design*, John Wiley & Sons, (2002).

[5] I. Richardson, *H.264 and MPEG-4 Video Compression: Video Coding for Next-Generation Multimedia*, John Wiley & Sons, (2003).

[6] J. A. Alzubi, O. A. Alzubi, and T. M. Chen, *A Forward Error Correction Based On Algebraic-Geometric Theory*, 1st ed. Springer International Publishing, (2014).

[7] J. Cai, and W. Pan, *On fast and accurate block-based motion estimation algorithms using particle swarm optimization*, 3rd ed. Elsevier Info. Sci., (2012) 53-64.

[8] E. Cuevasa, D. Zaldvara, M. Prez-Cisnerosa, H. Sossab, and V. Osunab, *Block matching algorithm for motion estimation based on Artificial Bee Colony (ABC)*, Elsevier Appl. Soft Comput., 13(6), (2013) 3047-3059.

[9] E. Cuevasa, D. Zaldvara, M. Prez-Cisnerosa, and D. Olivab, *Block-matching algorithm based on differential evolution for motion estimation*, Elsevier Eng. Appl. Artif. Intel., 26(1) (2013) 488-498.

[10] Z. Cui,G. Jiang, S. Yang, and C. Wu, *A new fast motion estimation algorithm based on the loopepipolar constraint for multiview video coding*, Elsevier Signal Process. Image Commun., 27(2), (2012) 172-179 .

[11] J. Fabrizioa, S. Dubuissonb, and D. Brziatb, *Motion compensation based on tangent distance prediction for video compression*, Elsevier Signal Process. Image Commun., 27(2), (2012) 153171.

[12] C. Je, and H-M. Park, *Optimized hierarchical block matching for fast and accurate image registration*, Elsevier Signal Process. Image Commun.,28(7), (2013) 779-791.

[13] K. Lai, Y. Chan, C. Fu, and W. Siu, *Hybrid motion estimation scheme for secondary SP-frame coding using inter-frame correlation and FMO*, Elsevier Signal Process. Image Commun., 27(1), (2012) 115.

[14] F. Di Martinoa, V. Loiaa, and S. Sessab, *Fuzzy transforms for compression and decompression of color videos*, Elsevier Info. Sci., 180 (20), (2010) 3914-3931.

[15] S.N. Basah, A. Bab-Hadiashar, and and R. Hoseinnezhad, *Conditions for motion-background segmentation using fundamental matrix*, IET Comput. Vis., 3(4), (2009) 189200.

[16] P. Kuhn, G. Diebel, S. Herrmann, A. Keil, H. Mooshofer, A. Kaup, R. Mayer, and W. Stechele, *Complexity and PSNR-Comparison of several Fast Motion Estimation Algorithms for MPEG-4*, In: Proceedings of SPIE, (1998) 486489.

[17] V. Argyriou, and T. Vlachos, *Quad-Tree Motion Estimation in the Frequency Domain Using Gradient Correlation*, IEEE Transactions on Multimedia, 9(6), (2007) 1147 1154.

[18] C. Erdem, G.Karabulut, E.Yanmaz, and E.Anarim, *Motion estimation in the frequency domain using fuzzy c-planes clustering*, IEEE Transcations on Image Processing, 10(12), (2001) 1873-1879.

[19] L. Lucchese, G. Doretto, and G. Cortelazzo, *Frequency domain estimation of 3-D rigid motion based on range and intensity data*, In Proceedings of the International Conference on Recent Advances in 3-D Digital Imaging and Modeling, (1997) 107-112.

[20] A. Briassouli, and N. Ahuja, *Integrated spatial and frequency domain 2D motion segmentation and estimation*, The Tenth IEEE International Conference on Computer Vision ICCV, (2005) 244-250.

[21] G. Tzimiropoulos, V. Argyriou, and T. Stathaki, *Symmetry detection using frequency domain motion estimation techniques*, The IEEE International Conference on Acoustics, Speech and Signal Processing, (2008) 861-864.

[22] M. Pingault, and D. Pellerin, *Motion estimation of transparent objects in the frequency domain*, Harlow, England: Addison-Wesley, Journal of signal processing, 84(8), (2004).

[23] R. Young, and N. Kingsbury, *Frequency-domain motion estimation using a complex lapped transform*, IEEE Transactions on Image Processing, 2(1), (1993) 2-17.

[24] G. Torres, *3-D Kalman Filter for Image Motion Estimation*, 3rd ed. ACM Trans. Math. Softw. 37(2010) 1-16.

[25] J. Kim, and J. Woods, *A Guide to LaTeX*, 3rd ed. IEEE Trans. Image Process. 3, (1998) 42-52.

[26] C. Kuo, C. Chao, and C.Hsieh, *A new motion estimation algorithm for video coding using adaptive Kalman filter*, 3rd ed. J. Real-Time Imaging., 8, (2002) 387-398.

[27] C. Kuo, C. Chao, and C. Hsieh, *An Efficient Motion Estimation Algorithm for Video Coding Using Kalman Filter*, 3rd ed. J. Real-Time Imaging. 8(2002) 253-264.

[28] C. Kuo, S. Chung, and P. Shih, *Kalman filtering-based rate-constrained motion estimation for very low bit rate video coding*, EEE Trans. Circuits Syst. Video Technol. 16, (2006) 3-18.

[29] C. Kuo, C. Hsieh, and C. Chao, *Multiresolution Video Coding Based on Kalman Filtering Motion Estimation*, J. Visual Commun. Image Represent. 13, (2002) 348-362.

[30] A. Sholiyi, J. A. Alzubi, O. A. Alzubi, O. Almomani, and T. O'Farrell, *Near Capacity Irregular Turbo Code*, Indian Journal of Science and Technology, 8 (23), (2015).

[31] R. Gonzales, and R. Woods, *Digital Image Processing*, 3rd ed. Prentice Hall (2008).

[32] T. Koga, K. Iinuma, A. Hirano, Y. Iijima, and T. Ishiguro, *Motion compensated interframe coding for video conferencing*, Proc. Nut. Telecommun. Conf. (1981).

[33] J. R. Jain, and A. K. Jain, *Displacement measurement and its application in interframe image coding*, IEEE Trans. Commun. 29(12), (1981) 1799-1808.

[34] S. Zhu, and K. K. Ma, *A new diamond search algorithm for fast block matching motion Estimation*, IEEE Trans. Image Process. 9(2), (2000) 287-290.

[35] X. Jing, and L. P. Chau, *An efficient three-step search algorithm for block motion estimation*, IEEE Trans. Multimedia 6(3), (2004) 435-438.

[36] S. Tedmori, and N. Al-Najdawi, *Hierarchical stochastic fast search motion estimation algorithm*, IET Comput. Vis. 6(1), (2012) 21-28.

[37] N. Al-Najdawi, M. N. Al-Najdawi, and S. Tedmori, *Employing a novel cross-diamond search in a modified hierarchical search motion estimation algorithm for video compression*, Elsevier Inform. Sci., 268, (2014) 425435.

A Survey On Interactivity in Topic Models

Patrik Ehrencrona Kjellin
School of Software Engineering
Tongji University
China

Yan Liu
School of Software Engineering
Tongji University
China

***Abstract*—Trying to make sense and gain deeper insight from large sets of data is becoming a task very central to computer science in general. Topic models, capable of uncovering the semantic themes pervading through large collections of documents, have seen a surge in popularity in recent years. However, topic models are high level statistical tools; their output is given in terms of probability distributions, suited neither for simple interpretation nor deep analysis. Interpreting the fitted topic models in an intuitive manner requires visual and interactive tools. Additionally, some measure of human interaction is typically required for refining the output offered by such models. In the research, this area remains relatively unexplored – only recently has this aspect been receiving more attention. In this paper, the literature is surveyed as it pertains to interactivity and visualisation within the context of topic models, with the goal of finding current research trends in this area.**

***Keywords*—topic model; latent dirichlet allocation; LDA; interactive; visualisation; IVA; survey; review**

I. Introduction

With the advent of the Internet, a fundamental change has been experienced in the way we access and use information. For instance, a scientist of today wishing to research some subject may be faced with thousands of relevant articles retrieved from journals spanning numerous decades.

With this in mind, it is of increasing importance to be able to extract useful information and make sense of large collections of documents by the means of computation, a task that has come to be very central to computer science in general. *Topic models* have in recent years emerged as a powerful set of techniques for discovering the underlying semantic structure of large, unstructured collections of documents [1].

Topic models are typically Bayesian or linear algebraic models able to extract abstract topics pervading through large corpora. Through the results of such analysis, the individual documents can then in turn be organised in accordance with the themes.

Powerful as they are, topic models do suffer from some problems that may deter some users, or at the very least prevent them from reaping the full benefits of the methods. Often, the models are treated as "black box" approaches without regard for the underlying assumptions they are based on. Parameter tuning can prove difficult without a full understanding of the specific technique to be employed [2]. Additionally, the emerging topics are by no means guaranteed to be sensible to a human reader – motivating the use of human knowledge and user interaction as an additional step toward more coherent and sensible results [3].

Furthermore, the raw, numerical output of topic models may not always lend itself to easy interpretation. Interactive visual analysis in general has proven a useful tool for interpreting and gaining insight from the results of topic models in an intuitive manner. Despite the fact that topic models in general have been subject to a great deal of research in recent years, the visualisation of topic models is still a relatively unexplored area [4].

The *objective* of this paper is to survey the use of visual and interactive data analysis in conjunction with topic models in the literature. In particular, the author is interested in finding out when, how, and for what purposes interactive visual analysis have been used to enhance topic models, and in which ways visualisation can be used to interpret its results.

In Section II and III, respectively, we provide the survey methodology employed and a brief background on topic modeling techniques and interactive visual analysis. In Section IV, a survey of the literature related to interactivity and interactive visual analysis within the context of topic modeling is presented. Finally, in Section V, we conclude the survey with what we perceive to be future work in the integration of visualisation and interaction in the context of topic models.

II. Survey Methodology

Here we describe the methodology for finding and selecting papers for review, and the reasoning behind it.

Visualisation of fitted topic models is a relatively young field; while many topic model papers include some degree of output visualisation, it is rarely the main focus of the paper. Papers purely dealing with the subject are somewhat sparse, and to the knowledge of the author, no summarised overview of such papers exist. Additionally, we are interested in techniques and methods that not only visualise topic models, but also provide the user with some degree of interactivity.

Papers candidate for review have been found primarily through common search engines (i.e., Google) and the digital libraries of **ACM** and **IEEE**. Search terms used are various mixtures of {topic, model, IR, interactive, visualisation, visual, IVA}. Papers have then been selected based on their relevancy, as deemed by the author upon glancing over the contents. As a basic criteria for relevancy, the papers must be focused on topic modeling, while also including some aspect related to interactive visualisation, possibly incorporating human algorithm supervision.

This survey does not attempt to compare the methods surveyed against other methods of visualisation that do not

include much of an interactive component, as no such papers are reviewed. It simply attempts to create an overview of current research trends within this specific subset of visualisation techniques, as they relate to topic models.

III. Background

Here we provide some brief descriptions of topic modeling techniques and interactive visual analysis in general. For a more in-depth, comprehensive view of these topics, we refer to relevant papers.

A. Topic Models in Brief

In information retrieval (IR), the general term *topic model* refers to a suite of algorithmic approaches to discovering the latent topics present in a collection of documents. Some basic vocabulary is necessary for describing the general concept of topic models. Formally,

- A word w is a basic unit of data (for instance, a string of alphanumerical characters, but topic modeling can also be applied to other domains than natural language processing)
- A document d is an ordered sequence of N words, $w_1, w_2, \ldots, w_N$.
- A corpus is an unordered set of M documents, denoted by $D = \{d_1, d_2, \ldots, d_M\}$.

Topic modeling then consists of taking a corpus D as input and computing K topics (typically in terms of multinomial distributions over the words in the vocabulary), and associating each document with the relevant topic (again, in terms of a multinomial distribution over the different emerging topics).

Early attempts at tackling this problem were however largely concerned with creating a *term-document matrix*, describing the relative frequency of words in each document $d \in D$. This method is useful for many applications, but insufficient in terms of topic modeling, as such a matrix provides little size reduction w.r.t the original corpus, and does not take into account the relationships between words within a document or documents within a corpus [5].

Further research resulted in the strictly linear algebraic approach Latent Semantic Indexing (LSI), which uses singular-value decomposition in order to significantly minimise the term-document matrix [6]. This was later on extended by Probabilistic LSI (PLSI) [7], an early generative model attempting to correct some of the statistically unsound aspects of LSI.

1) Latent Dirichlet Allocation: Latent Dirichlet Allocation (LDA) is a generative statistical model loosely based on earlier work on LSI and probabilistic variations thereof. LDA attempts to address some perceived shortcomings found in the previous generative model pLSI.

Namely, in pLSI, parameters to be estimated grow linearly with the size of the corpus. It also has a strong tendency for overfitting, and of even greater consequence, the model is unable to generalize topic mixtures onto previously unseen documents (not part of the training data) [2], [5]. Through correcting these problems with a truly generative model, LDA has seen a surge in popularity and has acted like a springboard for numerous other advancements in IR.

In LDA, documents are regarded as mixtures of a finite set of K underlying topics, where the parameter K must be specified either by the user or determined through computational inference on the corpus to be analysed. Topics, in turn, are seen as multinomial distributions over the words of the vocabulary.

Inherent to LDA is the assumption that each document $w \in D$ is generated accordingly [5];

$$N \sim Poisson(\xi) \tag{1}$$

$$\theta \sim Dir(\alpha) \tag{2}$$

$$\forall n \in \{1, 2, \ldots, N\} : \tag{3}$$

(a) Choose a topic $z_n \sim Multinomial(\theta)$
(b) Choose a word w_n from $p(w_n z_n, \beta)$,
a multinomial probability conditioned
on the topic z_n.

Here, α and β are smoothing factors for document-topic and topic-term distributions, respectively. LDA is then concerned with inferring the relevant posterior distribution (i.e., given the terms present in the corpus, what are the topics) through a *latent variable model* (the latent variables being the topics).

Further details on the inner workings of LDA is not necessary for this paper, and is therefore omitted. For a more in depth description of LDA, see [5], for instance.

B. Interactive Visual Analysis

Interactive visual analysis (IVA) is a set of techniques incorporating visual analytics (VA) and user interaction in computational or statistical analysis.

Typically, IVA is employed in the task of analysing and attempting to obtain deeper insights from large and possibly complex sets of data, where certain information may not easily be extracted from looking at the data set alone.

It is particularly useful for hypotheses generation and validation, since it equips the user with tools enabling them to look at data sets in a variety of different ways and perspectives [8].

IV. Topic Models and Interactivity

In the surveyed articles, interactivity was most commonly incorporated for addressing one of two concerns:

1) **Human knowledge injection.** The first use case concerns integrating human knowledge in topic models in some manner. Parameter tuning and model constraints through user interactions can enhance models in various ways. Topic models like LDA rely on parameters that, while there are methods for doing so, can not easily be estimated through computation alone. Often, some emerging topics will be nonsensical to a human user [9]. Through interactivity, a topic model can be guided towards achieving more meaningful results.

2) **Topic visualisation.** Visualisation of the emerging topics generated by a topic model appeared in many of the surveyed articles. Graphical tools of many varieties have proved helpful in the task of exploring and attempting to make sense of the results of topic modelling, in order to get an overarching grasp of the various topics spanning some literary corpus. IVA, in the form of allowing users to navigate the corpus and discover the relationships between topics and documents, has been shown to allow users to gain deeper insight in studies [10].

Apart from this, many different task specific measures are to be found in topic model related papers. For instance, topic modeling for the purpose of source code analysis may benefit from visual interactive analysis for displaying the relationships between actual code, requirements documents, and change logs. Here, we focus on general-purpose methods.

A. Human Knowledge Injection in Topic Models

Some researchers have attempted to improve on the results offered by topic models by correcting some of its common shortcomings through incorporating human knowledge in the process. Shortcomings of topic models identified in previous research include non-sensible and incoherent topics [11], certain terms wrongfully belonging to a topic, terms not belonging to a specific topic when they sensibly should [12], et cetera. At its heart, the problem is due to the fact that the objective function subject to optimisation in LDA does not necessarily reflect the expectations on topic quality felt on behalf of a human [13].

Many different extensions to the normal methods (LDA, in particular) have been proposed for improving the results offered by different models. One such approach is by directly incorporating domain knowledge into the model, typically in an *a priori* fashion, thereby introducing a degree of supervision to an otherwise unsupervised model.

In [3], the authors describe constrained LDA (cLDA), a framework for allowing users to add constraints to a model in order to improve it iteratively. Here, constraints are defined on the documents in terms of **must link**, indicating that two documents semantically belong to the same topic, or **cannot link**, representing the opposite.

The general process of this semi-supervised learning, outlined in Figure 1, consists of first performing LDA analysis, then presenting selected documents to the user who adds constraints based on the output, upon which a specialised constrained LDA is computed. The constraints are here encoded as *soft constraints*, which is to say they will be satisfied to some specified degree, but not necessarily fully satisfied.

Similarly, in [12] the authors implement user interaction through allowing users to add constraints to the model formulated in first-order logic (FOL). Here, the FOL constraints are similar to the **must link** and **cannot link** constraints of [3], but defined on word-pairs rather than documents. In some cases, real-time interactive knowledge injection has been applied, such as in [9], where the authors have used similar concepts as in [12] to create a framework allowing users to iteratively and interactively improve topic modeling results.

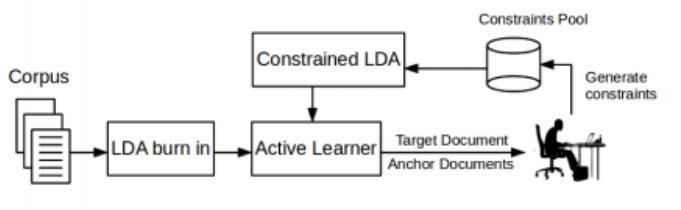


Fig. 1: Diagram from [3], outlining the general process of user-guided constrained topic models.

While the work in [12] and [9] are general-purpose solutions, many of the specialised variations of LDA which incorporate domain knowledge are custom-built, single-purpose methods. All the methods described here for semi-supervised, user guided LDA use only limited visualisation, in the form of matrices or word clouds.

B. Interactive Visualisation of Topic Models

Topic models are high level statistical tools; the raw numerical distributions produced alone are not particularly well suited for intuitive analysis [10]. The visualisation of topic models is an area previously relatively unexplored, which has come under more scrutiny in recent times. Here, some of the concepts and techniques found in the literature (summarised in Table I) are described in terms of their respective unique contributions. There are many ways of visualising topic models, however in this survey of interactive visualisations, the most common representations were found to be either graph based or matrix and text based, along with a few other novel visualisation techniques. The following subsections are organised accordingly.

1) Matrix & Text Based: Matrix or tabular representations are generally easily understood from a user perspective [19]. In *Termite*, the authors present a visual analysis system for quickly assessing fitted topic models computed with LDA, for the specific purpose of user-guided, iterative topic modeling. A corpus is here represented in matrix form, wherein rows correspond to words, and columns to topics. While, in its current state, *Termite* is merely a visual tool, the authors outline future work consisting of expanding it into a complete framework for user-guided, iterative topic modeling with the addition of user interaction (in terms of topic deletion and merging, model parameter adjustments, et cetera) [14].

As mentioned, LDA requires several input parameters, the smoothing variables β, α and the number of topics K. There are no strict guidelines for setting these parameters; tuning

TABLE I: Summary of Surveyed Papers

Type	Papers
Matrix & text based	Termite [14], A. Chaney, et al. [10], The Topic Browser [15], H. Yuening, et al. [9], Y. Yang, et al. [3], D. Andrzejewski, et al. [12]
Graph based	Topicnets [16], ParallelTopics [17], LDAVis [18], LDAexplore [19], TopicPanorama [20], S. Rönnqvist, et al. [4], Hierarchie [21]
Time Visualisation	TextFlow [22], TIARA [23], ThemeRiver [24], RoseRiver [25]

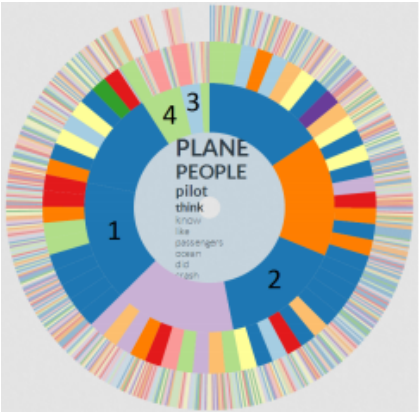


Fig. 2: Sunburst chart from Hierarchie [21].

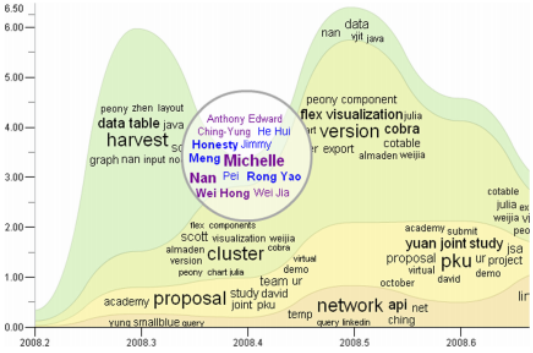


Fig. 3: River-flow representation from TIARA [23].

is usually done through experimentation. This emphasizes the need for good visualisation, allowing users to quickly evaluate the results of their model and tuning parameters accordingly, as is the ambition presented in [14].

In [15], an interactive tool *The Topic Browser* is introduced, with the addition of incorporating document attributes (such as date and authors). Additionally, in their method, a variety of different topic and document metrics are computed and displayed – ranging from simple word counts to pairwise topic and document correlations. Visualisation is done through a mixture of *word clouds* (i.e., terms listed with font sizes determined by their respective probability within a topic) and other text-based views, which may be filtered.

Though many of the existing approaches serve to give a good overview of topics, they seldom capture the relationships between individual documents present in the data. In [10], a topic navigation method for fitted topic models is presented where, in comparison with other methods, greater emphasis is put on individual documents, rather than just topics. Moreover, contrary to previous similar methods, the authors attempt to use visuals rather than numerical data to convey meaning.

Here, the authors provide not only a summarised overview of the corpus as a whole, but also an interactive method for uncovering the discovered structure of the corpus in more detail; in terms of document-document and document-topic relationships. Visualisation is here done entirely through several tabular, text-based views. The technique is validated through a (small) user survey, indicating that the interactive visualisation gives rise to additional insight and discovery.

2) Graph Based: In the topic browser *TopicNets* [16], among other things, a novel visualisation approach is presented allowing users to navigate the corpus through a high-level graph-based representation of topics, wherein semantically similar topics are positionally clustered. Another interesting addition seen in TopicNets is the ability of users to perform additional topic modeling on subsets of the corpus for more fine-grained analysis.

In [4], a corpus is similarly organized and explored through a graph-based visual approach; topics are displayed along with relevant terms, and are linked together with similar topics through shared keywords. In [4], the authors note that topic models are imperfect; review by domain experts is often necessary for perfecting the fitted models – a fact that should be accounted for in further research on topic model visualisation.

ParallelTopics [17] presents several novel representations of fitted topic models generated through LDA. The main distinguishing feature of *ParallelTopics* is that it displays documents in terms of the number of topics pervading through them; documents are plotted in accordance with the number of associated topics, and a document's distribution over different topics can be viewed in more detail. An additional interactive view exists in *ParallelTopics*, which presents topics in terms of their evolution over time. Here, users gain an overview of the pervasiveness of each topic at some particular moment in time, and are able to "zoom in" on specific periods and topics, thereby accessing documents of that time period that have a high probability of containing said topic.

Not uncommonly, the resulting topics are displayed simply by listing the n most probable terms from each topic, and analogously, listing the m most common topics present in each document. This method often leaves a lot to be desired, as it does not comprehensively capture the document and topic relationships discovered in a way that is easily interpreted. In [18], a user study suggests that measuring word relevance purely on the basis of word probability is suboptimal for topic interpretation, as common terms may then appear at the top of several topics. The authors present *LDAvis*, wherein new ideas are introduced on defining *term relevancy* in a more useful way. The topic browser of [18] allows users to visually explore a corpus using such relevancy scores.

TopicPanorama [20] differs from previous graph-based approaches in that it visualises not just one, but several corpus. Here, a topic graph is generated for each corpus through a topic model algorithm Correlated Topic Models (CTM). These are then combined through graph-matching. The authors wish to address the concern that many topic model visualisation tools are unfit to scale for growing data sets.

Hierarchical LDA (hLDA) is a variation of LDA which, contrary to LDA, captures the relationships between different topics [26]. Effectively, the method results in *topic trees* allowing for simpler analysis and greater scalability for large data sets. Recently, some studies have proposed new visual and interactive tools specifically based on such models. For instance, *Hierarchie* [21] uses a sunburst chart (see Figure 2) for displaying the hierarchical topic trees in a compact and

simple fashion. Users may explore the topics of the sunburst chart in terms of keywords, through selecting individual slices.

3) Changes Over Time: Beyond simply visualising fitted topic models statically, recently, plenty of research has been conducted on the visualisation of topics changing over time [22]. Examples of early such methods include *ThemeRiver* [24], where topic evolution over time is displayed in terms of a metaphorical river (see Figure 3 for example) made up of smaller streams (topics). Set against a time line, the river provides users with an intuitive overview of how a corpus has changed and at which point specific topics are more or less pervasive in the associated documents. In *TIARA* [23], a tool that resembled the work of [24] in terms of visuals, a river is similarly used as metaphor for the changing of topics over time. However, *TIARA* also includes a rich set of interactive tools for further analysis; users may zoom in on selected topics or topic segments for further analysis. Additionally, by selecting some keyword in the river view, a user can retrieve relevant documents for further examination.

In [22], *TextFlow* was introduced using a novel approach for LDA output analysis. Here, in contrast to previous research, topics are not only displayed as they progress over time (again, in terms of a river), but the splitting and merging of topics is also captured in the visualisation. It is also highly interactive, allowing users to discover what causes the birth, death, splitting and merging of topics throughout the time period of the associated corpus. *Roseriver* [25] further builds upon the work in [22], using a similar river-flow visual representation, but employs a hierarchical topic model in order to better describe large corpus, and for providing users with different overview levels as desired.

V. Conclusion & Future Work

Topic models have seen a surge in popularity in recent years and have provided a new way of discerning useful information from big, complex sets of data, with applications in several different fields.

Recently, much of the effort put into researching topic models, as has been summarised in this survey, is focused on providing users with tools for interacting with and visualising topic models, both in order to improve results in terms of topic coherence and sensibility, and also for allowing users to fully comprehend and benefit from the model outputs.

Many of the attempts at visualising the results of topic models are not limited to simple visualisation, they also provide varying degrees of user interaction with demonstrably improved results [15], suggesting that IVA may play a central role in making these models more available and intuitive to end users.

Research suggests that different representations may aid in different tasks and lead to discovery at different levels [14]. Whereas an overarching graph or matrix based topic overview may provide a deeper understanding of the corpus as a whole, other visuals displaying word relatedness and topic-topic or topic-document relationships may aid in providing other forms of insight or discovery. Currently, most studies focus on a specific level or representation. In future work, several of the proposed representations could be integrated for a more comprehensive view.

Future work in this area may also include more comprehensive frameworks in terms of combining the interactive elements of semi-supervised LDA described in Section IV-A with interactive visual aids for output analysis – both have demonstrable value in terms of increased usability.

Acknowledgment

This work was supported by the Innovation Project at Tongji University Graduate Education (2014JYJG016).

References

[1] D. M. Blei, "Probabilistic topic models," *Commun. ACM*, vol. 55, no. 4, pp. 77–84, Apr. 2012. [Online]. Available: http://doi.acm.org/10.1145/2133806.2133826

[2] T.-H. Chen, S. W. Thomas, and A. E. Hassan, "A survey on the use of topic models when mining software repositories," *Empirical Software Engineering*, pp. 1–77, 2015. [Online]. Available: http://dx.doi.org/10.1007/s10664-015-9402-8

[3] Y. Yang, S. Pan, D. Downey, and K. Zhang, "Active learning with constrained topic model," in *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*. Baltimore, Maryland, USA: Association for Computational Linguistics, June 2014, pp. 30–33. [Online]. Available: http://www.aclweb.org/anthology/W/W14/W14-3104

[4] S. Rönnqvist, X. Wang, and P. Sarlin, "Interactive visual exploration of topic models using graphs," in *The Eurographics Conference on Visualization (EuroVis)*, R. S. Laramee and M. Chen, Eds. Eurographics, 2014.

[5] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003. [Online]. Available: http://dl.acm.org/citation.cfm?id=944919.944937

[6] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE*, vol. 41, no. 6, pp. 391–407, 1990.

[7] T. Hofmann, "Probabilistic latent semantic indexing," in *Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '99. New York, NY, USA: ACM, 1999, pp. 50–57. [Online]. Available: http://doi.acm.org/10.1145/312624.312649

[8] S. Oeltze, H. Doleisch, H. Hauser, and G. Weber, "Interactive visual analysis of scientific data," Tutorial at the IEEE VisWeek 2012, October 2012. [Online]. Available: http://visweek.org/visweek/2012/tutorial/interactive-visual-analysis-scientific-data

[9] Y. Hu, J. Boyd-Graber, and B. Satinoff, "Interactive topic modeling," in *Association for Computational Linguistics*, 2011.

[10] A. J.-B. Chaney and D. M. Blei, "Visualizing topic models." in *ICWSM*, J. G. Breslin, N. B. Ellison, J. G. Shanahan, and Z. Tufekci, Eds. The AAAI Press, 2012. [Online]. Available: http://dblp.uni-trier.de/db/conf/icwsm/icwsm2012.htmlChaneyB12

[11] D. Newman, J. H. Lau, K. Grieser, and T. Baldwin, "Automatic evaluation of topic coherence," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, ser. HLT '10. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010, pp. 100–108. [Online]. Available: http://dl.acm.org/citation.cfm?id=1857999.1858011

[12] D. Andrzejewski, X. Zhu, M. Craven, and B. Recht, "A framework for incorporating general domain knowledge into latent dirichlet allocation using first-order logic," in *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume Two*, ser. IJCAI'11. AAAI Press, 2011, pp. 1171–1177. [Online]. Available: http://dx.doi.org/10.5591/978-1-57735-516-8/IJCAI11-200

[13] J. Boyd-Graber, J. Chang, S. Gerrish, C. Wang, and D. Blei, "Reading tea leaves: How humans interpret topic models," in *Neural Information Processing Systems (NIPS)*, 2009.

[14] J. Chuang, C. D. Manning, and J. Heer, "Termite: Visualization techniques for assessing textual topic models," in *Proceedings of the International Working Conference on Advanced Visual Interfaces*, ser. AVI '12. New York, NY, USA: ACM, 2012, pp. 74–77. [Online]. Available: http://doi.acm.org/10.1145/2254556.2254572

[15] M. J. Gardner, J. Lutes, J. Lund, and J. Hansen, "The topic browser: An interactive tool for browsing topic models," 2010.

[16] B. Gretarsson, J. O'Donovan, S. Bostandjiev, T. Höllerer, A. Asuncion, D. Newman, and P. Smyth, "Topicnets: Visual analysis of large text corpora with topic modeling," *ACM Trans. Intell. Syst. Technol.*, vol. 3, no. 2, pp. 23:1–23:26, Feb. 2012. [Online]. Available: http://doi.acm.org/10.1145/2089094.2089099

[17] W. Dou, X. Wang, R. Chang, and W. Ribarsky, "Paralleltopics: A probabilistic approach to exploring document collections," in *Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on*, Oct 2011, pp. 231–240.

[18] C. Sievert and K. E. Shirley, "Ldavis: A method for visualizing and interpreting topics," in *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, 2014. [Online]. Available: https://github.com/cpsievert/LDAvis

[19] A. Ganesan, K. Brantley, S. Pan, and J. Chen, "Ldaexplore: Visualizing topic models generated using latent dirichlet allocation," *CoRR*, vol. abs/1507.06593, 2015. [Online]. Available: http://arxiv.org/abs/1507.06593

[20] S. Liu, X. Wang, J. Chen, J. Zhu, and B. Guo, "Topicpanorama: A full picture of relevant topics," in *Visual Analytics Science and Technology (VAST), 2014 IEEE Conference on*, Oct 2014, pp. 183–192.

[21] A. Smith, T. Hawes, and M. Myers, "Hiearchie: Visualization for hierarchical topic models," in *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*. Baltimore, Maryland, USA: Association for Computational Linguistics, June 2014, pp. 71–78. [Online]. Available: http://www.aclweb.org/anthology/W/W14/W14-3111

[22] W. Cui, S. Liu, L. Tan, C. Shi, Y. Song, Z. Gao, H. Qu, and X. Tong, "Textflow: Towards better understanding of evolving topics in text," *IEEE Transactions on Visualization and Computer Graphics*, vol. 17, no. 12, pp. 2412–2421, Dec 2011.

[23] S. Liu, M. X. Zhou, S. Pan, W. Qian, W. Cai, and X. Lian, "Interactive, topic-based visual text summarization and analysis," in *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, ser. CIKM '09. New York, NY, USA: ACM, 2009, pp. 543–552. [Online]. Available: http://doi.acm.org/10.1145/1645953.1646023

[24] S. Havre, E. Hetzler, P. Whitney, and L. Nowell, "Themeriver: visualizing thematic changes in large document collections," *IEEE Transactions on Visualization and Computer Graphics*, vol. 8, no. 1, pp. 9–20, Jan 2002.

[25] W. Cui, S. Liu, Z. Wu, and H. Wei, "How hierarchical topics evolve in large text corpora," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 12, pp. 2281–2290, Dec 2014.

[26] D. M. Blei, T. L. Griffiths, M. I. Jordan, and J. B. Tenenbaum, "Hierarchical topic models and the nested chinese restaurant process," in *Advances in Neural Information Processing Systems*. MIT Press, 2004.

Answer Extraction System Based on Latent Dirichlet Allocation

Mohammed A. S. Ali[1], Sherif M. Abdou[2]
Information Technology Department
Faculty of Computers and Iinformation,Cairo University
Cairo, Egypt

***Abstract*—Question Answering (QA) task is still an active area of research in information retrieval. A variety of methods which have been proposed in the literature during the last few decades to solve this task have achieved mixed success. However, such methods developed in the Arabic language are scarce and do not have a good performance record. This is due to the challenges of Arabic language. QA based on Frequently Asked Questions is an important branch of QA in which a question is answered based on pre-answered ones. In this paper, the aim is to build a question answering system that responds to a user inquiry based on pre-answered questions. The proposed approach is based on Latent Dirichlet Allocation. Firstly, the dataset, pairs of questions and associated answers, will be grouped into several clusters of related documents. Next, when a new question to be answered is posed to the system, it,therefore, starts to assign this question to its appropriate cluster, then, use a similarity measure to get the top ten closest possible answers. Preliminary results show that the proposed method is achieving a good level of performance.**

Keywords*—Question Answering; frequently asked questions; information retrieval; artificial intelligence;

I. Introduction

Question Answering (QA) is a task which has been created to satisfy the specific and urgent need of a user to get a direct answer to a given question. Generally speaking, QA tasks can be classified, from an information retrieval perspective, into two separated types: question answering based on retrieving and forming an answer from flat documents, and the second one is based on retrieving an answer to a similar pre-answered question. Both types are considered active research topics in information retrieval. However, this paper is concerned only with QA based on the Frequently Asked Questions (FAQ) task.

According to the literature, great efforts have been made to build a reliable QA system. However, few of these attempts have been made for the Arabic language. And among those only a few of them are oriented to QA based on a FAQ task. This lack of such systems is due to challenges presented by the Arabic language.

Arabic is a Semitic language spoken as a native language by more than 330 million people [1]. Arabic is a morphologically complex, highly derivational and inflectional language. Moreover, Arabic is rich in the use of affixes and clitics and, usually, disambiguating short vowels and other orthographic diacritics in standard orthography are omitted [2]. Therefore, it has been difficult, to some extent, to build a reliable QA system.

In this paper, a system for QA based on Latent Dirichlet Allocation (LDA) [3] has been presented. The LDA has been exploited, as a clustering algorithm, to divide the dataset into related document groups. Then, its estimated models parameters has been also exploited to calculate the similarity between the new question and each question-answer pair in its closest group [4].

The domain in which this application will be applied is Islamic Fatwa. A Fatwa is a formal Islamic legal opinion issued by expert scholar(s) (mufti or committee) in response to a question from an individual. In Fatwa, mufti clarifies an issue based on evidence from Shariah [5]. The Fatwa is considered as an Islamic religion verdict, therefore, Muslims all over the globe are interested in them and seek them out on a daily basis. Moreover, the field is very sensitive, so, mistakes are not allowed. The official Fatwa organizations are responsible for receiving, handling and replaying these daily questions.

Due to the limitation of human resources, these organizations are unable to handle this barrage of questions within a reasonable time frame. Meanwhile, many newly posed questions have similar answered ones in the database. Unfortunately, there are no effective and reliable systems yet built to automatically retrieve such a type of questions.

A. Previous work

Several pieces of research have been proposed in the literature in the field of question answering based on already pre-answered ones.

In [6] R. D. Buke el.al. have proposed a system to fetch a similar question to a newly posed question. This system was called FAQFinder. Their system is based on a vector space model, and included a semantic definition of similarity between words based on the concept of hierarchy in WordNet as well.

Keliang Jia et.al. in [7] have built a QA system for remote learning applications, so as to enhance the communication facilities between teachers and their students. They calculated the similarity between questions by integrating both similarity between domain keywords using a domain knowledge dictionary and similarity between common words using HowNet.

Zhiguo Wang et.al in [8] have tried to address the issue of FAQ-based QA via word alignment. They started with extracting a feature vector, including (similarity, dispersion, penalty, 5 important words, reverse and some spare lexical

features), from pair (query, candidate), then used a neural network to calculate the similarity between such a pair.

None of the previously mentioned works is concerned with Arabic language. However, in [9], [10] Islam Elhelwany et.al. have proposed an Arabic Fatwa Intelligent system based on textual case based rezoning which was firstly used in [11]. In their system, they started by extracting a representative term for each cluster which were later called clusters attractors. Then, the cases clustered around these attractors. Eventually, they used Jensen-Shannon divergence to assign a newly posed question to its appropriate cluster and, subsequently, to find the closest possible question among questions in such a cluster. Unfortunately, no results or evaluation are presented in these works and the data sets are not available for comparison. In general, none of the existing works efficiently addresses the task of Arabic QA based on FAQ which is going to be address in this work.

The rest of this paper is organized as follows: Section 2 introduces the proposed method; the evaluation and experimental results are discussed in Section 3; and finally, in Section 4, our findings are summarized and some future work is propose.

II. Approach

A. Latent Dirichlet Allocation (LDA) model estimation

Latent Dirichlet Allocation (LDA) is an unsupervised, statistical approach for document modeling that discovers latent topics in a collections of text documents, in this case each document is Fatwa (question and answer). LDA considers a word as a basic unit of information, and it assumes that documents that discuss similar topics use a similar collection of words. In other words, documents are modeled as distribution of topics (θ), and each topic is modeled as a distribution of words (ϕ). topics are thus discovered by recognizing collections of words which frequently occur together within documents [3]. In figure 1 a graphical representation of LDA is shown. As depicted in the figure [2] M is the number of documents of arbitrary length in the collection, T topics and V words forming the vocabulary. Here, the topic distribution per document and the per-topic word distributions are sampled from $Dir(\alpha)$ and $Dir(\beta)$ respectively. The LDA model estimation goes through these steps:

1) Choose number of topics T and LDA hyperparamters α and β.
2) For each document
   a) Choose the number of words N.
   b) For each word:
      i) Sample z from $\theta^{(j)}$, where j is the index of the current document.
      ii) Sample w from $\phi^{(z)}$

In LDA the goal is to estimate the distribution $p(z/w)$. Unfortunately, exact estimation of LDA parameters is an intractable problem. The solution to this problem is to use an approximation estimation algorithm; common methods to do so include Expectation propagation and Gibbs sampling [12] , which is more common and is followed here.

We will present only the most important equation used by the algorithm for topic sampling for words. Let $\vec{w}$ and $\vec{z}$ be the vectors of all words and their topic assignment of the whole documents collection W respectively. The topic assignment for a particular word depends on the current topic assignment of all the other word positions. More specifically, the topic assignment of a particular word t is sampled from the following distribution:

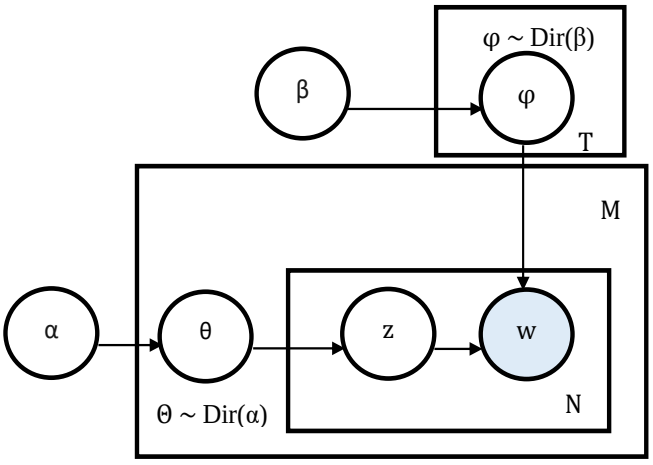


Fig. 1: LDAs graphical representation [3] shaded nodes represent observed variables whereas other nodes represent latent ones.

$$p(z_i = k \mid \vec{z_{\neg i}}, \vec{w}) = \frac{n^{(t)}_{k,\neg i} + \beta_t}{[\sum_{v=1}^{V} n^{(v)}_k + \beta_v] - 1} \cdot \frac{n^{(k)}_{m,\neg i} + \alpha_k}{[\sum_{j=1}^{K} n^{(j)}_m + \alpha_j] - 1}$$

where $n^{(t)}_{k,\neg i}$ is the number of times the word t is assigned to topic k except the current assignment; $\sum_{v=1}^{V} n^{(v)}_k - 1$ is the total number of words assigned to topic k except the current assignment;$n^{(k)}_{m,\neg i}$ is the number of words in document m assigned to topic k except the current assignment; and $\sum_{j=1}^{K} n^{(j)}_m - 1$ is the total number of words in document m except the current word t. Here, α and β are hyperparamters of LDA and they describe the nature of the priors of θ and ϕ respectively.The choice of priors has an important implication for the result. For instance, choose high value for β can be expected to decrease the number of topics, whereas smaller β values will generate more topics [13].

Once p(z/w) is estimated using a sufficient number of Gibbs sampling iteration, the distributions ϕ and θ can be easily estimated using the following formulas:

$$\phi_{k,t} = \frac{n^{(t)}_k + \beta_t}{\sum_{v=1}^{V} n^{(v)}_k + \beta_v}$$

$$\theta_{m,k} = \frac{n^{(k)}_m + \alpha_k}{\sum_{j=1}^{K} n^{(j)}_m + \alpha_j}$$

once the LDA model is estimated, that is to say that the corpora have been clustered and the estimated model can be used to infer a new document as well.

B. Semantic answer extraction using LDA estimated model

A similarity between two documents d1 and d2 is computed by multiplying both, the similarities between the topic distribution per-document ($\theta\ d1$ and $\theta\ d2$) and the per-topic word distributions ($\phi\ t1$ and $\phi\ t2$) [4]

The similarities between the topic distribution per-document can be calculated using the following equations:

$$IR(p,q) = \sum_{i=1}^{T} p_i\ log\frac{2xp_i}{p_i+q_i} + \sum_{i=1}^{T} q_i\ log\frac{2xq_i}{p_i+q_i} \quad (1)$$

such that p and q are the document distribution over topics for $d1$ and $d2$ respectively.

As IR measures the distance between two documents, it can be transformed into similarity measure using the following equation:

$$SIM(p,q) = 10^{-\delta IR(p,q)} \quad (2)$$

Meanwhile, each word has a certain contribution to a topic (word distributions per-topic $\phi\ t1$ and $\phi\ t2$). Based on these contributions, word-to-word semantic similarity is defined. The word-to-word semantic similarity measure based on LDA is further used in conjunction with an optimal matching method to measure similarity given two documents. The similarities between word distributions per-topic can be considered as an assignment problem. Given a complete bipartite graph, $G = (D1, D2, E)$, with n document 1 vertices (words) $(D1)$, n document 2 vertices (words) $(D2)$, and each edge $e_{d1\in D1, d2\in D2} \in E$ a non-negative weight (similarity between the two words). The aim to find matching M from $D1$ to $D2$ with maximum similarity. Such an assignment is called optimum assignment. Method in [14] is used to solve this assignment and can be formulated as finding a permutation π for which $q_{OPT} = \sum_{i=1}^{n} word\text{-}sim(d1_i, d2_{\pi(i)})$ is maximum such that $word\text{-}sim$ is word-to-word similarity measure based on LDA and can be calculated using the following equation (Hellinger distance)

$$HD(w1,w2) = \frac{1}{\sqrt{2}} + \sqrt{\sum_{1}^{T} (\sqrt{w1_i} - \sqrt{w2_i})^2} \quad (3)$$

Such that $w1$ and $w2$ are the word distributions per-topic for $d1$ and $d2$ respectively.

Briefly, the proposed method can be summarized in these steps:

1) Performed data pre-processing.
2) Estimate LDA model for the collection of documents.
3) Cluster the document collection based on estimated LDA model.
4) When a new question is posed, assign it to its appropriate cluster using LDA inferencer.
5) Once the new question is assigned to its cluster, retrieve the ten closest answers possible using measures mentioned in section II-B
6) Display results

III. Results and discussion

The dataset that has been used was collected from the well-known website that introduces Islamic Fatwas "IslamWeb"[1]. The total number of documents is 11109. Each one of these documents represents a Fatwa which contains a question and associated answer. All documents in this collection are used to estimate LDA model in step II-A described in the methodology. For the test, 110 non-answered questions were posed to the system and the result obtained shown to seven educated users. The users were then asked to tell how much they agreed with the following statement: "this answer fits my question and I am satisfied with it". The users rated their degree of agreement on a 5-point Likert scale where 1 indicates strong disagreement and 5 indicates strong agreement.

It should be noticed that all results, of the proposed method, presented in this section are based on the following parameters which have been experimentally set: the Dirichlet hyper-parameters α and β were chosen to be 0.5 and 0.1 respectively and the number of topics was chosen to be 100. Gibbs sampling is stopped after 1000 steps.

It is difficult to compare the proposed approach and the various approaches described in Section I-A because 'the software applications and the textual resources used in the experiments are unavailable'. Moreover, the results of the respective experiments are not conclusive.

For example, the work presented by Islam et.al [9] does not measure the effectiveness of the presented approach. Moreover, in another work presented by the same authors [10], the only results shown are clustering results. However, neither the software applications nor the textual resources used in the experiments are available for comparison.

The rest of the works are oriented to other languages but not Arabic. Therefore, the effectiveness of the proposed approach will be evaluated by comparison with Google search engine, where the top ten retrieved results are collected manually and compared to the top ten retrieved results by the proposed method as shown in section III.

To estimate the performance, the average Likert scale and average retrieval time are calculated. The average Likert scale is defined as follows: let U and Q be the total number of users and total number of questions respectively. $LS = \frac{1}{Q\times U}\sum_{n=1}^{Q}\sum_{m=1}^{U} S_{n,m}$ such that $S_{n,m}$ is a score given by a user n to a question m. To test the inter-rater reliability, the Kappa measure has been calculated. As shown in the table I, the performance of the proposed system is better. This success is mitigated, though, by the fact that the Google average response time is better than ours by orders of magnitude. Nonetheless, it is commonly assumed that Kappa values between 0.4 and 0.6 offer a moderate level of agreement, and therefore, both of them, LDA-Optimal and Google get a moderate agreement. In Figure 2 a diverging stacked bar chart shows the raw results based on user evaluation of the proposed system and of the Google search engine. It presents the Likert scale results of the criteria "this answer fits my question and I am satisfied with it". As it can be seen in the figure, that number of answers with which the users 'strongly agree' and 'agree' in the proposed approach is

[1] www.islamweb.net/fatwa/

TABLE I: average Likert scale, average response time and Kappa measures of LDA-Optimal and google

Method	Average 5-point Likert scale	Average retrieval time (Second)	Kappa
Google	2.65	**0.85**	**0.58**
LDA-Optimal	**3.75**	22.4	0.55

clearly greater than those found through Google. Meanwhile, according to user evaluation, more than half of the questions are not answered through Google.

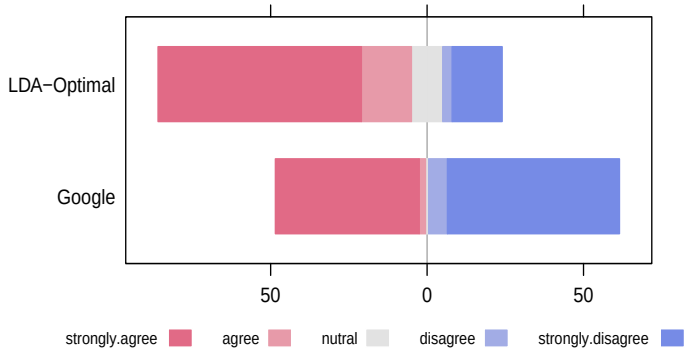


Fig. 2: Results of criteria this answer fits my question and I am satisfied with

After analysis of results for both methods, it has been found that, Google was able to handle those questions which contain only small number of words, 3-7 words, better than LDA-Optimal. On the other hand, the proposed approach has an ability to handle those questions with longer scripts, while Google has a lesser ability to do so and sometimes fails when the number of characters exceeds its limit.

IV. Conclusion and Future Work

With the boom of the Web's content, an inevitable need for an effective information retrieval system is required. In particular, the possibility of extracting a direct answer to a specific question. This process is called question answering and is currently one of the most active research areas in the field of information retrieval. The QA based on FAQ is the task in which a new question is answered based on already pre-answered ones.

In this paper, a new methodology is proposed to accomplish the task of QA based on FAQ. This approach assumes that an answer is a contextual expansion of its corresponding question. Therefore, the question and its associated answer is treated as one document. Since organization of documents into clusters of related documents has been shown to significantly improve the results of information retrieval systems, the approach first started to cluster the corpora into several clusters of related documents. Such clustering is achieved by the LDA model. When a new question to be answered is posed to the system, it is inferred, and assigned to an appropriate cluster using LDA inferencer.

Up to now, there is the question to be answered and its associated cluster. A similarity measure based on LDA estimated distributions is used to retrieve the closest possible answers to a given question.

In spite of all the advantages and possibilities of the proposed method, it has several limitations that could be improved in the future. First, the proposed approach does not consider the type of question, so future improvements to the accuracy of the system will involve a question analysis step so as to determine the type of question. Second, a different sophisticated similarity measure can be used instead of the current one. Finally, the current proposed approach does not handle negation, this may be dealt with in future researches.

References

[1] A. Farghaly and K. Shaalan, "Arabic natural language processing: Challenges and solutions," *ACM Transactions on Asian Language Information Processing (TALIP)*, vol. 8, no. 4, p. 14, 2009.

[2] N. Habash and O. Rambow, "Arabic tokenization, part-of-speech tagging and morphological disambiguation in one fell swoop," in *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ser. ACL '05. Stroudsburg, PA, USA: Association for Computational Linguistics, 2005, pp. 573–580.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.

[4] V. Rus, N. Niraula, and R. Banjade, *Computational Linguistics and Intelligent Text Processing: 14th International Conference, CICLing 2013, Samos, Greece, March 24-30, 2013, Proceedings, Part I*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, ch. Similarity Measures Based on Latent Dirichlet Allocation, pp. 459–470.

[5] J. Esposito and A. Sachedina, *The Islamic World: Hizbullah-Ottomon empire*, ser. The Islamic World. Oxford University Press, 2004. [Online]. Available: https://books.google.com.eg/books?id=GSUZAQAAIAAJ

[6] R. D. Burke, K. J. Hammond, V. Kulyukin, S. L. Lytinen, N. Tomuro, and S. Schoenberg, "Question answering from frequently asked question files: Experiences with the faq finder system," *AI magazine*, vol. 18, no. 2, p. 57, 1997.

[7] K. Jia, X. Pang, and Z. Li, "Question answering system in network education based on faq," in *Young Computer Scientists, 2008. ICYCS 2008. The 9th International Conference for*, Nov 2008, pp. 2577–2581.

[8] Z. Wang and A. Ittycheriah, "Faq-based question answering via word alignment," *CoRR*, vol. abs/1507.02628, 2015.

[9] I. Elhalwany, A. Mohammed, K. Wassif, and H. Hefny, "Using textual case-based reasoning in intelligent fatawa qa system," *International Arab Journal of Information Technology (IAJIT)*, vol. 12, no. 5, 2015.

[10] I. Elhalwany, A. Mohammed, K. T. Wassif, and H. A. Hefny, "Enhancements to knowledge discovery framework of {SOPHIA} textual case-based reasoning," *Egyptian Informatics Journal*, vol. 15, no. 3, pp. 211 – 220, 2014.

[11] D. Patterson, N. Rooney, M. Galushka, V. Dobrynin, and E. Smirnova, "Sophia-tcbr: A knowledge discovery framework for textual case-based reasoning," *Knowledge-Based Systems*, vol. 21, no. 5, pp. 404 – 414, 2008.

[12] X.-H. Phan, L.-M. Nguyen, and S. Horiguchi, "Learning to classify short and sparse text & web with hidden topics from large-scale data collections," in *Proceedings of the 17th International Conference on World Wide Web*, ser. WWW '08. New York, NY, USA: ACM, 2008, pp. 91–100.

[13] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proceedings of the National Academy of Sciences*, vol. 101, no. suppl 1, pp. 5228–5235, 2004.

[14] H. W. Kuhn, "The hungarian method for the assignment problem," *Naval Research Logistics Quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.

Computational Intelligence Optimization Algorithm Based on Meta-heuristic Social-Spider: Case Study on CT Liver Tumor Diagnosis

Mohamed Abu ElSoud
Faculty of Computers and Information,
Mansoura University, Egypt,
Tabouk University, Saudi Arabia.

Ahmed M. Anter
Faculty of Computers and Information,
Mansoura University, Egypt,
Benisuef University, Egypt
Scientific Research Group in Egypt (SRGE)

***Abstract*—Feature selection is an importance step in classification phase and directly affects the classification performance. Feature selection algorithm explores the data to eliminate noisy, redundant, irrelevant data, and optimize the classification performance. This paper addresses a new subset feature selection performed by a new Social Spider Optimizer algorithm (SSOA) to find optimal regions of the complex search space through the interaction of individuals in the population. SSOA is a new natural meta-heuristic computation algorithm which mimics the behavior of cooperative social-spiders based on the biological laws of the cooperative colony. Different combinatorial set of feature extraction is obtained from different methods in order to keep and achieve optimal accuracy. Normalization function is applied to smooth features between [0,1] and decrease gap between features. SSOA based on feature selection and reduction compared with other methods over CT liver tumor dataset, the proposed approach proves better performance in both feature size reduction and classification accuracy. Improvements are observed consistently among 4 classification methods. A theoretical analysis that models the number of correctly classified data is proposed using Confusion Matrix, Precision, Recall, and Accuracy. The achieved accuracy is 99.27%, precision is 99.37%, and recall is 99.19%. The results show that, the mechanism of SSOA provides very good exploration, exploitation and local minima avoidance.**

***Keywords*—Liver; CT; Social-Spider Optimization; Meta-heuristics; Support Vector Machine; Random Selection Features; Classification; Sequential Forward Floating Search; Optimization.**

I. Introduction

Liver cancer is a serious disease and it is the third commonest cancer followed by stomach and lung cancer [1]. As reported in [2], liver cancer in Egypt is continues to be the second highest cause of cancer incidence and mortality. The most effective way to reduce deaths due to liver cancer is to treat the disease in the early stages. Early treatment requires early diagnosis based on an accurate and reliable diagnostic procedure. One of the most common and robust imaging techniques for the detection of hepatic lesions is CT. The classification of benign and malignant patterns in CT is one of the most significant processes during the diagnosis of liver cancer. Computer aided liver diagnosis (CAD) is a technique that can help radiologists to accurately identify abnormality and help in reducing the risk of liver surgery [3].

The high volume of CT liver tumor data requires some helpful classification approaches to support the analysis of this data. Discriminate analysis is now widely used to distinguishing between normal and abnormal tumor tissues [4]. A critical issue in discriminate analysis is feature selection, instead of using all available features in the dataset, chooses only subset of features to be used in discriminate system [5], [6].

The feature selection stage is one of the important components in any classification system. The performance of a classifier depends directly on the choice of feature selection. The feature selection stage is designed to obtain a compact, relevant, non-redundant and meaningful representation of the data. These selected features are used by the classifier to classify features. It is assumed that, a classifiers that uses smaller and relevant features need less memory, computation speed, and prediction accuracy increased which is desirable for any real-time system [7]. When selecting a small subset of features, their biological relationship with the target tumor is more easily identified. Selecting an effective and more representative feature set is the objective of this paper.

Generally, feature reduction techniques can be divided into two main approaches: filter approach and wrapper approach. Filter approaches are not dependent on machine learning techniques and they are computationally inexpensive and it is more common than wrapper approach. Wrapper approach contains a machine learning techniques as part of the evaluation function, when combine filter approach with wrapper approach usually gives best results [8], [9].

The main challenging problem in feature selection and reduction is the huge search space. The size of the search space increase respect to the available number of features in the dataset. Thus, an exhaustive search is impossible in most cases. Many different search methods have been used to feature selection such as mutual information (MI) [10], document frequency (DF) [11], information gain (IG) [12]],Relief [11], [12], Sequential Backward Search (SBS) [12], Sequential Forward Search (SFS) [13], Sequential Forward Floating Search (SFFS) [13], and Random Selection Features (RSFS) [13], almost all these techniques are suffer from the stuck in local minima and computationally expensive. In order to further improve effect of feature selection global search algorithm is needed, many researches try to add intelligent optimization

algorithms into feature selection method, such as improved Gray wolf optimization algorithm [14] and genetic algorithms [15]. The current research work is focused on the determination of an optimal subset feature selection from CT liver tumor dataset using a new intelligence swarm model called social spider optimization algorithm (SSOA) in order to improve the diagnosis accuracy. The choice is a trade-off between computational time and quality of the generated feature subset solutions.

Swarm intelligence is a research field that models the collective behavior in swarms of insects or animals. Several algorithms inspired from the insects and animals behavior to solve a wide range of complex optimization problems [15]. The SSOA algorithm is based on the simulation of cooperative behavior of social-spiders is proposed to optimize our problem. In a social-spider colony, each member depending on its gender and executes a variety of tasks such as ferocity, mating, web design, and social interaction. The communal web is important part of the colony because it is a communication channel among them [16], [17], [18].

SSOA differ to other Evolutionary algorithms (EA). SSOA has a strong capability to search in the problem space and can efficiently find minimal reductions. This algorithm considers two different search agents (spiders): male and female. Depending on gender, each individual is conducted by a set of different evolutionary operators which mimic different cooperative behaviors within the colony. Depending on gender computational mechanisms are applied to avoid the critical flaws such as the premature convergence and the incorrect exploration-exploitation balance. The individuals who have achieved efficient exploration (female spiders) and individuals that verify extensive exploitation (male spiders) [15].

Though the studies mentioned above have contributed extremely to our understanding of the severity of the liver cancer problem, they are lacking to quantitative system to diagnosis these patients. Therefore, the main objective of this study is to develop computerized image analysis system to assist radiologists in interpretation of liver tumor. Multi-classifiers are used in conducting the liver tumor diagnostic problem. A new feature reduction and subset selection approach is used based on natural meta-heuristic model SSOA. Compared with other feature selection methods. SSOA yields more efficient results than any of the other methods tested in this paper.

The reminder of this paper is ordered as follows. Section II discusses the related work for liver tumor characterization. Details of the proposed swarming SSOA model and Texture feature extraction method based on fractal dimension are presented in Section III. In Section IV, the proposed liver tumor diagnosis approach is presented. Section V shows the dataset used and experimental results with discussion. Finally, Conclusion and future work are discussed in the end of this paper.

II. Related works

Computer Aided Diagnosis (CAD) plays a key role in the early detection and diagnosis of liver cancer. CAD system is a set of automatic or semi-automatic tools developed to assist radiologists in the diagnosis of liver tumor. Some of the recent classification results obtained by other studies for liver disease dataset are presented below:

Gletsos et al. [19], proposed first order statistics, SGLDM, gray level difference method, Laws' texture energy features, and fractal dimension measurements methods to extract features from liver tumors. Feature reduction is applied using Genetic Algorithm. Neural Network is applied for classification. Classification performance achieved 91%. Cavouras et al. [20], calculate twenty textural features from the CT density matrix of 20 hemangiomas (benign) and 36 liver metastases (malignant) and were used to train a multilayer perception neural network classifier and four statistical classifiers are used. The performance achieved 83%.

Chen et al. [21], the neural network is included to classify liver tumors. It is implemented by a modified probabilistic neural network (PNN) [MPNN] in conjunction with feature descriptors which are generated by fractal feature information and the gray-level co-occurrence matrix. NFB feature values, spatial gray level dependence matrices give better performance. It is texture based. 30 patients (20 malignant, 10 benign). Classification rate is 83%.

Mougiakakou et al. [22], proposed for each ROI, five distinct sets of texture features to characterize liver based on FOS, SGLDM, GLDM, TEM, and FDM. The genetic algorithm-based feature selection is applied to reduct features. The fused feature set was obtained after feature selection applied. 97 samples is used (38 healthy and 59 abnormal). Weighted voting scheme for 5 classifiers is used. The best performance achieved is 84%.

Kumar et al. [23], proposed Wavelet and Fast Discrete Curvelet Transform (FDHCC) for feature extraction, and to distinguish between benign and malignant tumors the Feed Forward Neural Network classifier is used. The accuracy achieved for Curvelet Transform is 93.3%, and Wavelet is 88.9%.

Duda et al. [24], proposed approach to texture characterization from dynamic CT scans of the liver. The methods applied to recognizing features from hepatic primary tumors are RLM 8 features, COM 11 features, and entropy of image after filtering it with 14 features Laws filters. Experiments with various sets of texture parameters show that the classification accuracy was greater than 90% using Support Vector Machines.

Kumar et al. [25], Improved his work by apply texture features using Gray-Level first-order statistics (GLFOS), Gray level co-occurrence matrix, Contour let coefficient first-order statistics (CCFOS), Contour let coefficient co- occurrence matrices (CCCMs) and for feature selection applied PCA. The classification accuracy based on PNN to classify liver tumor into HCC and Hemangioma. The results obtained from this CAD system for FOS, GLCM, CCFOS, and CCCM are 79%, 86%, 93%, 94% respectively with total accuracy 88%.

The accuracy obtained from above researches are very low and computationally expensive. Therefore, the intelligent optimizations are needed to increase the efficiency and reduce the computation of the methods used. Some of the recently authors are working to optimize hard problems using Social Spider Optimization in different application such as:

James et al. [26], proposed framework based on the foraging strategy of social spiders. SSO can tackle a wide range of different continuous optimization problems and has the potential to be employed to solve real world problems. A set of 20 benchmark functions were used to evaluate the performance of SSO which cover a large variety of different optimization problem types. SSO compared with some widely computational intelligence. Results indicate the effectiveness and robustness of the proposed algorithm to solve optimal hard problem.

Djemame et al. [27], proposed approach to improve segmentation process based on Social Spider optimization. The spiders seem sensitive to the topology of the image, so it is possible to guide spiders movement with a gradient or a laplacian. Indeed, these measures will provide information on the possible presence of contours. It would be then possible to use spiders in two ways: 1) Gradient would be repellent which would partition a colony of spiders in a region. 2) On the contrary, the gradient could have an attractive effect. In this case, spiders would be used to detect the contours of regions.

Pazhaniraja et al. [28], proposed a novel scheme of Discovering new services using SSO. The services that get scattered in the UDDI registry can be discovered by using SSO technique. The SSO method can be used to retrieve more appropriate service from number of services. The proposed approach used to embed the bio-inspired algorithm (SSO) into the web service.The result achieved dynamic web service response to the service requester.

Lenin et al. [29], proposed an improved spider algorithm (ISA) to solve the optimal reactive power dispatch (ORPD) Problem. The structure is based on the foraging social spiders, which make use of the vibrations spread over the communal web to decide the position of preys. The simulation results demonstrate high quality performance of ISA in solving an optimal reactive power dispatch problem. Results indicate the effectiveness and robustness of the proposed algorithm to solve optimal reactive power dispatch problem.

Computational Intelligence based on Bio-inspired SSOA algorithm is used in this paper to decreases time consuming, extract and select relevant, optimal and few features from a huge number of features which are sufficient. Computations also are reduced while prediction accuracy is increased via effective feature selection.

III. METHODOLOGY

1) Swarming Model based on Bio-inspired Social-Spider Optimization algorithm (SSOA): Social-Spider Optimization algorithm can be defined as population-based and algorithmic search meta-heuristic methods that mimic natural evolution process of social spider colony for brief description and more details in [30], [31], [32], [33].

A majority of the spiders are solitary which means that they spend most of their lives without interacting with others. Among the 35 000 spider species observed and described by scientists, some species are social. These spiders live in groups. Based on these social spiders, social spider optimization algorithm (SSOA) is developed to optimize the problems [30], [31]. There are two fundamental components of a social spider colony, social members and communal web. The social members are divided into males and females. Each spider in the problem represents the solution. Each attribute of features distributed randomly to these spiders. The number of females N is randomly selected within the range of 65% - 90% and calculated by the following equation:

$$N_f = floor[(0.9 - rand\,(0,1)\,.0.25)\,.N] \tag{1}$$

Where S is a population size, and N is number of spider positions (solution). The population contains of females f_i and males m_i. The number of male spiders N_m is calculated as follows:

$$N_m = N - \ N_f \tag{2}$$

Generate females and males positions randomly on dimension space. The position for female spider calculated as follow:

Generate females and males positions randomly on dimension space. The position for female spider calculated as follow:

$$f_{i,j}^0 = P_j^{low} + rand\,(0,1)\,.\left(P_j^{high} - \ P_j^{low}\right) \tag{3}$$

$$Where\ i = 1, 2, \ldots\ldots,\ N_f; j = 1, 2,\ \ldots.., n$$

Where $f_i$is the female spider position, p^{low} lower initial parameter bound and p^{high} upper initial parameter bound.

The position for male spider m_i is generated randomly as follow:

$$m_{i,j}^0 = P_j^{low} + rand\,(0,1)\,.\left(P_j^{high} - \ P_j^{low}\right) \tag{4}$$

$$Where\ i = 1, 2, \ldots\ldots,\ N_m; j = 1, 2,\ \ldots.., n$$

The evaluations of females and males spiders are defined and weights assigned to each spider. The weighted function for each spider which represents the solution is calculated as follow:

$$w_i = \frac{J\,(s_i) - \ worst_s}{best_s - \ worst_s} \tag{5}$$

Where $J(s_i)$ is the fitness value obtained of the spider position s_i, the values of worst and bests are the maximum and minimum values of the solution in the population respectively. In SSO, the communal web represents the dimension of search space. The search space of the optimization problem seen as a hyper-dimensional spider web. Each solution within the search space represents a spider position. The weight of each spider represents the fitness value of the solution [32]. The information among the colony members is transmitted through the communal web and encoded as a small vibrations. The vibrations depend on the weight and distance of the spider which has generated them [30]. The information transmitted (vibrations) perceived by the individual i from member j are modeled as follow:

$$V_i b_{i,j} = w_j . e^{-d_{i,j}^2} \tag{6}$$

Where the d_{ij} is the Euclidean distance between the spiders *i* and *j*.

In each iterations. Female spider presents an attraction or dislike to other spiders according to their vibrations based on the weight and distance of the spiders. Female spiders start looking for any stronger vibration. If there's someone more attractive, the Euclidean distance is calculated. Then the shortest distance between around spiders are calculated and index for the shortest distance. Then female spider do movement and an attraction based on the strong vibration and distance [30], [31]. If r_m is smaller than a threshold *PF*, an attraction movement is generated; otherwise, a dislike movement is produced as follows.

$$f_i^{t+1} = \begin{cases} f_i^t + \alpha.V_i b_{ci}.(S_c - f_i^t) + \beta.V_i b_{bi}.(S_b - f_i^t) \\ +gamma.(rand - 0.5)\, with\ probability\ PF \\ f_i^t - \alpha.V_i b_{ci}.(S_c - f_i^t) + \beta.V_i b_{bi}.(S_b - f_i^t) \\ +\gamma.(rand - 0.5)\, with\ probability\ 1 - PF \end{cases} \tag{7}$$

Where r_m is random number generated between [0 1], α,β,δ and rand are random numbers between [0, 1], PF threshold =.7 and sc and sb represent the nearest member to i that holds a higher weight and the best spider of the entire population.

Male spiders are divided into two classes, dominate and non-dominate male spiders. Dominant male spiders have weight value above the median value of the male population. Non-dominate male have weights under the median value [31]. The position of the male spider can be modeled as follows:

$$m_i^{t+1} = \begin{cases} m_i^t + \alpha.V_i b_{fi}.(S_f - m_i^t) + \delta.(rand - 0.5) \\ if\ W_{N_{f+i}} > W_{N_{f+m}} \\ m_i^t - \alpha.\left(\frac{\sum_{h=1}^{N_m} m_h^t.w_{N_{f+h}}}{\sum_{h=1}^{N_m} w_{N_{f+h}}} - m_i^t\right) \end{cases} \tag{8}$$

Where s_f represents the nearest female spider to the male spider i and *W* is the median weight of male spider population.

The mating in a social spider colony is performed by the dominant males and the female members. Only the Male spiders above median are mating. When a dominant male m_g spider locates a set of female members within a specific range r (range of mating), it mates and forming a new brood [30]. The mating operation calculated as follow:

$$r = \frac{\sum_{j=1}^{n}(P_j^{high} - P_j^{low})}{2.n} \tag{9}$$

Where n is the dimension of the problem, and ${l_j}^{high}$ and ${l_j}^{low}$ are the upper and lower bounds. Once the new spider is formed, it is compared to the worst spider of the colony. If the new spider is better, the worst spider is replaced by the new one. This process is iterated until get the best weighted for each spider and convergence to optimum solution. All weights above 50% will have value '1' that indicates the particular feature indexed by the position of the '1' is selected. If it is '0', the feature is not selected for evaluation process.

2) Texture Feature Extraction based Fractal Dimension: Geometric primitives that are self-similar and irregular in nature are termed as fractals. Fractal Geometry was introduced to the world of research in 1982 by Mandelbrot [34]. Liver tumor texture is a combination of repeated patterns with regular/irregular frequency. The tumor structure exhibit similar behavior, it has maximum disparity in intensity texture inside and along boundary which serves as a major problem in its segmentation and classification. Fractal dimension reflects the measure of complexity of a surface and the scaling properties of the fractal i.e. how its structure changes when it is magnified. Thus fractal dimension gives a measure of the irregularity of a structure. In fact, the concept of fractal dimension can be used in a large number of applications, such as shape analysis and image segmentation. Segmentation-based Fractal Texture Analysis (SFTA) algorithm consists of decomposing the image into a set of binary images from which the fractal dimensions are computed to describe segmented texture. In order to decompose the image, a new algorithm two-threshold binary decomposition (TTBD) is proposed [35]. Then SFTA feature vector is constructed as the resulting binary images size, mean gray level and bound arias fractal dimension. The fractal measurements are employed to describe the boundary complexity of objects and structures segmented in the input image using box counting algorithm.

IV. PROPOSED CT LIVER TUMOR DIAGNOSIS APPROACH

The proposed approach consists of five main phases to classify CT liver tumors into benign and malignant as shown in Figure 1.

Feature Extraction phase: in this phase feature extraction methods are used such as Gray Level Co-occurrence Matrix (GLCM) [36], [37], First Order Statistics (FOs) [36], Local Binary Pattern (LBP) [37], SAFTA [35], and Feature Fusion to discriminate between benign and malignant tissues.

Normalization phase: the dataset will preprocessed using a normalization technique to decrease the gap between features and smooth data range between [0,1] to increase the classification rate.

Feature Selection phase: The main purpose of feature reduction is to determine a small set of features from a whole features in the problem. The features extracted have irrelevant, redundant, misleading, and noisy features. Remove these data that affects the prediction and classifiers accuracy can be useful. The proposed SSOA is based on K-nearest neighbor (KNN) as fitness function for selecting the optimal feature set as seen in Figure 2. The principles of social spider optimization is used for the optimal feature selection problem. Eventually, they should converge to optimal solution. The solution represents all possible features. Each feature can be seen as a position represented by male and female spider. The optimal position is the subset with high fitness and highest classification accuracy.

The SSOA makes iterations of exploration using female spider for new regions in the feature space and exploitation

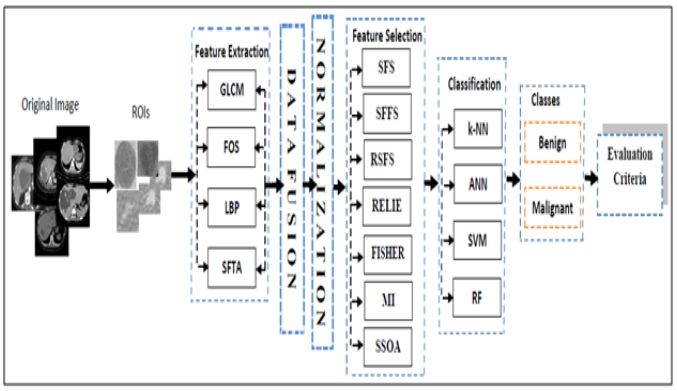


Fig. 1: The proposed liver tumor diagnosis approach.

using male spider until reaching near-optimal solution. To decide if a feature will be selected or not, constant threshold is used. All selected features weights above 50% will have value '1' that indicates the particular feature indexed by the position of the '1' is selected. If it is '0', the feature is not selected for evaluation. The best features with high weight and fitness are selected for evaluation and classification using KNN. The best subset features with high classification accuracy are indexed for classification system. Two fitness functions are used, which are weighted function to measure the weights for each spider in each iteration and changed till reaching the satisfactory solution, and KNN resembling the well-known forward selection. Algorithm 1 shows the steps of the proposed approach for feature subset selection using SSOA and Figure 2, shows the visual representation of the main steps of the proposed system based on bio-inspired SSOA for liver tumor diagnosis.

Classification phase: in this phase the classifiers K-Nearest Neighbor (KNN) [38], Artificial Neural Network (ANN) [39], Support Vector Machine (SVM) [40] and Decision Tree Classifier (DT) [41] are used to classify abnormality into two classes Benign and Malignant tumors. Multi-classifier system are used to obtain high accuracy and to increase the efficiency of our proposed system.

Analysis and evaluation: evaluation criteria for classifiers performance are calculated using confusion matrix, ROC, TP, FP, TN, FN, Precision, Recall, Accuracy, and Over-all accuracy [12], [42].

V. Experimental Results and Discussion

In this section, the performance of the proposed approach is presented. The simulations were programmed by Matlab 7.9 and performed on Intel (R) Core (TM) i7 CPU 2670QM-2.2 GHz and memory 8GB personal computer and a Microsoft Windows 7.

3) Data set collection: CT scanning is a diagnostic imaging procedure that uses X-rays in order to present cross-sectional images ("slices") of the body. The proposed CAD system will be work on difficult dataset. The dataset divided into benign and malignant categories depend on tumor type. The expert

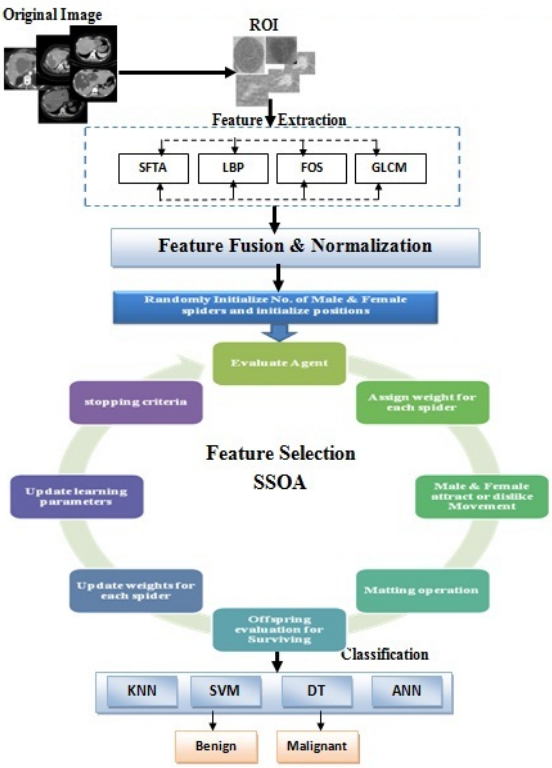


Fig. 2: The steps of the proposed bio-inspired social-spider optimization algorithm for subset feature selection.

physician select from data set 482 region of interests (ROIs) represent benign cases, and 350 ROIs represent malignant cases. Each ROI has dimension size 64×64 pixels [43].

4) Experimental Results: In this paper we developed a new approach for liver tumor diagnosis based on meta-heuristic social spider optimizer algorithm to select optimal features with no noise and redundancy.

Firstly, texture features are extracted from each ROIs which represent tumor. The feature extraction methods GLCM, FOs, LBP, and SFTA are proposed. Gray level co-occurrence matrix is constructed to extract feature vector with 68 values which used to represent each ROI. Seventeen features used to represent abnormality Energy, Entropy, homogeneity, Contrast, Dissimilarity, Angular Second Moment (ASM), Correlation, Variance, Maximum Probability (MP), Cluster Tendency, Cluster Shade, Cluster prominence, Sum Average, Sum variance, Sum Entropy, Entropy Difference, and Difference Variance. Four directions (0, 45, 90, and 135) and one distance between pixels equal to 1 are used. LBP is constructed to extract feature vector using mean, variance, skewness and kurtosis, these 4 values used to represent each ROI. FOs is used to extract feature vector with 4 values to represent abnormality (mean, variance, skewness and kurtosis). SFTA is constructed to extract feature vector from liver abnormality with 36 values. The feature vector constructed from SFTA features. The vector corresponds to the number of binary images obtained by TTBD multiplied by three, and from each binary image the following measurements are computed: fractal dimension, mean gray level and size. Finally, the future fusion is applied between

Algorithm 1 Feature subset selection using SSOA.

Initialize spidn, itern, Dims /* Define spider no., Iter. no. & problem dim. */
Initialize pflow, pfhigh , Xd, Xu /* Lower & upper Female Percent & space dim. */
Intialize rm=0, PF=.7 /* Reset Random generator & initialize proper tuning */
Intialize Nf, Nm /* The population of females and males */
Do Equation (2), (3)
Initlize fefit, mafit, Fewei, mawei /* Initlize fitness and weigths Female & males spider */
/* Population Initialization */
Generate female and male positions.
According to Equations (4), (5)
Assign weight for each spider on the colony.
According to Equation (6)
Start iterations for each feature
Female and male spiders movement according to the vibrations (V) depend on the weight and distance of the spiders.
Female spider movement and an attraction or dislike occur according to stronger vibration
According to Equations (7), (8)
For male spider movement. Select male spiders above median
Start looking for nearest female with higher vibration and shortest distance.
According to Equations (7), (9)
Mating operation: occurs after check whether male spider is good or not (above median).
According to Equation (10)
If(Fe is nearest distance) /* Mating occur and produce new offspring's*/
If(offspring $>$ Worst spider) /*Eval. offspring and calc. worst spider*/
Delete(Worst spider)
Do steps 9-15 again and recalculate the weight for each spider and check again the best and worst fitness
End For //iteration
Calculate the accuracy for each feature from KNN classifier

these features to construct feature vector with 112 values used to represent each ROI. After texture features are extracted. The normalization technique are applied on feature values to enhance, smooth, reduce gape between features and return feature values between [0,1].

Feature subsets are selected with no noise and redundancy and dimensionality reduced using social spider optimization algorithm. A social-spider members maintain a set of complex cooperative behaviors. SSOA can be defined as population-based and algorithmic search meta-heuristic methods that mimic natural evolution process of social spider colony. Each spider in the colony executes a variety of tasks depending on its gender. The fitness's of the spiders are evaluated using a function commonly referred to objective function in Equation (5). The fitness function reports numerical values which are used in ranking the best spider in the population. The space of the problem solution represents all possible selections of features. Each feature subset values can be seen as a position in this space represented by male and female spider. The numbers of female and male spiders are determined from feature space. In this application number of spiders is equal to number of values in the feature vector which will be 112 spiders. Also the number of iterations are suggested from 5, 10, 50 and 100 iterations to check the SSOA performance and the convergence to optimal solution. Table I shows the initial setting parameters that used in SSOA.

SSOA starts iterations to search for optimal solution, in each one the spiders move to attract or dislike presented in Equation (7) and Equation (8). Each female checked all the spiders and start looking for any stronger vibration presented in Equation (6). These strong vibrations based on the spider characteristics fitness and weight presented in Equation (5). If there's someone more attractive, the Euclidean distance is calculated. Then the shortest distance between the around spiders are calculated and indexed. An attraction is done based on the strong vibration and distance coming from nearest spider. Also repulsion or dislike is done based the gender. The median of spiders is calculated, the male spiders above mean median is start looking for a female with short distance. The spiders below median, go to weighted mean. In mating operation, we check whether male spider is good or not (above median) to generate the offspring's. Only the male spiders above median are mating and the radio (range of mating) is calculated as presented in Equation (10). Then start looking, if there's a good female near. The mating occurs and produces new offspring's. Then the offspring is evaluated and worst spider on the colony calculated. If the fitness of the offspring is better than the worst spider. Then the worst spider removed from the colony. This operation is iterated based on mating and movement occurs till convergence to optimal solution.

After feature values are optimized and the best weights with better fitness are calculated. All weights above threshold 50% will have value '1' that indicates the feature is selected for evaluation and if it is '0' below threshold, the feature is not selected and discard to get the best features for evaluation. The best features with high weights and fitness are selected for evaluation and classification using KNN. The best subset features with high classification accuracy are indexed for classification system. Two fitness functions are used, which are weighted function to measure the weights for each spider in each iteration and changed till reaching the satisfactory solution, and KNN work as forward selection. Forward selection starts with an empty feature set and searches for a feature that achieves the highest classification performance. Then the classification accuracy for the next optimized feature is calculated. Subset Features are selected if achieves the highest improvement in classification accuracy. Then the unused features are removed from features array.

In classification phase the training and testing cases are randomly selected by our algorithm, the number of training cases is 70% and 30% for testing cases. Four classifiers are applied to check the accuracy of the features selection. K-nearest neighbor, Decision Tree, Support Vector Machine, and Artificial neural Network are applied. To prepare KNN we

used k=1, in ANN we applied feed-forward back-propagation network with 7 layers in hidden neurons, input neurons depend on the feature sub set selection algorithm, output neurons 2 classes benign and malignant, number of training epochs is 10000. In SVM, we used regularization parameter for weight lambda=1, and we used linear kernel. The DT classifier is much simpler and faster in comparison with the neural network classifier. Each classifier has pros and cons in term of time execution and accuracy for feature vectors as shown in Tables (II-V).

The selection of relevant features and eliminate irrelevant ones is a great problem before classification is applied to train dataset. The redundant and irrelevant features with noise decreases the classification accuracy and makes the computation very complex as shown in Table II.

The normalization approach is applied as shown in Table III, all feature values in the vectors are normalized between [0,1] to decrease gap between values and increase the classifier performance. The visual representations for over-all accuracy, precision, and recall obtained from classification algorithms for non-enhanced dataset compared with normalized dataset are shown in Figure 3.

In this paper an accuracy of classification from different feature reduction methods is applied on abdominal CT liver dataset. In Table IV many different search methods used for feature reduction and selection such as Sequential Forward Search (SFS), Sequential Forward Floating Search (SFFS), Random Feature selection (RFS), Mutual information (MI), Relief, Fisher. Figure 4, shows the precision, recall, and accuracy results of using feature fusion and subset feature selection extracted from classical feature selection methods on the liver tumor dataset.

From the results, all these techniques suffer and stuck in local optima and computationally expensive. For these reasons we improve the effect of feature selection using global intelligent optimization search algorithm SSOA.

SSOA always converge to the optimal or near optimal solution. SSOA model shows a good balance between exploration and exploitation, critical flaws are avoidance such as premature convergence and local minima avoidance. The individuals divided into different search based on gender. The female spiders achieved efficient exploration and male spiders verify extensive exploitation. This assists a meta-heuristic to explore the search space extensively. From the results SSOA can search in the feature space until the optimal solution is converge.

In Table V, shows the results of using SSOA for feature reduction and selection. The number of iterations in SSOA is decreased by normalization phase to 5 iterations, because all features values are smoothed between [0, 1]. Also the time is decreased in all classifiers algorithm, this is because the proposed SSOA algorithm reducted the features and extracted only the optimal ones which are 13 features from 114 features. In Figure 5 we can see that the SSOA gives high accuracy in all classifiers used to diagnosis benign and malignant tumor. SSOA compared with the selection features methods applied in this paper as shown in Figure 6. The results of our proposed approach high, excellent and near to optimum. The achieved accuracy is 99.27%, precision is 99.37%, and recall is 99.19%.

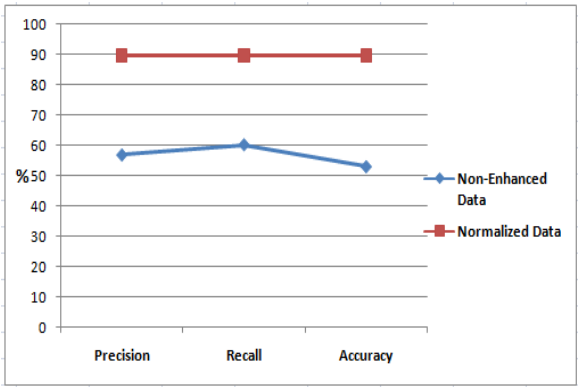


Fig. 3: Results of over-all precision, recall, and accuracy for non-enhanced features (irrelevant, redundant, and noise features) and normalized features.

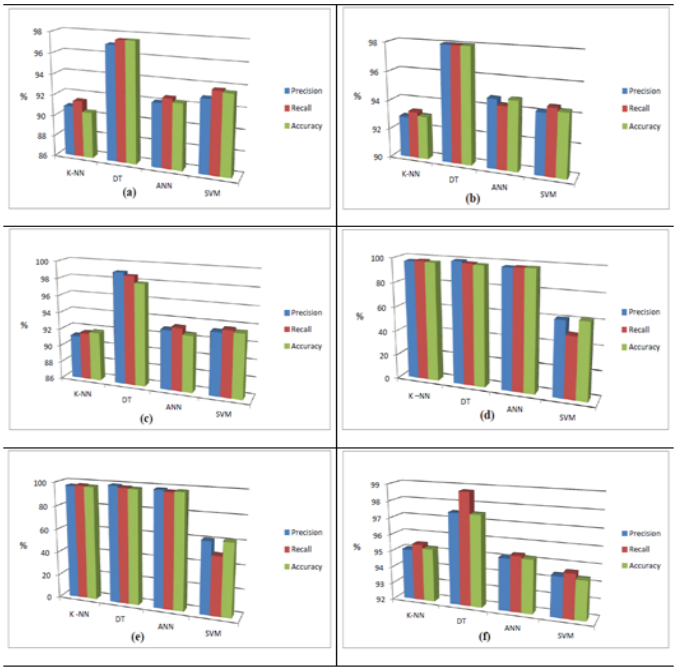


Fig. 4: Accuracy, precision, and recall results using feature fusion and classical feature selection methods on liver tumor dataset ((a) Fisher feature selection method, (b) Relief, (c) Mutual information, (d) SFS, (e) SFFS, (f) RSFS).

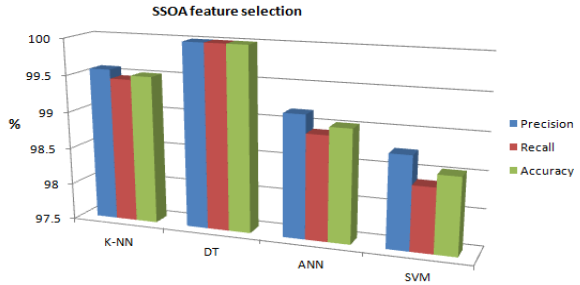


Fig. 5: Results of using the proposed social-spider optimization algorithm to select optimal features.

TABLE I: Initial parameters for Social Spider Optimization Algorithm.

Parameters	Description	Initial value
Spidn	No. of Spider	112
Itern	No. of Iterations	10, 50, 100
X_d	Lower space dim	-800
X_u	Upper space dim	800
${p_f}^{low}$	Lower Female Percent	0.65
${p_f}^{high}$	Upper Female Percent	0.9
Dims	Dimension Space	1
PF	Attraction or dislike	0.7
r_m	Random number	[0,1]
S_d	Search Domain	[0 1]

TABLE II: Classification results using Feature Fusion without normalization.

Classi.	Precision		Recall		Accuracy		Time/s
	B.	M.	B.	M.	B.	M.	
KNN	**58.04**	**47.06**	**64.84**	**40**	**61.25**	**43.24**	**0.51**
DT	**60**	**52.05**	**72.66**	**38**	**65.72**	**43.93**	**9.08**
ANN	**56.19**	**50**	**89.22**	**50**	**71.75**	**60**	**5.95**
SVM	**56.70**	**75**	**87.22**	**40**	**72.16**	**5.77**	**0.13**

TABLE III: Classification results using Feature Fusion with data normalization.

Classi.	Precision		Recall		Accuracy		Time/s
	B.	M.	B.	M.	B.	M.	
KNN	**92.29**	**86.28**	**87.84**	**92**	**88.87**	**86.89**	**2.31**
DT	**95**	**91.13**	**93**	**91**	**93**	**94**	**1.42**
ANN	**89.93**	**84.47**	**86.18**	**89**	**88.02**	**86.68**	**9.53**
SVM	**92.36**	**86.45**	**87.75**	**92**	**92**	**89.15**	**0.23**

TABLE IV: Classification results of using fused features with SFS, SFFS, RSFS, Relief, and Mutual Information.

Classi.	Precision		Recall		Accuracy		Time/s
	B.	M.	B.	M.	B.	M.	
***********Fusion*****RSFS*****Subsets(17)*******							
KNN	97	93.15	93.88	97	95.41	95.04	241.31
DT	97	98	97.5	100	97	98	1.14
ANN	96.21	94.06	94.66	96	95.42	95.02	1.95
SVM	97.5	91.34	92.31	97	94.6	94.09	0.29
***********Fusion*****SFS*****Subsets(7)*******							
KNN	97	97	97.5	98	97	97	27.12
DT	98.5	100	98	98	97.5	97	0.943
ANN	97	97	98	97	97	98	2.51
SVM	56.89	65	70	30	72.52	50.83	0.179
***********Fusion*****SFFS*****Subsets(7)*******							
KNN	97	97	97.5	98	97	97	27.23
DT	98.5	100	98	98	98	97	0.946
ANN	96.22	100	98	96	97.61	97.50	2.51
SVM	56.89	65	70	30	72.52	50.83	0.197
***********Fusion*****Relief*****Subsets(38)*******							
KNN	95.37	90.33	91.53	95	93.41	92.61	0.21
DT	98	98	98	98	98	98	1.185
ANN	95.43	94.03	93.66	95	95.04	94.51	4.64
SVM	96.19	92.19	93.09	96	94.62	94.06	0.21
***********Fusion*****Fisher*****Subsets(85)*******							
KNN	95.46	86.18	87.84	95	91.49	89.38	0.534
DT	96.5	97.56	97	98	97	98	1.39
ANN	95.52	88.65	90.19	95	92.78	91.72	6.867
SVM	96.17	89.67	90.97	96.5	93.97	93.19	0.24
***********Fusion*****MI*****Subsets(72)*******							
KNN	93.72	88.57	90.19	93	92.4	91.2	0.48
DT	98	100	100	97.34	97.80	98	1.33
ANN	95.56	90.38	91.75	95	93.62	91.63	4.77
SVM	95.39	91.23	92.31	95	93.83	93.08	0.23

TABLE V: Classification results using fused feature and subset feature selection using SSOA.

Feat. Selec.	Iter.	Fe.No.	Class.	Precision		Recall		Accuracy		Time/s
				B.	M.	B.	M.	B.	M.	
Fusion + SSOA	**5**	**13**	**KNN**	99.15	100	100	98.91	99.57	99.45	0.27
			DT	100	100	100	100	100	100	1.59
			ANN	98.32	100	100	97.83	99.15	98.90	1.49
			SVM	97.5	100	100	96.74	98.73	98.34	0.39

Fig. 6: Results of the proposed SSOA compared with classical feature selection.

A. Discussion

The proposed approach was tested and evaluated on difficult abdominal liver CT data set. The best overall accuracy obtained from Feature Fusion method and KNN classifier is 94%, and the accuracy obtained from other features extraction methods GLCM, LBP, and FD are 85%, 80%, and 87% respectively. To increase the performance of the proposed approach for liver tumor diagnosis. A new subset feature selection algorithm based on Meta-heuristic Bio-inspired Social Spider Optimization algorithm was proposed to select subset of relevant features and eliminate irrelevant ones. The solution space represents all possible selections of features. Each feature subset values can be seen as a position in such a space represented by male and female spider. We used GLCM, FOs, LBP, and SFTA for feature extraction phase. Normalization function is applied to enhance, smooth data and reduce gap between features. SSOA is used to select features with no-noise, no-redundancy and dimensionality reduction. The global intelligent optimization search for subset feature selection SSOA is compared with SFS, SFFS, RFS, MI, Relief, and Fisher. The best optimal features extrated from SSOA with high accuracy and less time consuming are 13 features and 38, 85, 72, 7, 7, 17 from Relief, Fisher, MI, SFS, SFFS, RSFS methods respectively. From results show that, all these techniques suffer from the issues of stuck in local optima and computationally expensive.

SSOA shows a good balance between exploration and exploitation and the results in high local minima avoidance. The female spiders achieved efficient exploration and male spiders verify extensive exploitation. This assists a meta-heuristic to explore the search space extensively. The mechanism of SSOA provides very good exploration, local minima avoidance, and exploitation simultaneously. The proposed approach is high,

TABLE VI: Comparison between the proposed approach and previous works on CT liver tumor diagnosis.

Authors	Year	Dataset	Accuracy
Cavouras et al. [20]	1996	56	83%
Chen et al. [21]	1998	30	83%
Gletsos et al. [19]	2003	147	91%
Bilello et al. [44]	2004	51	80%
Mougiakakou et al. [22]	2007	97	84%
Zhang et al. [45]	2008	44	97.7%
Proposed approach	**2015**	**832**	**99.27%**

excellent and near to optimum solution 99.27%, 99.37%, and 99.19% for accuracy, precision, recall respectively.

Comparing with the other previous works which diagnosis liver abnormality, the works of Cavouras et al. [20] and Chen et al. [21] reached 83% for liver abnormality classification without using feature reduction and selection methods. The results obtained with time computation cost and less accuracy, Gletsos et al. [19] used evolutionary genetic algorithm (GA) for feature reduction and achieved an overall correct classification of 91%. Mougiakakou et al. [22], used genetic algorithm-based feature selection to reduct features and achieved an overall correct classification of 84%. GA gives good results and converge to optimal solution, but has some problems such as pramature convergance, crossover, mutation and stuck in local minima. Kumar et al. [25], applied PCA for feature selection and achieved total accuracy 88%. A 99.27% of correct classification and perfect agreement were obtained in our experiments with large dataset as seen in Table VI.

VI. CONCLUSION AND FUTURE WORK

In this paper, we have demonstrated the effectiveness of feature extraction and selection in CT liver tumor classification system. A CAD intelligent system has been built with multi-classifier to achieve high accuracy. The feature fusion extracted from GLCM, FOs, LBP, and SFTA were further subjected to new swarm algorithm called Social Spider Optimization Algorithm (SSOA) to bring the best features with no noise and redundancy for optimal accuracy. SSOA has a strong capability to search in the problem space and can efficiently find minimal reducts. This algorithm considers two different search agents male and female. Depending on gender computational mechanisms are applied to avoid premature convergence and balance between exploration and exploitation. SSOA has selected only 13 features from 112 features to be used in classification phase. The feature selection algorithm SSOA compared with SFS, SFFS, RFS, MI, Relief, and Fisher. From experimental results SSOA prove much better performance, much robustness, and fast convergence speed. with excellent and near optimum solution on abdominal liver CT dataset with accuracy 99.27%, precision 99.37%, and recall 99.19%. In future work, we will apply our new meta-heuristic SSOA algorithm for feature sub set selection on another huge datasets to insure the performance and accuracy.

REFERENCES

[1] Park, C. J., Cho, E. K., Kwon, Y. H., Park, M. S., Park, J. W. (2005). Automatic separate algorithm of vein and artery for auto-segmentation liver-vessel from abdominal mdct image using morphological filtering. In Advances in Natural Computation, Springer Berlin Heidelberg, v.3612, pp.1069-1078.

[2] Arafa, N., El Hoseiny, M., Rekacewicz, C., Bakr, I., El-Kafrawy, S., El Daly, M., Fontanet, A. (2005). Changing pattern of hepatitis C virus spread in rural areas of Egypt. Journal of hepatology, 43(3), 418-424.

[3] Calle-Alonso, F., Perez, C. J., Arias-Nicolas, J. P., Martin, J. (2013). Computer-aided diagnosis system: A Bayesian hybrid classification method. Computer methods and programs in biomedicine, 112(1), 104-113.

[4] Huang, S. H., Wulsin, L. R., Li, H., Guo, J. (2009). Dimensionality reduction for knowledge discovery in medical claims database: application to antidepressant medication utilization study. Computer methods and programs in biomedicine, 93(2), 115-123.

[5] Costa, A. F., Humpire-Mamani, G., Traina, A. J. M. (2012). An efficient algorithm for fractal analysis of textures. In Graphics, Patterns and Images (SIBGRAPI), 2012 25th SIBGRAPI Conference on IEEE, pp.39-46.

[6] Ding, C., Peng, H. (2005). Minimum redundancy feature selection from microarray gene expression data. Journal of bioinformatics and computational biology, 3(02), 185-205.

[7] Gunasundari, S., Janakiraman, S. (2013). A study of textural analysis methods for the diagnosis of liver diseases from abdominal computed tomography. International Journal of Computer Applications, 74(11), 59-67.

[8] Kohavi, R., John, G. H. (1997). Wrappers for feature subset selection. Artificial intelligence, 97(1), 273-324.

[9] Kojadinovic, I., Wottka, T. (2000). Comparison between a filter and a wrapper approach to variable subset selection in regression problems. In Proc. European Symposium on Intelligent Techniques (ESIT).

[10] Peng, H., Long, F., Ding, C. (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 27(8), 1226-1238.

[11] Lu, Y., Liang, M., Ye, Z., Cao, L. (2015). Improved particle swarm optimization algorithm and its application in text feature selection. Applied Soft Computing, 35, 629-636.

[12] Azar, A. T., Elshazly, H. I., Hassanien, A. E., Elkorany, A. M. (2014). A random forest classifier for lymph diseases. Computer methods and programs in biomedicine, 113(2), 465-473.

[13] Pohjalainen, J., Rsnen, O., Kadioglu, S. (2015). Feature selection methods and their combinations in high-dimensional classification of speaker likability, intelligibility and personality traits. Computer Speech and Language, 29(1), 145-171.

[14] Emary, E., Zawbaa, H. M., Grosan, C., Hassenian, A. E. (2015). Feature Subset Selection Approach by Gray-Wolf Optimization. In Afro-European Conference for Industrial Advancement, Springer International Publishing, pp.1-13.

[15] James, J. Q., Li, V. O. (2015). A social spider algorithm for global optimization. Applied Soft Computing, 30, 614-627.

[16] Salomon, M., Sponarski, C., Larocque, A., Aviles, L. (2010). Social organization of the colonial spider Leucauge sp. in the Neotropics: vertical stratification within colonies. Journal of Arachnology, 38(3), 446-451.

[17] Yip, E. C., Powers, K. S., Aviles, L. (2008). Cooperative capture of large prey solves scaling challenge faced by spider societies. Proceedings of the National Academy of Sciences, 105(33), 11818-11822.

[18] Cuevas, E., Cienfuegos, M., Zaldvar, D., Perez-Cisneros, M. (2013). A swarm optimization algorithm inspired in the behavior of the social-spider. Expert Systems with Applications, 40(16), 6374-6384.

[19] Gletsos, M., Mougiakakou, S. G., Matsopoulos, G. K., Nikita, K. S., Nikita, A. S., Kelekis, D. (2003). A computer-aided diagnostic system to characterize CT focal liver lesions: design and optimization of a neural network classifier. Information Technology in Biomedicine, IEEE Transactions on, 7(3), 153-162.

[20] Cavouras, D., Prassopoulos, P., Karangellis, G., Raissaki, M., Kostaridou, L., Panayiotakis, G. (1996). Application of a neural network and four statistical classifiers in characterizing small focal liver lesions on CT. In Engineering in Medicine and Biology Society, 1996. Bridging Disciplines for Biomedicine. Proceedings of the 18th Annual International Conference of the IEEE, vol.3, pp. 1145-1146.

[21] Chen, E. L., Chung, P. C., Chen, C. L., Tsai, H. M., Chang, C. I.

(1998). An automatic diagnostic system for CT liver image classification. Biomedical Engineering, IEEE Transactions on, 45(6), 783-794.

[22] Mougiakakou, S. G., Valavanis, I. K., Nikita, A., Nikita, K. S. (2007). Differential diagnosis of CT focal liver lesions using texture features, feature selection and ensemble driven classifiers. Artificial Intelligence in Medicine, 41(1), 25-37.

[23] Kumar, S. S., Moni, R. S. (2010). Diagnosis of liver tumor from CT images using curvelet transform. International Journal of Computer Science and Engineering, 2(4), 1173-1178.

[24] Duda, D. O. R. O. T. A., Kretowski, M., Bezy-Wendling, J. O. H. A. N. N. E. (2006). Texture characterization for hepatic tumor recognition in multiphase CT. Biocybernetics and Biomedical Engineering, 26(4), 15.

[25] Kumar, S. S., Moni, R. S., Rajeesh, J. (2012). Liver tumor diagnosis by gray level and contourlet coefficients texture analysis. In Computing, Electronics and Electrical Technologies (ICCEET), 2012 International Conference on IEEE, pp.557-562.

[26] James, J. Q., Li, V. O. (2015). A social spider algorithm for global optimization. Applied Soft Computing, 30, 614-627.

[27] Djemame, S., Nekkache, M., Batouche, M. (2009, December). A Multi-Agent System for Image Segmentation A Bio-Inpired Approach. In Proceedings of the Third international conference on Innovation and Information and Communication Technology (pp. 17-17). British Computer Society.

[28] Pazhaniraja N., Anbusivam P., Raguraman S., Rajeshkanna M. (2014). Service Discovery and Selection using the Bio Inspired Approach. I. Journal of Engineering And Science (IJES), v.3, no.4, pp.42-48.

[29] Lenin K., Ravindhranath B. (2014). Improved Spider Algorithm for Solving Optimal Reactive Power Dispatch Problem. International Journal of Recent Research in Interdisciplinary Sciences (IJRRIS), vo.1, no.1, pp.35-46.

[30] Venkatesan, A., Parthiban, L. (2013). Hybridized Algorithms for Medical Image Segmentation. International Journal of Engineering and Advanced Technology (IJEAT), 3(2), pp.305-307.

[31] Wang, S., Xu, Y., Pang, Y. (2011). A fast underwater optical image segmentation algorithm based on a histogram weighted fuzzy C-means improved by PSO. Journal of Marine Science and Application, 10(1), 70-75.

[32] Deepa, G.(2012). Mammogram Image Segmentation Using Fuzzy Hybrid with Particle Swarm Optimization (PSO). International Journal of Engineering and Innovative Technology (IJEIT), 6(2), 167-171.

[33] Gopal, N., Karnan, M.(2010). Diagnose Brain Tumor through MRI using Image Processing Clustering Algorithms such as Fuzzy C Means along with Intelligent Optimization Techniques, International Conference and Computing Researcsh (ICCIC), pp.1-4.

[34] Mandelbrot, B. B. (1983). The fractal geometry of nature/Revised and enlarged edition. New York, WH Freeman and Co., pp.1-495.

[35] Costa, A. F., Humpire-Mamani, G., Traina, A. J. M. (2012). An efficient algorithm for fractal analysis of textures. In Graphics, Patterns and Images (SIBGRAPI), 2012 25th SIBGRAPI Conference on IEEE, pp.39-46.

[36] Aggarwal, N., Agrawal, R. K. (2012). First and second order statistics features for classification of magnetic resonance brain images, Journal of Signal and Information Processing, vol.3, pp.146-153.

[37] Anter, A. M., El Souod, M. A., Azar, A. T., Hassanien, A. E. (2014). A hybrid approach to diagnosis of hepatic tumors in computed tomography images. International Journal of Rough Sets and Data Analysis (IJRSDA), 1(2), 31-48.

[38] Kibriya, A. M., Frank, E. (2007). An empirical comparison of exact nearest neighbour algorithms. In Knowledge Discovery in Databases: PKDD 2007, Springer Berlin Heidelberg, pp. 140-151.

[39] Xu, H., Yu, B. (2010). Automatic thesaurus construction for spam filtering using revised back propagation neural network. Expert Systems with Applications, 37(1), 18-23.

[40] Wu, Q., Zhou, D. X. (2006). Analysis of Support Vector Machine Classification. Journal of Computational Analysis and Applications, 8(2).

[41] Bishop, C. M. (1995). Neural networks for pattern recognition. Oxford university press.

[42] Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochimica et Biophysica Acta (BBA)-Protein Structure, 405(2), 442-451.

[43] Anter, A. M., Hassanien, A. E., Schaefer, G. (2013). Automatic Segmentation and Classification of Liver Abnormalities Using Fractal Dimension. In Pattern Recognition (ACPR), 2013 2nd IAPR Asian Conference on IEEE, pp.937-941.

[44] Bilello, M., Gokturk, S. B., Desser, T., Napel, S., Jeffrey Jr, R. B., and Beaulieu, C. F. (2004). Automatic detection and classification of hypodense hepatic lesions on contrast-enhanced venous-phase CT. Medical physics, 31(9), 2584-2593.

[45] Zhang, X., Fujita, H., Qin, T., Zhao, J., Kanematsu, M., Hara, T., and Hoshi, H. (2008). CAD on liver using CT and MRI. In Medical Imaging and Informatics (pp. 367-376). Springer Berlin Heidelberg.

Containing a Confused Deputy on x86: A Survey of Privilege Escalation Mitigation Techniques

Scott Brookes
Thayer School of Engineering
Dartmouth College
Hanover, NH, USA

Stephen Taylor
Thayer School of Engineering
Dartmouth College
Hanover, NH, USA

***Abstract*—The weak separation between user- and kernel-space in modern operating systems facilitates several forms of privilege escalation. This paper provides a survey of protection techniques, both cutting-edge and time-tested, used to prevent common privilege escalation attacks. The techniques are compared against each other in terms of their effectiveness, their performance impact, the complexity of their implementation, and their impact on diversification techniques such as ASLR. Overall the literature provides a litany of disjoint techniques, each of which trades some performance cost for effectiveness against a particular isolated threat. No single technique was found to effectively mitigate all known and potential attack vectors with reasonable performance cost overhead.**

Keywords*—Protection & Security; Virtualization; Kernel ROP; ret2usr; Kernel Code Implant; rootkits; Operating Systems; Privilege Escalation

I. Introduction

The modern operating system kernel is one of the most basic building blocks of any complex computing or control system. It exists to provide a controlled interface to the hardware and to protect multiple processes and users from each others' actions. In order to accomplish these tasks securely, it must operate with a higher privilege level than user processes, making it an attractive target for attackers. As security research steadily enhances the security of individual processes, the kernel is being attacked more regularly. Despite the recent increase in popularity of attacking the kernel, system designers have long recognized the need for kernel security. MULTICS [1], [2] was one of the first operating systems to take security seriously and laid the groundwork for the most popular kernel security mechanisms still used today. In particular, it defined operating system "rings", designated by processor modes, and memory segmentation and paging structures with flexible read, write, and/or execute permission bits to allow memory partitioning and protection.

Unfortunately, almost all modern operating systems share a common vulnerability: a "weak" separation between kernel- and user-space. While the operating system provides a unique address space for each process in order to isolate processes from one-another, each address space must still allow access to kernel functionality. This is generally accomplished by sharing the address space of the kernel with each process. In contrast to the rare instances of "strong" separation between kernel- and user-space (such as the 4G/4G split Linux patch [3], 32-bit XNU [4], and certain systems using the hardware facilities provided by SPARC V9 hardware [5]), this weak separation protects the kernel from unauthorized access only with the mode of operation of the processor. A process that successfully manages to operate in supervisor mode has carte blanche access to all of the code and data of the kernel.

Often assisted by the weak separation of kernel- and user-space, all of the most popular kernels have been compromised by "rootkits" that give the attacker the highest level of privilege (i.e. "root") [6]–[8]. This survey is specifically interested in privilege escalation attacks that:

- *Hijack the facilities of the kernel* to create a "confused deputy" that is acting on behalf of the attacker [9]. This does not include attacks that are correctly exercising badly designed features of the kernel [10] or attacks that operate outside of the purview of the kernel [11].
- *Persist even without a specific kernel-level bug or design flaw.* Although most rootkits do require some kernel level bug (such as a buffer overflow) to be invoked, attacks that utilize a *specific* bug such as [12]–[14] are beyond the scope of this survey. Additionally, attacks such as [15] that are enabled by a specific kernel design flaw will not be considered. These cases typically have trivial solutions.
- *Elevate local privilege to root* rather than "horizontal" privilege escalation such as [16].
- *Effect x86 Architectures.* The focus of this article is on the x86 architecture because of its wide use in data centers and workstations [17]. However, some techniques specific to ARM will be examined because they do make valuable and interesting contributions to the state of the art.

The privilege escalation attacks that fit these criteria fall into three main categories: kernel code implants [18], kernel-mode return oriented programming (ROP) [19]–[21], and return-to-user (ret2usr) attacks [22].

Kernel code implants are attacks in which the adversary manages to overwrite existing code with (or inject) arbitrary instructions into the kernel space, and then direct the kernel to execute those instructions. Well-known examples of this type of attack include exploitation of classic buffer-overflow vulnerabilities associated with system calls [23]. If an attacker

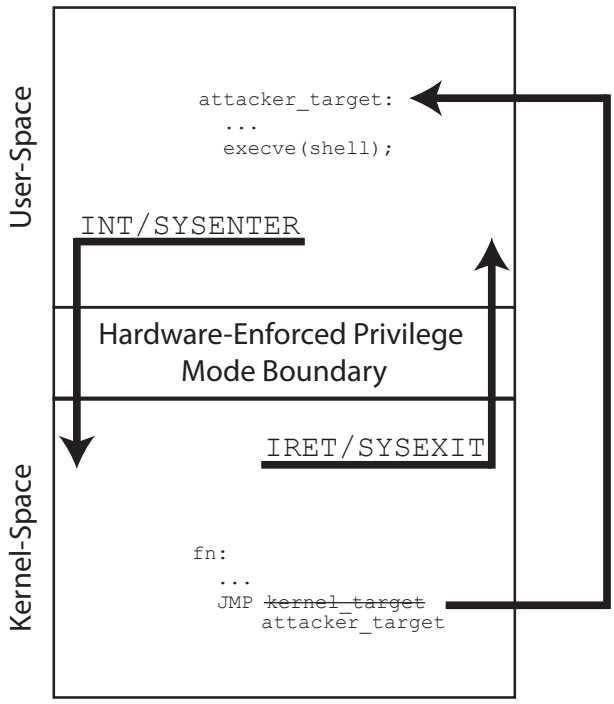


Fig. 1: Return-to-User Privilege Escalation Attack

manages to overflow a buffer on the kernel stack using some malformed arguments to a system call, it is possible to write shell-code onto the stack and overwrite a return-address so as to invoke the shell-code. This attack vector has largely been mitigated by techniques that mark the stack non-executable or provide canary code to detect overflows [24], [25] but it illustrates the core concept.

Kernel return-oriented programming (ROP) attacks defeat the use of a non-executable stack by using a payload, not of code directly on the stack, but of carefully crafted stack-frames that direct computation through a series of gadgets found in normal kernel code [19]–[21]. Research has shown that even small programs are likely to contain the gadgets necessary to generate a ROP Turing machine controlled only by a carefully crafted payload delivered to the stack [26]. All operating systems are large and complex enough to guarantee that the necessary gadgets will be present. As a result, an attacker with the appropriate knowledge can perform arbitrary computation using a ROP payload.

A return-to-user attack is enabled directly by weak kernel- and user-space separation. In this attack, illustrated in Figure 1, a user-controlled target associated with some kernel-code branch is set to an address in the normal user-space code. The compromised branch creates a path of execution that leaves kernel-code, entering user-code, without changing the CPU privilege level from supervisor mode to user mode. This attack results in the execution of user-controlled code with kernel-level privileges. Although hardware extensions such as Intel's SMEP [27] aim to mitigate this threat, these extensions are only slowly being adopted by operating systems and SMEP bypass techniques have already been demonstrated [28], [29].

Unfortunately, mitigation techniques for privilege escalation do not operate in isolation and it is important that they do not undermine other security features. For instance, it is easy to inadvertently inhibit techniques for enhancing security using non-determinism. This general class of technique was initially described by Cohen [30] and Forrest [31]. In the years since these seminal papers, many have explored the idea further. A recent survey of the area was presented in [32]. Address Space Layout Randomization (ASLR) is one of the most widely used applications of this technique. First implemented by the Linux PaX team [33], many other operating systems have implemented some form of ASLR including Mac OS X [34], Windows [35], and others. ASLR loads distinct memory regions including main program code, libraries, and the stack and heap at random locations within a program's virtual address space making it difficult to predict code entry points. More fine-grained techniques for diversifying the memory layout of a process [36], [37] require even more flexibility than traditional ASLR.

In Summary, this paper surveys the primary technologies presented in the literature to mitigate privilege escalation. It provides a comparative analysis based on their effectiveness, performance impact, and implementation complexity. It also specifically considers whether the technologies provide sufficient flexibility to coexist with state of the art address space layout randomization techniques. ASLR is chosen to provide a window to whether the techniques presented "play nicely" with other kernel security efforts because it is has widespread application on real systems and requires flexibility in order to be implemented fully. Section II examines techniques based on hypervisors and virtualization while the remaining techniques are discussed in section III. These techniques are compared and contrasted in section IV. Finally, some proposals are more accurately described as architectures than techniques. These are not directly comparable to the primary methods because they involve a dramatic paradigm shift. These approaches are briefly reviewed in Section V.

II. Mitigation Techniques based on Virtualization

Virtualization has dramatically changed the face of computing, not simply in terms of security and the way individual users interact with computers, but also by enabling cloud computing by allowing virtual machines to be migrated between servers. By adding a layer to the standard software stack, known as a hypervisor [38] or Virtual Machine Monitor (VMM) [39], an abstraction layer is introduced to isolate the operating system kernel from the hardware. In many ways, the hypervisor is to an operating system what an operating system is to a user process - serving to protect virtual machines from each other just as a kernel isolates user processes. The following approaches use virtualization as a means to deliver security guarantees to the kernel.

A. NICKLE

NICKLE [40] provides memory integrity to kernel code and thereby denies the execution of kernel code implants. It uses a VMM to maintain a "shadow" copy of memory that is verified when any kernel-code is loaded. This is achieved by comparing the memory to be loaded against a pre-computed cryptographic hash of the "clean" code distributed by the

manufacturer or developer of the code. At boot time, a known clean copy of the kernel is loaded into the shadow memory and whenever a kernel module is loaded at runtime, it is verified and added to the shadow memory.

With the integrity of the shadow memory guaranteed by off-line a priori cryptographic hashes of trusted code, NICKLE can ensure that no unauthorized kernel code is executed by directing all memory accesses targeting kernel code to retrieve from the shadow memory rather than from regular memory. Although no attempt is made to deny an attacker from modifying or injecting kernel code, kernel-mode execution is contained within trusted memory.

This is achieved transparently to the operating system kernel, allowing for commodity operating systems to be executed with NICKLE with no modification of kernel code. Additionally, NICKLE permits the mixing of kernel code and data within memory pages; this distinguishes NICKLE from many alternative approaches that require code and data to be loaded onto unique pages.

Unfortunately, NICKLE requires the off-line computation of cryptographic hashes for any code that may be executed; this poses a significant logistical issue for maintaining NICKLE on real systems and adds additional vulnerabilities associated with protection and distribution of hash values. NICKLE imposes a "minimal to moderate impact on system performance, relative to that of the respective original VMMs" averaging 1%-5% [40].

B. SecVisor

SecVisor [41] is an alternative virtualization technology leveraging hardware facilities to virtualize physical memory associated with modern processors. By utilizing this additional layer of translation from "guest physical" to "real physical" memory addresses, additional hardware memory protections can be enforced. This capability typically provides additional flexibility in creating memory access security; namely, any combination of read, write, and execute permissions can be allowed or denied on a particular page of memory [42].

SecVisor uses physical memory virtualization to mark only one of kernel- and user-space executable at a time. When a violation of security rules is detected, the protections can be swapped if the CPU has indeed changed privilege level, but are otherwise denied. This defeats ret2usr attacks by preventing unauthorized processor mode switches as shown in Figure 2. Additionally, the same virtualization allows SecVisor to enforce standard W$\oplus$X rules on all kernel code pages that the user has approved. This mitigates the possibility of a kernel code implant by verifying that all executable kernel code is non-writable and has been approved for execution by the user.

The security benefits of SecVisor are packaged in a tiny VMM that provides a small attack surface: only 4092 lines of source code in total. Unfortunately, SecVisor does have several weaknesses. The kernel running on top of SecVisor must guarantee that it does not share code and data on a single page. Additionally, the kernel has to be modified to cooperate with SecVisor by issuing VMCALLs to designate that it is loading or unloading kernel code. Finally, it imposes an overhead as high as 97% due to the additional translation

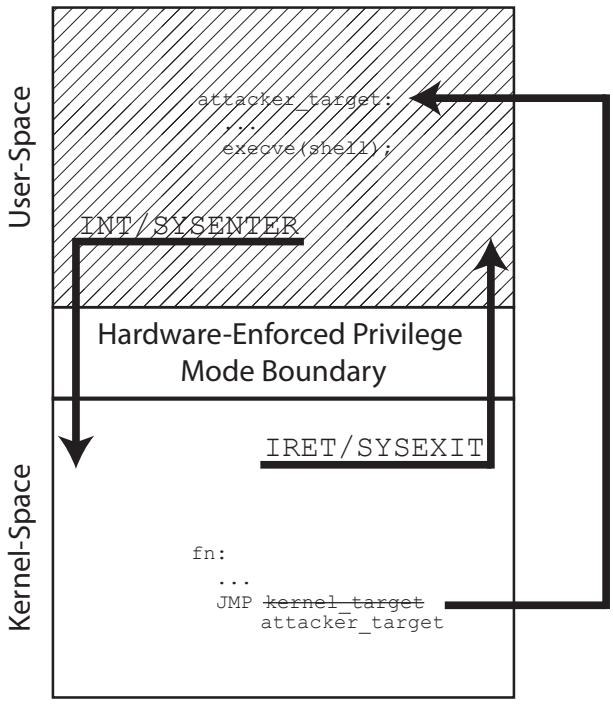


Fig. 2: SecVisor's Protection Against ret2usr Attacks

required by the virtualization of physical memory. For a full discussion of performance overhead costs, the interested reader should consult [41].

It is worth noting that [43] discovered two different bugs in the SecVisor implementation that allowed an attacker to violate rules that SecVisor claimed to enforce. Although these were implementation rather than design issues, and easily remedied, it is clear that even in a small code base security properties are difficult to reason about and correctly enforce.

C. SVA

The Secure Virtual Architecture (SVA) [44] is a set of architecture independent instructions that allow an operating system to interact with hardware. A kernel is ported to use these instructions, similar to porting a kernel to any new hardware architecture. Offline, an SVA compiler produces SVA byte-code from the kernel source code. This compiler has advanced features to provide memory safety and control-flow integrity at compile-time, similar to "safe" programming languages such as Java. The byte-code is distributed to users and executed on top of a virtualized SVA interpreter that performs the final step of translating to native target-dependent machine code.

The effort required to port an operating system to execute on SVA and the large performance cost are balanced by a promise of a substantial increase in security. Guaranteed memory safety and control-flow integrity deny common methods used to initiate ret2usr, kernel ROP, and kernel code implant attacks. An important point is that SVA does not set out to deny these attacks explicitly. Instead, it attempts to deny the vulnerabilities that enable these forms of attack, such as buffer overflows. Unfortunately, the infrastructure needed to support

SVA presents a significant hurdle. In addition to porting a kernel to a new architecture, SVA imposes restrictions on the kernel's memory allocation mechanisms that are likely to require modifications in kernel subsystems such as `kmalloc`. The performance cost is high, measured at approximately 50% on average, but at times reaching a 4-fold reduction.

D. KCoFI

Kernel Control Flow Integrity (KCoFI) [45] leverages the mechanics of the SVA implementation discussed previously, but offers only control flow integrity. Specifically, KCoFI ensures that function calls always enter at the beginning of some function's code, and that all returns from a particular function target the location of a possible call site. In order to prevent user-space applications from imitating the labels that KCoFI uses to validate branches, allowable address transitions are restricted to those within a certain pre-defined "kernel" range of virtual addresses. This limits the capabilities of advanced load-time randomization schemes. KCoFI also provides advanced treatment for the issues that make control flow integrity particularly difficult in the context of operating systems. In particular, it takes special care to handle interrupts, signals, DMA/devices, incomplete branch target information at compile-time, and page faults.

By verifying all branches at run-time, while the processor is in kernel mode, KCoFI manages to deny each of the three primary privilege escalation techniques described in this survey. Unfortunately, as with SVA, there is a large performance cost. Although the average performance impact on a standard application was 13%, worst-case costs up to 3.5-fold were reported. In addition, the method shares the SVA framework and therefore also requires porting the OS to a new "architecture," and pre-compiling the kernel and all of its modules with specialized SVA compilers.

E. SBCFI

State-based control-flow integrity (SBCFI) [46] provides course grained control-flow integrity for the operating system kernel. It sets itself apart from traditional control-flow integrity solutions, such as [47], in two ways. First, it implements monitoring externally from the kernel, in a hypervisor. Additionally, it assumes that attackers will generate persistent control-flow violations, therefore necessitating that kernel state is checked only periodically. Consequently, its introspection techniques allow SBCFI to detect any attack that persistently modifies the kernel's known control-flow graph.

The authors of [46] argue that trading strict security rules for performance by using SBCFI instead of complete CFI is acceptable because SBCFI will still detect most rootkits. In particular, they examined 25 rootkits found "in the wild" on Linux and found that all but one were detected by SBCFI. They suggest that attacker goals such as packet-sniffing or keystroke logging demand persistent rather than transient control-flow changes.

Unfortunately, SBCFI focuses on detection rather than prevention. This, combined with the focus on only persistent control-flow changes, leaves many avenues open to the attacker. SBCFI verifies the state of the kernel by checking a pre-computed hash of the kernel code and checking all function pointers stored in the kernel heap to verify that nothing has been changed. These checks would not detect a process that has achieved escalated privilege via a ret2usr attack or a kernel ROP payload. Overall, SBCFI manages to effectively deny persistent modifications to the kernel control-flow graph with minimal performance costs of less than 1% on average. However, it fails to address the general threat associated with privilege escalation.

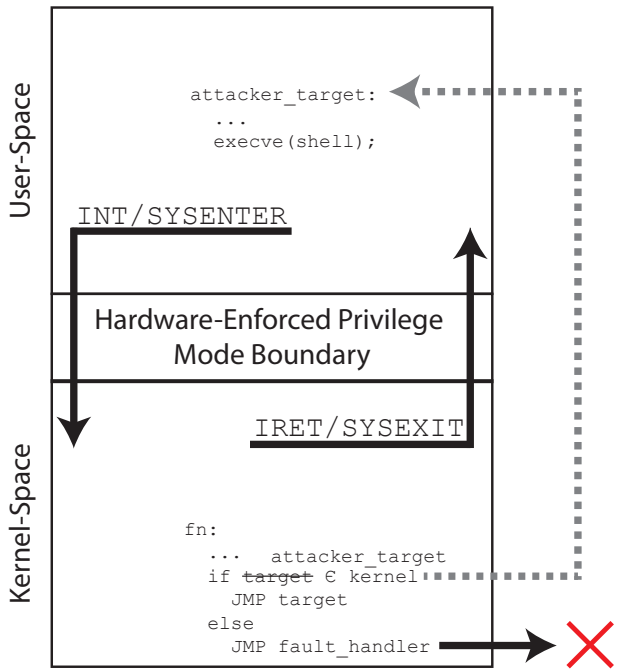


Fig. 3: kGuard's Protection Against ret2usr Attacks

III. Other Techniques

A. kGuard

kGuard [22] aims to deny ret2usr attacks by inserting guards on the kernel's control-flow at compile time as shown in Figure 3. On the x86 platform, the `call`, `jmp`, and `ret` instructions all redirect control-flow and therefore are vulnerable to being hijacked in order to redirect kernel execution into user-controlled code. kGuard places an inline check before any of these instructions. The checks are provided in two different forms depending on whether the target address is stored in a register or in memory. The checks simply verify that the branch target lies within kernel-space. Unfortunately, if an attacker controls the target of two branches, he can direct the first to jump directly to the second branch, bypassing the kGuard check completely. Since the second branch is in kernel-space, the check on the first branch would allow the control transfer. To avoid this attack, kGuard includes a compile-time code diversification mechanism that makes it difficult for the attacker to locate the address of the second branch.

One of the most significant advantages of kGuard is that, as a purely compile-time technique, it is portable to any operating system on any target hardware. It does not require any special hardware or impose many restrictions on the implementation of kernel features. Additionally, its average performance cost

is low at approximately 1%, making it deployable on existing systems. Unfortunately, kGuard does suffer from a variety of weaknesses. Although its simplicity lends itself to easy deployment, it is unable to protect against kernel-code implants or kernel-ROP. Although these are outside of the scope of kGuard, a kernel code implant could be used to create a ret2usr attack by implanting an unguarded jump into a user-space region. Therefore, another technique must be used in combination with kGuard to deny the possibility of a ret2usr scenario. This quickly increases in complexity and performance cost as multiple techniques need to be deployed on the same system. Additionally, kGuard's inline checks verify that the target of a control-flow transfer lies in kernel-space only by checking that it falls within a predefined range. This limits the capacity for deploying advanced code randomization during the loading of the kernel.

B. Return-less Kernels

Recall that Kernel ROP attacks requires "return" instructions in order to move from one gadget to another. In [48] the author utilizes "return indirection," introducing additional jumps at compile-time to disrupt this mechanism and defeat kernel ROP attacks. This approach uses a pre-computed table of all legal return addresses. Rather than pulling a return address from the stack and jumping to it at the end of a function, this method reads an address from the specified index in the return address table. If this table is trusted, the attacker could only modify which legal return address is used. It is assumed that most gadgets begin in a location other than a legal return address, and as a result this technique defeats the possibility of an attacker to craft a malicious payload.

In addition to introducing return indirection, [48] introduces compiler modifications to avoid instructions with an embedded "return" opcode. On an architecture such as x86, variable length instructions make it possible to read different instructions if the instruction pointer is offset some distance into an opcode. Without taking care at compile-time to avoid these scenarios, an attacker could still create gadgets by indexing the instruction pointer at unintended positions in the middle of the intended instruction.

The idea of using a return-less kernel is a clearly beneficial. It effectively mitigates a very particular risk with reasonable overhead, assessed at approximately 6%. Unfortunately, it does require modification to the kernel source. Functionality provided by compiling a higher-level language, such as C, does not need to be modified, but any functionality defined in assembly language must be manually modified to follow return-less principles. Since the kernel interfaces with hardware directly, there is a non-trivial level of assembly code included in most kernel implementations.

C. PaX

One of the first kernel-hardening efforts was implemented on Linux by the PaX team [49] circa 2000. UDEREF [50] utilizes segmentation to create a stricter separation between kernel- and user-space (denying ret2usr), while the PAGEXEC and Restricted `mprotect()` features essentially generate and enforce typical W$\oplus$X security rules on kernel code and data to mitigate kernel code implants.

PaX is valuable as a case study in hardening kernels. Unfortunately, it is less valuable as a mechanism for protecting modern kernels on today's hardware. Its protection mechanisms were based on Linux-specific software mechanisms (such as `mprotect()`) and x86-32-specific hardware features (such as segmentation). Additionally, the performance cost was significant, according to [22]. PaX-reported data about performance cost was available at the time of writing.

D. Sprobes and TZ-RKP

Sprobes [51] and TZ-RKP [52] both utilize the ARM TrustZone [53] hardware facilities included in modern ARM processors. TrustZone is a hardware-protected context that can run tangentially to the regular operation of the processor. The hardware disables the normal processor context from accessing anything within the "secure world" created by TrustZone and transitions between the regular context and TrustZone's secure context are limited by hardware to a small well-specified interface.

Sprobes [51] utilizes TrustZone by installing an introspection handler in the secure world and installing, at load- or run-time, special instructions that invoke the secure world at predetermined points in the execution of the kernel. When one of these probes is executed, control transfers to the secure world in which kernel state can be interrogated, control flow or memory contents verified, or any other number of actions can be taken. Furthermore, restrictions are placed on the normal world's ability to manipulate the virtual memory settings of the processor. The requirement that these systems be updated by the secure world guarantees that a kernel cannot manipulate virtual memory in order to bypass the probes.

TrustZone-based Real-time Kernel Protection (TZ-RKP) [52] is a similar approach that forces vital control operations involving the virtual memory layer to be routed through the secure world. TZ-RKP forgoes the probes provided by Sprobes, but takes a more extreme approach by limiting the kernel's control over important system state such as virtual memory. TZ-RKP forces all attempts to control virtual memory and other hardware resources through the secure world, providing a mechanism to verify any changes to the system state. With a controlled and static system state, it is easier to make claims about what an attacker may do to manipulate the kernel state.

Both [52] and [51] are built on the TrustZone architecture. The hardware underlying their implementation allows each to be implemented with a reasonable performance cost (typically 10%). TrustZone is also attractive because it manages to avoid the "turtles all the way down" problem in which software layer x is protected by introducing software layer $x-1$, which simply becomes the new target for attackers and instantiates the same problem again. Traditional virtualization can be criticized for this problem, but TrustZone holds itself off to the side of layer x rather than existing underneath it.

Unfortunately, TrustZone is an ARM-specific technology. Although ARM is used extensively in mobile and embedded applications, the x86 architecture continues to dominate desktop and server applications. Although these techniques are interesting, their utility is limited by a reliance on specialized hardware.

TABLE I: Summary of Examined Attack Mitigation Methods

Project	Kernel Code Implant	Kernel ROP	ret2usr	Typical Reported Performance Cost	Maximum Reported Performance Cost
NICKLE [40]	✓	✗	✗	1-5%	19.03%
KCoFI [45]	✓	✓	✓	13%	3.5×
SVA [44]	✓	✓	✓	50%	4×
SecVisor [41]	✓	✗	✓	20%	97%
SBCFI [46]	✓	✗	✗	<1%	13%
kGuard [22]	✗	✗	✓	1%	23.5%
PaX [49], [50]	✓	✗	✓	No Data	No Data
Return-less Kernel [48]	✗	✓	✗	6%	17.32%
Sprobes [51]	✓	✓	✓	10%	10%
TZ-RKP [52]	✓	✗	✓	3%	7.65%

TABLE II: Further Characteristics of Examined Methods

Project	x86-64 Compatible	(M)emory and/or (C]ontrol [F)low Integrity	Code-Diversity Compatible	LoC
NICKLE [40]	✓	M	✓	932
KCoFI [45]	✓	CF	✗	5579
SVA [44]	✓	M & CF	✓	No Data
SecVisor [41]	✓	M	✓	4092
SBCFI [46]	✓	CF	✓	No Data
kGuard [22]	✓	CF	✗	1000
PaX [49], [50]	✗	M & CF	✗	No Data
Return-less Kernel [48]	✓	CF	✓	2100
Sprobes [51]	✗	M & CF	✓	No Data
TZ-RKP [52]	✗	M	✓	No Data

IV. Comparison

Table I summarizes and compares the techniques discussed in the previous sections on the basis of their ability to mitigate privilege escalations and their expected cost:

- *Kernel Code Implant/Kernel ROP/ret2usr*: Does this technique mitigate the risk of privilege escalation associated with these particular attack vectors?
- *Typical/Maximum Performance Cost*: What is the typical and worst-case reported performance costs?

The performance costs listed represent only the maximum performance cost and an estimated average used only to illustrate differences between the techniques. In some cases these come from micro-benchmarks corresponding to small code segments, in other cases they come from macro-benchmarks corresponding to full applications. For the estimated average, they are often a mix of these tests. Each of the techniques offers thorough performance cost analyses that could not be summarized in a simple table. Interested readers should consult the original paper for each technique for a more complete treatment.

Table II compares the techniques on the basis of general observations regarding their operation:

- *x86-64 compatible*: Most desktop and server-class systems use the 64-bit x86 architecture. Is the technique viable with the hardware provided by the x86-64 hardware?
- *Memory and/or Control Flow Integrity*: Which is the primary mechanism by which the tool delivers its security guarantees?
- *Code-Diversity Compatible*: Is the technique sufficiently flexible to allow for advanced fine-grained address space layout randomization techniques?
- *Code Size*: How many lines of code (LoC), as a measure of the attack surface presented, are used in the implementation of the technique as presented?

It is clear from Table I that while KCoFI and SVA offer the most protection against the three different techniques associated with privilege escalation, they also come with dramatically more performance overhead than the other methods. This conforms to expectations in that the more thorough the security measure, the higher its performance impact. Sprobes and TZ-RKP appear exceptional as they enjoy the lowest performance costs and strong security claims. Unfortunately, each utilizes the ARM TrustZone architecture and consequently are unavailable on the Intel x86 architecture. Additionally, vulnerabilities have already been discovered in some TrustZone hardware implementations [54].

V. Paradigm-Shift Techniques

The techniques compared in Tables I and II each provide a modification to some part of the conventional kernel design, implementation, or build process that mitigates a particular threat. There are a few approaches, however, that attempt to offer similar security benefits by redefining the security paradigm rather than simply patching the status quo best practices. This radical departure from the current state of the art means that they cannot easily be compared to the previously described techniques. In all cases, it also means that they have not yet been widely accepted.

A. Microkernels

The idea of a microkernel departs from the standard "monolithic" kernel architecture by emphasizing a small codebase for the operating system kernel. There have been several examples of microkernels presented in the literature such as Mach [55], Minix [56], L4 [57], QNX [58], Bear [59], and many others.

All microkernels aim to minimize the source code in order to decrease the likelihood of vulnerabilities [60]. Additionally, a small code base allows for the possibility of using formal analysis and formal verification techniques [61], [62]. In order to keep the microkernel small, core functionality such as device drivers are migrated into user level processes. Additionally, many microkernels use message-passing for all communication between two processes or a process and the kernel. This provides a more easily verified and secured narrow interface between components.

By exporting core functionality, such as device drivers, into user-space microkernels struggle to offer the same levels of performance as monolithic kernels. Consequently, they have yet to replace monolithic kernels in common applications on commodity hardware.

B. ExoKernel

The ExoKernel [63] suggests redefining the nature of the kernel entirely. Rather than providing abstractions that the application developer can use to access hardware, the ExoKernel provides only the thinnest possible layer to manage the multiplexing of hardware resources. Therefore, the ExoKernel circumvents tasks normally reserved for the kernel such as buffering network communications, interrupt or exception handling, virtual memory management, and other normal kernel functions. Instead, each individual application must define its own abstractions to handle these tasks.

Although likely to offer more security for a system overall, the ExoKernel appears significantly complicate application development. Many of the tasks that a secure kernel can provide to protect all processes, such as virtual memory management, become the responsibility of the application developer. This is likely to make individual applications less secure since application programmers may lack the technical sophistication to interact directly with hardware, interrupts, atomicity, and concurrency. These central parts of the operating system exist to provide applications with well-defined interfaces to this complex functionality. The ExoKernel eliminates those interfaces by design.

C. Unikernels

Unikernels trade flexibility for security and performance by running a single process within a single address space [64]. Eliminating the requirement to support multiple processes and/or multiple users simplifies the code base required to implement a unikernel and reduces the overhead required to complete a single unit of useful work. Several examples have been deployed alongside virtualization technologies in cloud applications [65]–[67]. Despite their proven usefulness for providing fast, highly focused applications, unikernels don't, in isolation, provide protection from most of the attack vectors discussed in this paper. Additionally, in order to support the multiple-user multiple-job paradigm that conventional applications require to operate effectively, they require a hypervisor for scheduling and other process-management type tasks. In a sense, this is simply asking the hypervisor to act as an operating system and the same issues with conventional operating system design will simply move one layer deeper in the software stack.

VI. Conclusion

Each of the techniques examined in this survey makes valuable contributions to the security of modern operating systems. Those that offer the most comprehensive security suffer from high performance costs or specialty hardware requirements. On the other hand, many mitigate a specific, focused risk to kernel security while suffering only a small performance cost. Unfortunately, there is no single solution that offers both acceptable performance and comprehensive security coverage on the popular x86 platform. The impact of combining the techniques to improve coverage is not well understood in terms of complexity, performance, or security. This survey has also examined techniques that, rather than presenting incremental improvements on the status quo, attempt to dramatically redefine the notion of an operating system. These techniques also suffer from nontrivial performance costs in addition to the logistical challenges associated with a paradigm shift.

Overall, the kernel developer has a wide variety of techniques to choose from, but must balance individual strengths in privilege escalation prevention with the associated penalties in performance and complexity. The authors believe that future work aimed at mitigating privilege escalation will continue to have performance issues without some change in the underlying hardware or kernel design paradigms. Modern commodity operating systems are so highly developed that there is unlikely to be some technique hiding in a dark corner that will not decreasing perforance by requiring extra work.

Notice

The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the Defense Advanced Research Projects Agency (DARPA) or the U.S. Government. This material is based on research sponsored by DARPA under agreement number: FA8750-11-2-0257.

References

[1] R. C. Daley and J. B. Dennis, "Virtual memory, processes, and sharing in multics," *Communications of the ACM*, vol. 11, no. 5, pp. 306–312, 1968.

[2] F. J. Corbató and V. A. Vyssotsky, "Introduction and overview of the multics system," in *Proceedings of the November 30–December 1, 1965, Fall Joint Computer Conference, Part I*, ser. AFIPS '65 (Fall, part I). New York, NY, USA: ACM, 1965, pp. 185–196. [Online]. Available: http://doi.acm.org/10.1145/1463891.1463912

[3] I. Molnar, "4G/4G split on x86, 64 GB RAM (and more) support," July 2003. [Online]. Available: https://lwn.net/Articles/39283/

[4] D. Keuper, "Xnu: a security evaluation," December 2012. [Online]. Available: http://essay.utwente.nl/62852/

[5] R. McDougall and J. Mauro, *Solaris internals: Solaris 10 and OpenSolaris kernel architecture*. Pearson Education, 2006.

[6] K. Way, "Lastore-daemon in deepin 15 results in privilege escalation," February 2016. [Online]. Available: https://www.exploit-db.com/exploits/39433/

[7] J.-J. Khalife, "MS15-010/CVE-2015-0057 win32k Local Privilege Escalation," December 2015. [Online]. Available: https://www.exploit-db.com/exploits/39035/

[8] rebel, "issetugid() + rsh + libmalloc osx local root," July 2015. [Online]. Available: https://www.exploit-db.com/exploits/38371/

[9] N. Hardy, "The confused deputy: (or why capabilities might have been invented)," *SIGOPS Oper. Syst. Rev.*, vol. 22, no. 4, pp. 36–38, Oct. 1988. [Online]. Available: http://dl.acm.org/citation.cfm?id=54289.871709

[10] L. Davi, A. Dmitrienko, A.-R. Sadeghi, and M. Winandy, "Privilege escalation attacks on android," in *Information Security*, ser. Lecture Notes in Computer Science, M. Burmester, G. Tsudik, S. Magliveras, and I. Ili, Eds. Springer Berlin Heidelberg, 2011, vol. 6531, pp. 346–360. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-18178-8_30

[11] Kdm, "NTIllusion: A portable Win32 userland rootkit," 62. [Online]. Available: http://phrack.org/issues/62/12.html

[12] "CVE-2016-0728," January 2016. [Online]. Available: http://www.cve.mitre.org/cgi-bin/cvename.cgi?name=2016-0728

[13] "CVE-2013-2094," May 2013. [Online]. Available: http://www.cve.mitre.org/cgi-bin/cvename.cgi?name=2013-2094

[14] metasploit, "Chkroot local privilege escalation," November 2015. [Online]. Available: https://www.exploit-db.com/exploits/38775/

[15] V. P. Kemerlis, M. Polychronakis, and A. D. Keromytis, "Ret2dir: Rethinking kernel isolation," in *Proceedings of the 23rd USENIX Conference on Security Symposium*, ser. SEC'14. Berkeley, CA, USA: USENIX Association, 2014, pp. 957–972. [Online]. Available: http://dl.acm.org/citation.cfm?id=2671225.2671286

[16] "CWE-639: Authorization Bypass Through User-Controlled Key," September 2008. [Online]. Available: https://cwe.mitre.org/data/definitions/639.html

[17] T. P. Morgan, "x86 servers dominate the datacenter - for now," June 2015. [Online]. Available: http://www.nextplatform.com/2015/06/04/x86-servers-dominate-the-datacenter-for-now/

[18] A. Lineberry, "Malicious code injection via/dev/mem," 2009.

[19] E. Buchanan, R. Roemer, S. Savage, and H. Shacham, "Return-Oriented Programming: Exploitation without Code Injection," 2008. [Online]. Available: https://www.blackhat.com/presentations/bh-usa-08/Shacham/BH_US_08_Shacham_Return_Oriented_Programming.pdf

[20] H. Shacham, "The geometry of innocent flesh on the bone: Return-into-libc without function calls (on the x86)," in *Proceedings of the 14th ACM Conference on Computer and Communications Security*, ser. CCS '07. New York, NY, USA: ACM, 2007, pp. 552–561. [Online]. Available: http://doi.acm.org/10.1145/1315245.1315313

[21] R. Hund, T. Holz, and F. C. Freiling, "Return-oriented rootkits: Bypassing kernel code integrity protection mechanisms," in *Proceedings of the 18th Conference on USENIX Security Symposium*, ser. SSYM'09. Berkeley, CA, USA: USENIX Association, 2009, pp. 383–398. [Online]. Available: http://dl.acm.org/citation.cfm?id=1855768.1855792

[22] V. P. Kemerlis, G. Portokalidis, and A. D. Keromytis, "kGuard: Lightweight Kernel Protection Against Return-to-user Attacks," in *Proceedings of the 21st USENIX Conference on Security Symposium*, ser. Security'12. Berkeley, CA, USA: USENIX Association, 2012, pp. 39–39. [Online]. Available: http://dl.acm.org/citation.cfm?id=2362793.2362832

[23] A. One, "Smashing the stack for fun and profit," *Phrack*, vol. 7, no. 49, November 1996. [Online]. Available: http://www.phrack.com/issues.html?issue=49&id=14

[24] C. Cowan, C. Pu, D. Maier, H. Hintony, J. Walpole, P. Bakke, S. Beattie, A. Grier, P. Wagle, and Q. Zhang, "Stackguard: Automatic adaptive detection and prevention of buffer-overflow attacks," in *Proceedings of the 7th Conference on USENIX Security Symposium - Volume 7*, ser. SSYM'98. Berkeley, CA, USA: USENIX Association, 1998, pp. 5–5. [Online]. Available: http://dl.acm.org/citation.cfm?id=1267549.1267554

[25] C. Cowan, "Non-executable stack," 1997.

[26] "Microgadgets: Size does matter in turing-complete return-oriented programming," in *Presented as part of the 6th USENIX Workshop on Offensive Technologies*. Berkeley, CA: USENIX, 2012. [Online]. Available: https://www.usenix.org/conference/woot12/workshop-program/presentation/Homescu

[27] S. Fischer, "Supervisor mode execution protection," 2011, nSA Trusted Computing Conference and Exposition. [Online]. Available: https://www.ncsi.com/nsatc11/presentations/wednesday/emerging_technologies/fischer.pdf

[28] D. Rosenburg, "SMEP: What is it, and How to Beat it on Linux." June 2011. [Online]. Available: http://vulnfactory.org/blog/2011/06/05/smep-what-is-it-and-how-to-beat-it-on-linux/

[29] keegan, "Attacking Hardened Linux Systems with Kernel JIT Spraying," June 2011. [Online]. Available: http://mainisusuallyafunction.blogspot.com/2012/11/attacking-hardened-linux-systems-with.html

[30] F. B. Cohen, "Operating system protection through program evolution," *Computers & Security*, vol. 12, no. 6, pp. 565–584, 1993.

[31] S. Forrest, A. Somayaji, and D. Ackley, "Building diverse computer systems," in *Proceedings of the 6th Workshop on Hot Topics in Operating Systems (HotOS-VI)*, ser. HOTOS '97. Washington, DC, USA: IEEE Computer Society, 1997, pp. 67–. [Online]. Available: http://dl.acm.org/citation.cfm?id=822075.822408

[32] P. Larsen, A. Homescu, S. Brunthaler, and M. Franz, "Sok: Automated software diversity," in *Security and Privacy (SP), 2014 IEEE Symposium on*, May 2014, pp. 276–291.

[33] "Address Space Layout Randomization," PaX Team, Tech. Rep., 2001. [Online]. Available: https://pax.grsecurity.net/docs/aslr.txt

[34] "OS X Mavericks Core Technologies Overview," Apple, Tech. Rep., October 2013. [Online]. Available: http://www.apple.com/media/us/osx/2013/docs/OSX_Mavericks_Core_Technology_Overview.pdf

[35] O. Whitehouse, "An Analysis of Address Space Layout Randomization on Windows Vista," Symantec, Tech. Rep., 2007.

[36] M. Kanter and S. Taylor, "Attack Mitigation through Diversity," in *Military Communications Conference, MILCOM 2013 - 2013 IEEE*, Nov 2013, pp. 1410–1415.

[37] ——, "Diversity in Cloud Systems Through Runtime and Compile-time Relocation," in *Proceedings of the 13th IEEE Conference on Technologies for Homeland Security*, 2013, pp. 396–402.

[38] P. Barham, B. Dragovic, K. Fraser, S. Hand, T. Harris, A. Ho, R. Neugebauer, I. Pratt, and A. Warfield, "Xen and the art of virtualization," in *Proceedings of the Nineteenth ACM Symposium on Operating Systems Principles*, ser. SOSP '03. New York, NY, USA: ACM, 2003, pp. 164–177. [Online]. Available: http://doi.acm.org/10.1145/945445.945462

[39] R. P. Goldberg, "Survey of virtual machine research," *Computer*, vol. 7, no. 9, pp. 34–45, Sep. 1974. [Online]. Available: http://dx.doi.org/10.1109/MC.1974.6323581

[40] R. Riley, X. Jiang, and D. Xu, "Guest-transparent prevention of kernel rootkits with vmm-based memory shadowing," in *Proceedings of the 11th International Symposium on Recent Advances in Intrusion Detection*, ser. RAID '08. Berlin, Heidelberg: Springer-Verlag, 2008, pp. 1–20. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-87403-4_1

[41] A. Seshadri, M. Luk, N. Qu, and A. Perrig, "SecVisor: A Tiny Hypervisor to Provide Lifetime Kernel Code Integrity for Commodity OSes," *SIGOPS Oper. Syst. Rev.*, vol. 41, no. 6, pp. 335–350, Oct. 2007. [Online]. Available: http://doi.acm.org/10.1145/1323293.1294294

[42] *Intel 64 and IA-32 Architectures Software Developer's Manual Combined Volumes: 1, 2A, 2B, 2C, 3A, 3B, and 3C*, Intel, June 2014.

[43] J. Franklin, A. Seshadri, N. Qu, S. Chaki, and A. Datta, "Attacking, repairing, and verifying secvisor: A retrospective on the security of a hypervisor," Carnegie Mellon University, Tech. Rep., 2008.

[44] J. Criswell, A. Lenharth, D. Dhurjati, and V. Adve, "Secure virtual architecture: A safe execution environment for commodity operating systems," in *Proceedings of Twenty-first ACM SIGOPS Symposium on Operating Systems Principles*, ser. SOSP '07. New York, NY, USA: ACM, 2007, pp. 351–366. [Online]. Available: http://doi.acm.org/10.1145/1294261.1294295

[45] J. Criswell, N. Dautenhahn, and V. Adve, "Kcofi: Complete control-flow integrity for commodity operating system kernels," in *Security and Privacy (SP), 2014 IEEE Symposium on*, May 2014, pp. 292–307.

[46] N. L. Petroni, Jr. and M. Hicks, "Automated detection of persistent kernel control-flow attacks," in *Proceedings of the 14th ACM Conference on Computer and Communications Security*, ser. CCS '07. New York, NY, USA: ACM, 2007, pp. 103–115. [Online]. Available: http://doi.acm.org/10.1145/1315245.1315260

[47] U. Erlingsson and F. B. Schneider, "Sasi enforcement of security policies: A retrospective," in *Proceedings of the 1999 Workshop on New Security Paradigms*, ser. NSPW '99. New York, NY, USA: ACM, 2000, pp. 87–95. [Online]. Available: http://doi.acm.org/10.1145/335169.335201

[48] J. Li, Z. Wang, X. Jiang, M. Grace, and S. Bahram, "Defeating return-oriented rootkits with "return-less" kernels," in *Proceedings of the 5th European Conference on Computer Systems*, ser. EuroSys '10. New York, NY, USA: ACM, 2010, pp. 195–208. [Online]. Available: http://doi.acm.org/10.1145/1755913.1755934

[49] PaX, "Homepage of the PaX Team," 2013. [Online]. Available: http://pax.grsecurity.net

[50] B. Spengler, "uderef," 2007. [Online]. Available: https://grsecurity.net/~spender/uderef.txt

[51] X. Ge, H. Vijayakumar, and T. Jaeger, "Sprobes: Enforcing kernel code integrity on the trustzone architecture," *arXiv preprint arXiv:1410.7747*, 2014.

[52] A. M. Azab, P. Ning, J. Shah, Q. Chen, R. Bhutkar, G. Ganesh, J. Ma, and W. Shen, "Hypervision across worlds: Real-time kernel protection

from the arm trustzone secure world," in *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS '14. New York, NY, USA: ACM, 2014, pp. 90–102. [Online]. Available: http://doi.acm.org/10.1145/2660267.2660350

[53] "Building a Secure System using TrustZone Technology," ARM, Tech. Rep. [Online]. Available: http://infocenter.arm.com/help/topic/com.arm.doc.prd29-genc-009492c/PRD29-GENC-009492C_trustzone_security_whitepaper.pdf

[54] D. Rosenberg, "Qsee trustzone kernel integer over flow vulnerability," 2014.

[55] M. Accetta, R. Baron, W. Bolosky, D. Golub, R. Rashid, A. Tevanian, and M. Young, "Mach: A new kernel foundation for unix development," 1986, pp. 93–112.

[56] J. N. Herder, H. Bos, B. Gras, P. Homburg, and A. S. Tanenbaum, "Minix 3: A highly reliable, self-repairing operating system," *SIGOPS Oper. Syst. Rev.*, vol. 40, no. 3, pp. 80–89, Jul. 2006. [Online]. Available: http://doi.acm.org/10.1145/1151374.1151391

[57] D. Potts, S. Winwood, and G. Heiser, "Design and implementation of the l4 microkernel for alpha multiprocessors," 2002.

[58] D. Hildebrand, "An architectural overview of qnx," in *Proceedings of the Workshop on Micro-kernels and Other Kernel Architectures*. Berkeley, CA, USA: USENIX Association, 1992, pp. 113–126. [Online]. Available: http://dl.acm.org/citation.cfm?id=646405.759105

[59] C. Nichols, M. Kanter, and S. Taylor, "Bear – a resilient kernel for tactical missions," in *Military Communications Conference, MILCOM 2013 - 2013 IEEE*, Nov 2013, pp. 1416–1421.

[60] R. K. Pandey and V. Tiwari, "Article: Reliability issues in open source software," *International Journal of Computer Applications*, vol. 34, no. 1, pp. 34–38, November 2011, full text available.

[61] C. Baumann, B. Beckert, H. Blasum, and T. Bormer, "Formal verification of a microkernel used in dependable software systems," in *Computer Safety, Reliability, and Security*, ser. Lecture Notes in Computer Science, B. Buth, G. Rabe, and T. Seyfarth, Eds. Springer Berlin Heidelberg, 2009, vol. 5775, pp. 187–200. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-04468-7_16

[62] G. Klein, J. Andronick, K. Elphinstone, T. Murray, T. Sewell, R. Kolanski, and G. Heiser, "Comprehensive formal verification of an os microkernel," *ACM Trans. Comput. Syst.*, vol. 32, no. 1, pp. 2:1–2:70, Feb. 2014. [Online]. Available: http://doi.acm.org/10.1145/2560537

[63] D. R. Engler, M. F. Kaashoek, and J. O'Toole, Jr., "Exokernel: An operating system architecture for application-level resource management," in *Proceedings of the Fifteenth ACM Symposium on Operating Systems Principles*, ser. SOSP '95. New York, NY, USA: ACM, 1995, pp. 251–266. [Online]. Available: http://doi.acm.org/10.1145/224056.224076

[64] A. Madhavapeddy and D. J. Scott, "Unikernels: Rise of the virtual library operating system," *Queue*, vol. 11, no. 11, p. 30, 2013.

[65] A. Bratterud, A.-A. Walla, P. E. Engelstad, K. Begnum *et al.*, "Includeos: A minimal, resource efficient unikernel for cloud services," in *2015 IEEE 7th International Conference on Cloud Computing Technology and Science (CloudCom)*. IEEE, 2015, pp. 250–257.

[66] A. Kivity, D. Laor, G. Costa, P. Enberg, N. HarEl, D. Marti, and V. Zolotarov, "Osvoptimizing the operating system for virtual machines," in *2014 usenix annual technical conference (usenix atc 14)*, 2014, pp. 61–72.

[67] J. Martins, M. Ahmed, C. Raiciu, and F. Huici, "Enabling fast, dynamic network processing with clickos," in *Proceedings of the second ACM SIGCOMM workshop on Hot topics in software defined networking*. ACM, 2013, pp. 67–72.

Data Security, Privacy, Availability and Integrity in Cloud Computing: Issues and Current Solutions

Sultan Aldossary*
Department of Computer Sciences and cypersecurity
Florida Institute of Technology
Melbourne, Florida 32901

William Allen
Department of Computer Sciences and cypersecurity
Florida Institute of Technology
Melbourne, Florida 32901

*Prince Sattam Bin Abdulaziz University

***Abstract*—Cloud computing changed the world around us. Now people are moving their data to the cloud since data is getting bigger and needs to be accessible from many devices. Therefore, storing the data on the cloud becomes a norm. However, there are many issues that counter data stored in the cloud starting from virtual machine which is the mean to share resources in cloud and ending on cloud storage itself issues. In this paper, we present those issues that are preventing people from adopting the cloud and give a survey on solutions that have been done to minimize risks of these issues. For example, the data stored in the cloud needs to be confidential, preserving integrity and available. Moreover, sharing the data stored in the cloud among many users is still an issue since the cloud service provider is untrustworthy to manage authentication and authorization. In this paper, we list issues related to data stored in cloud storage and solutions to those issues which differ from other papers which focus on cloud as general.**

***Index Terms*—Data security; Data Confidentiality; Data Privacy; Cloud Computing; Cloud Security**

I. Introduction

Cloud computing now is everywhere. In many cases, users are using the cloud without knowing they are using it. According to [1], small and medium organizations will move to cloud computing because it will support fast access to their application and reduce the cost of infrastructure. The Cloud computing is not only a technical solution but also a business model that computing power can be sold and rented. Cloud computing is focused on delivering services. Organization data are being hosted in the cloud. The ownership of data is decreasing while agility and responsiveness are increasing. Organizations now are trying to avoid focusing on IT infrastructure. They need to focus on their business process to increase profitability. Therefore, the importance of cloud computing is increasing, becoming a huge market and receiving much attention from the academic and industrial communities. Cloud computing was defined in [2] by the US National Institute of Standards and Technology (NIST). They defined a cloud computing in [2] as a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction. Schematic definition of cloud computing can be simple, such as seen in Figure 1 1 This

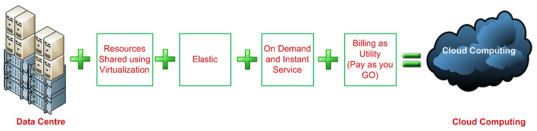


Fig. 1: Schematic definition of cloud computing [3]

cloud model is composed of five essential characteristics, three service models, and four deployment models as in the figure 2. In this technology users outsource their data to a server outside their premises, which is run by a cloud provider [4]. In addition, memory, processor, bandwidth and storage are visualized and can be accessed by a client using the Internet [5]. Cloud computing is composed of many technologies such as service oriented architecture, virtualization, web 2.0 and more. There are many security issues with cloud computing. However, the cloud is needed by organizations due to the need for abundant resources to be used in high demand and the lack of enough resources to satisfy this need. Also, cloud computing offers highly efficient data retrieval and availability. Cloud providers are taking the responsibility of resource optimization.

II. Characteristic of Cloud Computing:

There are five characteristics of cloud computing. The first one is on-demand self-service, where a consumer of services is provided the needed resources without human intervention and interaction with cloud provider. The second characteristic is broad network access, which means resources can be accessed from anywhere through a standard mechanism by thin or thick client platforms such mobile phone, laptop, and desktop computer. Another characteristic is resource pooling, which means the resources are pooled in order for multi-tenants to share the resources. In the multi-tenant model, resources are assigned dynamically to a consumer and after the consumer finishes it, it can be assigned to another one to respond to high resource demand. Even if consumers are assigned to resources on demand, they do not know the location of these

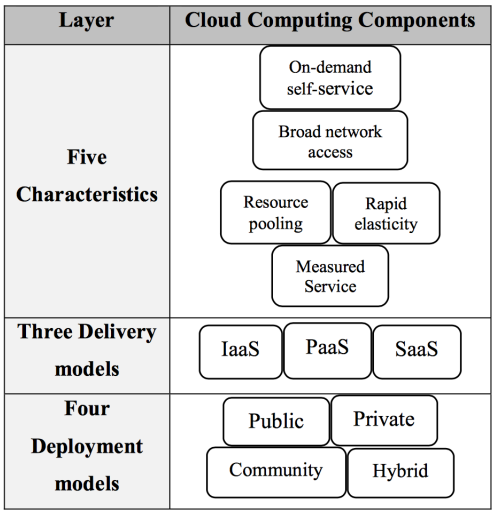


Fig. 2: Cloud environment architecture[6]

assigned resources. Sometimes they know the location at a high-level abstraction, such as country, state, and data center. Storage, processing, memory, and network are the kind of resources that are assigned. Rapid elasticity is also one of the cloud computing characteristics, which means that resources are dynamically increased when needed and decreased when there is no need. Also, one of characteristics that a consumer needs is measured service in order to know how much is consumed. Also, it is needed by the cloud provider in order to know how much the consumer has used in order to bill him or her.

III. Service Models

According to [2], there are three models. Those models differ in the capabilities that are offered to the consumer. It can be software, a platform, or infrastructure. In figure 3, it is comparison between those models with the traditional model.

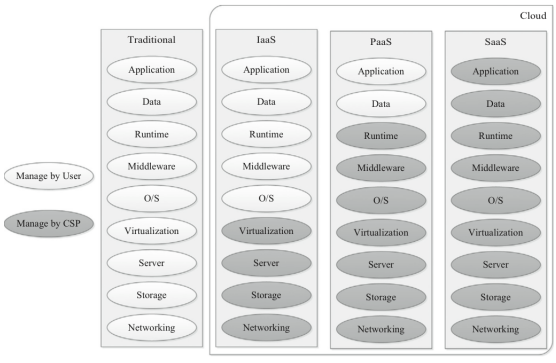


Fig. 3: Service oriented cloud computing architecture[7]

A. Software as a Service (SaaS)

In this service, the cloud service provider provides software and the cloud infrastructure to the clients so they can use this software on the cloud infrastructure for their applications. Since the clients can only run the software and use it, the client does not have control over the underlying infrastructure and physical setting of the cloud such as network, operating system, and storage. The cloud service provider is responsible and is the only one who is in charge of controlling underlying physical setting without client intervention. The client can access this software as a thin client through a web browser.

B. Platform as a Service (PaaS)

This service is similar to SaaS in that the infrastructure is controlled by the cloud service provider but is different in that the users can deploy their software. In this model, the clients can install and deploy their customized applications by using the tool offered by the cloud service provider. Physical settings are controlled and restricted by the cloud service provider and application settings are given to each user to control them.

C. Infrastructure as a Service (IaaS)

In this service, computing resources such as processing, storage and networks can be provisioned. The client of IaaS can install and use any arbitrary operating system. Also, the clients can install and deploy their applications on this operating system. Cloud services such as Amazon EC2 are adopting this model and charging their clients according to the resources are being utilized.

IV. Deployment Models:

Cloud deployment models have been discussed in the literature [8], [9], [10], [11], [12], [13], [14], [15]. There are four deployment models mentioned in [2] as following:

A. Private cloud

In this model, the cloud provider provides cloud infrastructure to a single organization that has many consumers. This infrastructure is to be used exclusively for their use and need. The owner, manager, and operator of this cloud could be the organization itself, a third party, or the organization and third party together. This private cloud could be on premises or off premises.

B. Community Cloud:

In this model, the cloud provider provides cloud infrastructure to many organizations that forms community that shares mission, security requirements, compliance consideration, or policy. this infrastructure is to be used exclusively for their uses and needs. The owner, manager, and operator of this cloud could be one of organizations, a third party, or the organization and third party together. This Community cloud could be on premises or off premises.

C. Public Cloud

This model differs from the previous model in that it is open for the public; it is not private and not exclusively for community. In this model, a public cloud can be provisioned for public to use it to satisfy their needs. The owner, manager,

and operator of this cloud could be a government, private organization, a business or academic organization, and sometimes many of them can be in one cloud and get the service from the same provider.

D. Hybrid Cloud

This model comprises two or more deployment models (private, community, or public). The cloud infrastructure can be combination of those models. Data center within an organization, private cloud, and public cloud can be combined in order to get services and data from both in order to create a well managed and unified computing environment. A cloud can be considered hybrid if the data moves from a data center to a private cloud or public cloud or vice versa.

V. Cloud Security Issues:

Even with these many benefits of cloud computing, previously mentioned, users are reluctant to adopt this technology and move from conventional computing to cloud computing [4]. In cloud computing, security is a broad topic. It is a mix of technologies, controls to safeguard the data, and policies to protect the data, services, and infrastructure. This combination is a target of possible attacks. Therefore, there are new security requirements in the cloud compared to traditional environments. Traditional security architecture is broken because the customer does not own the infrastructure any more. Also, the overall security cloud-based system is equal to the security of the weakest entity [16]. By outsourcing, users lose their physical control over data when it is stored in a remote server and they delegate their control to an untrusted cloud provider or party [17], [18]. Despite powerful and reliable server compared to client processing power and reliability, there are many threats facing the cloud not only from an outsider but also from an insider which can utilize cloud vulnerabilities to do harm [19]. These threats may jeopardize data confidentiality, data integrity, and data availability. Some untrusted providers could hide data breaches to save their reputations or free some space by deleting the less used or accessed data [20].

VI. Top Threats to Cloud Computing

Cloud computing is facing a lot of issues. Those issues are listed as the following: data loss, data breaches, malicious insiders, insecure interfaces and APIs, account or Service hijacking, data location, and denial of Service.

A. Data Loss:

Companies are outsourcing their entire data to cloud service providers. Because of the low cost rate that the cloud offers, the customers should make sure not to expose their important data to risks because of the many ways to compromise their data. In cloud computing, the risks are going up because there are risks that is newly facing the cloud and did not happen to traditional computing, and challenges taking to avoid those risks.[3]. There are many possibilities of losing data due to a malicious attack and sometimes due to server crashes or unintentional deletion by the provider without having backups. Catastrophic events like an earthquake and fire could be the causes of loss. Also, any event that leads to harming the encryption keys could lead to data loss to[21]. In order to avoid losing the data, there are many solutions proposed by CSA[22]:

- Using a strong API for access control
- While the data is in transit, encrypting and protecting its integrity
- Analyzing data protection at run time and design time
- Using strong key generation, storage, destruction, and management practices
- Requiring the service provider to wipe the persistent media data before releasing it to the pool
- Specifying the back up and retention strategies

B. Data Breaches:

A cloud environment has various users and organizations, whose data are in the same place. Any breach to this cloud environment would expose all users' and organizations' data to be unclosed[1]. Because of multi-tenancy, customers using different applications on virtual machines could share the same database and any corruption event that happens to it is going to affect others sharing the same database[21]. Also, even SaaS providers have claimed that they provide more security to customers? data than conventional providers. An insider can access the data but in different ways; he or she is accessing the data indirectly by accessing a lot of information in their cloud and incident could make the cloud insecure and expose customers' data[1]. In [23], it was reported "2011 Data Breach Investigations Report" that hacking and malware are the common causes of data breaches, with 50% hacking and 49% malware.

C. Malicious Insiders:

Malicious insiders are the people who are authorized to manage the data such as database administrators or employees of the company offering cloud services[21], partners, and contractors who have access to the data. Those people can steal or corrupt the data whether they are getting paid by other companies or to just hurt a company. Even the cloud providers may not be aware of that because of their inability in managing their employees. There are many solutions proposed by CSA[22]:

- Conducting a comprehensive supplier assessment and making supply chain management ID stricter
- As part of the legal contract, defining human resources requirements
- Making information security and all cloud service practices more transparent
- creating a process to notify when data breaches happen

D. Insecure interfaces and APIs:

The communication between the cloud service provider and the client is through the API through which the clients can manage and control their data[21]. Therefore, those interfaces

should be secure to prevent any unauthorized access. If they are weak and security mechanism cannot defend them, this could lead to accessing resources even as privileged user. There are many solutions proposed by CSA[22] to avoid insecure interfaces and APIs:

- Analyzing the security model for interfaces of the cloud provider
- Making a strong access control and authentication when data is transmitted
- Understanding dependencies in API

E. Account or Service Hijacking:

Users are using passwords to access the cloud service resources so when their accounts are hijacked and stolen, the passwords are misused and altered unsurprisingly[21]. The unauthorized user who has a password can access the clients' data by stealing it, altering it, or deleting it, or for the benefit of selling it to others. There are many solutions proposed by CSA[22] to avoid account or service hijacking:

- Preventing users from sharing their credentials
- Using a two-factor authentication system
- Monitoring all activities to detect unauthorized access
- Understanding security policies and SLAs

F. Data Location:

Cloud providers have many centers widespread over many places. Data location is an issue in cloud computing since the users of clouds need to know where their data is stored. Some countries, according to jurisdiction, require their companies to store their data in their country. Also, there are regulations in some countries where the company can store their data. Also, the data location matters when the user data is stored in a location that is prone to wars and disasters.

G. Denial of Service:

Some organizations need their systems to be available all the time because availability is important to them due to the critical services they provide. The cloud services provider offers resources that are shared among many clients. If an attacker uses all available resources, others cannot use those resources, which leads to denial of service and could slow accessing those resources. Also, customers, who are using cloud service and affected by botnet, could work to affect availability of other providers.

VII. Multitenancy

In [2], the author did not consider multitenancy as an essential characteristic of cloud computing. However, in CSA [24] and ENISA [25], multi-tenancy is considered an important part of cloud computing. However, with the many benefits multi-tenancy offers, this leads to many challenges regarding having more than one tenant on one physical machine, which is required to utilize the infrastructure. Since tenants are in the same place, they could attack each other. Previously, an attack could be between two separate physical machine but now because two or more tenants are sharing the same hardware, an attacker and a victim can be in the same place. In figure 4, the difference between multi-tenancy and traditional cases is shown. The technology is used to keep tenants from each other by providing a boundary for each tenant by using virtualization. However, virtualization itself is suffering from many issues.

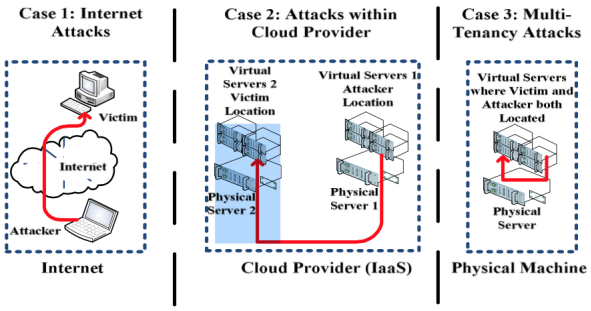


Fig. 4: Difference between Multi-Tenancy and Traditional Cases.[26]

VIII. Virtualization Security Issues

Virtualization is an important component of cloud computing. Now it is getting more attention from academic and industrial communities. Virtualization means separation of underlying hardware resources from provided resources. By using virtualization, two or more operating systems might run in the single machine with each having its own resources.

A. Cross Virtual Machine(VM) Side-Channel Attacks

This attack requires the attacker to be in another virtual machine on the same physical hardware with the victim. In this attack, the attacker and victim are using the same processor and same cache. When the attacker alternates with the victim's VM execution, the attacker can attain some information about the victim's behavior. In [27], there is an example of VM side-channel attack and how the attacker can infer some information about a victim. The timing side channel attack is one kind of VM side channel attacks[28]. This attack is based on determining the time needed by various computations. Determining this time can lead to leaking sensitive information such as described in[28]. This attack can help in leaking some sensitive information such as to the one who performs this computation or sometimes leaking information out of cloud provider itself. This attack is hard to detect because the owner of VM can check other VMs due privacy concern. Sometimes cloud providers can detect a side channel attack but to protect their reputation but they do not announce it. Moreover, there is another type of side channel attacks which is energy-consumption side channel [29].

B. VM Image Sharing

VM can be instantiated from a VM image. A shared image repository can be used to share VM images or a user can have his own VM image [30]. Since there is a repository for sharing VM images, some malicious users could take advantage of this

feature in order to inject a code inside a VM [31]. This will lead to a serious problem. For example, a VM image may contain malware. This malware is coming from the user who used it before[31]. If the image is returned without properly cleaning it, sensitive data could be leaked [30].

C. VM Isolation

Since VMs run in the same hardware, they share all components such as processor, memory, and storage. Isolating of VM logically to prevent one from intervening with another is not enough since they are sharing computation, memory, and storage. Therefore, the data may leak when it is in computation or memory or storage. This is a serious issue. Hence, isolation should be at the level of VM and hardware such as processor, memory, and storage [32].

D. VM escape

The VMs or a malicious user escape from the virtual machine manager(VMM) supervision [33]. VMM controls all VMs and it is the layer that controls how the VM or a user who uses the underlying resources such as hardware. One of the most serious scenarios is that malicious code can go through unnoticed from the VMM and then can interfere with the hypervisor or other guests [31].

E. VM Migration

VM migration process suspends the running VM, copies the status from the source Virtual Machine Monitor (VMM) to the destination VMM and resumes the VM at the destination[11]. In virtual machine migration, the running VM is suspended, has its status copied to the virtual machine monitor (VMM) from its source VMM, and is resumed on the destination VMM[34]. In [35], VM migration is defined as the moving of a VM from one physical machine to another while it is running without shutting it down. Fault tolerance, load balancing, and maintenance are some causes of VM migration [30], [36]. The data and the code of VM [35] are exposed when transferring in the network between two physical hardware locations when they are vulnerable to an attacker. Also, an attacker could let VM transfer to a vulnerable server in order to compromise it. When hen an attacker compromises the VMM, he can get a VM from this data center and migrate it to other centers. Therefore, he can access all resources as a legitimate VM[37]. Therefore, this process incurs more challenge and needs to be secured [30] In order to prevent attackers from benefiting.

F. VM Rollback

This is a process of rolling back a VM to its previous state. Since this process adds more flexibility to the user, it has more security issues. For example, a VM could be rolled back to previous vulnerable state that has not been fixed [38] or it can rolled back to an old security policy or old configuration [30]. In another example, a user could be disabled in a previous state and when the owner of the VM rolls back, the user can still have access [30].

G. Hypervisor Issues:

Hypervisor and virtual machine monitor are the main parts of virtualization. The virtual machine monitor is responsible for managing and isolating VMs from each other. The VMM is the intermediary between the hardware and VMs, so it is responsible for proving, managing, and assigning of the resources. Also, hypervisor with full control of hardware can access Vms' memory[39]. In [39], Jin et al. propose a hardware based soultion to protect VM's memory pages from the malicious hypervisor.

IX. Data Integrity Issues

Data that is stored in the cloud could suffer from the damage on transmitting to/from cloud data storage. Since the data and computation are outsourced to a remote server, the data integrity should be maintained and checked constantly in order to prove that data and computation are intact. Data integrity means data should be kept from unauthorized modification. Any modification to the data should be detected. Computation integrity means that program execution should be as expected and be kept from malware, an insider, or a malicious user that could change the program execution and render an incorrect result. Any deviation from normal computation should be detected. Integrity should be checked at the data level and computation level. Data integrity could help in getting lost data or notifying if there is data manipulation. The following is two examples of how the data integrity could be violated.

A. Data Loss or Manipulation

Users have a huge number of user files. Therefore, cloud providers provide Storage as Service(SasS). Those files can be accessed every day or sometimes rarely. Therefore, there is a strong need to keep them correct. This need is caused by the nature of cloud computing since the data is outsourced to a remote cloud, which is unsecured and unreliable. Since the cloud is untrustworthy, the data might be lost or modified by unauthorized users. In many cases, data could be altered intentional or accidentally. Also, there are many administrative errors that could cause losing data such as getting or restoring incorrect backups. The attacker could utilize the users outsourced data since they have lost the control over it.

B. Untrusted Remote Server Performing Computation on Behave of User

Cloud computing is not just about storage. Also, there are some intensive computations that need cloud processing power in order to perform their tasks. Therefore, users outsource their computations. Since the cloud provider is not in the security boundary and is not transparent to the owner of the tasks, no one will prove whether the computation integrity is intact or not. Sometimes, the cloud provider behaves in such a way that no one will discover a deviation of computation from normal execution. Because the resources have a value to the cloud provider, the cloud provider could not execute the task in a proper manner. Even if the cloud provider is considered more secure, there are many issues such as those coming from

the cloud provider's underlying systems, vulnerable code or misconfiguration.

X. Protecting Data Integrity

Tenants of cloud systems commonly assume that if their data is encrypted before outsourcing it to the cloud, it is secure enough. Although encryption is to provide solid confidentiality against attack from a cloud provider, it does not protect that data from corruption caused by configuration errors and software bugs. There are two traditional ways of proving the integrity of data outsourced in a remote server. Checking the integrity of data can be by a client or by a third party. The first one is downloading the file and then checking the hash value. In this way, a message authentication code algorithm is used. MAC algorithms take two inputs, which are a secret key and variable length of data, which produce one output, which is a MAC (tag). In this way this algorithm is run on the client side. After getting a MAC, the data owner outsources those data to the cloud. For checking its integrity, the data owner downloads the outsourced data and then calculates the MAC for it and compares it with the one calculated before outsourcing that data. By using this method accidental and intentional changes will be detected. Also, by using the key, the authenticity of data will be protected and only the one who has the key can check the data authenticity and integrity. For a large file, downloading and calculating the MAC of the file is an overwhelming process and takes a lot of time. Also, it is not practical since it consumes more bandwidth. Therefore, there is a need for using a lighter technique, which is calculating the hashing value.

The second one is to compute that hash value in the cloud by using a hash tree. In this technique, the hash tree is built from bottom to top where the leaves are the data and parents are also hashed together until the root is reached. The owner of data only stores the root. When the owner needs to check his data, he asks for just root value and compares it with the one he has. This is also to some extent is not practical because computing the hash value of a huge number of values consumes more computation. Sometimes, when the provided service is just storage without computation, the user download the file, the same as in the first case, or send it to third party, which will consume more bandwidth. Therefore, there is a need to find a way to check data integrity while saving bandwidth and computation power. Remote data auditing, by which the data integrity or correctness of remotely stored data is investigated, has been given more attention recently [40], [41], [42], [43], [44], [45]

A. Third Party Auditor

Third Party Auditor (TPA) is the person who has the skills and experience to carry out all auditing processes such as in the figure5. TPA scheme is used for checking the data integrity. Since there are many incidents and doubtful actions, users of cloud storage depend on third party auditors [46]. In [47], Balusamy et al. proposed a framework, which involves the data owner in checking the integrity of their outsourced data.

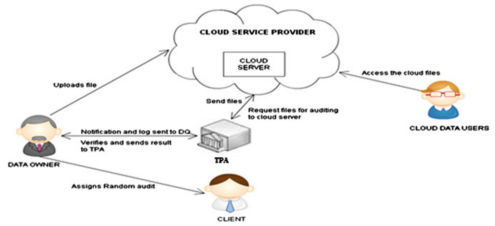


Fig. 5: Architecture of third-party auditing [47]

Their proposed scheme attains data integrity and assures the data owner of the data security. The owner is aware of all his resources on the cloud. Therefore, this scheme guarantees the integrity of data for all owner resources on the cloud. This scheme involves the data owner in the auditing process. First, TPA uses normal auditing processes. Once they discover any modification to the data, the owner is notified about those changes. The owner checks the logs of the auditing process to validate those changes. If the owner suspects that unusual actions have happened to his data, he can check his data by himself or by another auditor assigned by him. Therefore, the owner is always tracking any modification to his own data. There is an assigned threshold value that a response from the third party auditor should not exceed. The data owner validates all modifications lesser than or equal to this threshold. If the time exceeds this threshold, the data owner is supposed to do surprise auditing. The figure 6 shows this auditing process.

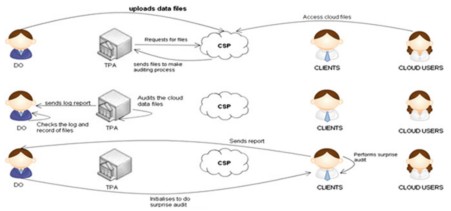


Fig. 6: Proposed scheme architecture [47]

B. Provable Data Possession

In [41] Ateniese et al. proposed the first the Provable Data Possession (PDP) scheme to investigate statically the correctness of the data outsourced to cloud storage without retrieving the data. In [41], the proposed model is to check that data stored in a remote server are still in its possession and that the server has the original data without retrieving it. This model is based on probabilistic proofs by randomly choosing a set of blocks from the server to prove the possession. They used a RSA-based homomorphic verifiable tag, which is combines tags in order to provide a message that the client can use to prove that the server has specific block regardless of whether the client has access to this specific block or not. Even with the advantages this scheme offers, they did not deal with dynamic data storage, and there is computation and communication overhead in the server because of the whole file RSA numbering. In the case of a prover that is untrusted

or has malicious intent, this scheme fails in proofing data possession [7].

In [42], Ateniese et al. overcome the limitation in [41]. By using symmetric cryptography, they proposed a PDP scheme that supports partial and dynamic verification. The limitation of this proposition is that it does not support auditability.

Since PDP schemes just check parts of the file for integrity, there is a need to correct blocks when they suffer from corruption due to hardware issue. In [48], Ateniese et al. propose a scheme to prove data possession with using Forward Error Checking(FEC). First, the file is encoded by using FEC. Then, the encoded file is used by PDP scheme. This methods help in finding the corruptions and mitigating them.

In [44], Wang et al. propose a new dynamic PDP for auditing remote dynamic data. They use the Merkle Hash Function(MHT) and the bilinear aggregate signature. They modify Merkle Hash Function structure by sort leafs node of MHK to be from left to right. This sorting will help in identifying the location of the update. However, this method incur more computation overhead when the file is large.

Sookhak et al.[49] propose a new method for dynamic remote data auditing by using algebraic signature and a new data structure called Divide and Conquer Table(DCT). DCT keep track of the data after appending, updating, insertions, and deletion. Therefore, The need of downloading the file for checking the integrity is avoided.

C. Proof of Retrievability

PDP differs from proof of retrievabilty in that PDP only detects when corruption happens to a large amount of data[50]. PDP protocols can be verified publicly or privately. In the protocol that is privately verifiable, only the owner of the key can verify the encoded data, while in publicly verifiable protocol, data integrity can be verified or audited by a third party. Proof of retrievability is a cryptographic approach based on a challenge response protocol in which a piece of data is proved to be intact and retrievable without retrieving it from the cloud. The the simplest form of proof of retrievability is taking the hash of block using a keyed hash function. Owner of data takes the hash values of the file by using keyed hash function. After getting the hash values, the data owner keep the key and the hash values. the data owner sends the file to a remote server. When the data owner needs to check his data retrievabilty, he sends his key and asks the server to send the hash values by using his key in order to compare them with the hash values that data owner has. The advantage of this solution is that it is simple and implementable. However, there are many disadvantages such that the data owner needs to store many keys in order to use one each time. Also, the number of checking is limited by the number of keys since the remote server could store all keys and the hash values and use them when it is asked to prove having that file. In addition, it costs more resources on the side of a client and server since the hash values need to be calculated each time when the proof is required. Moreover, some thin client such mobile device and PDA does not have the resources to calculate the hash values of big files.

In [50], They used an error correction code and spot checking to prove the possession and retrivability of the data. The verifer hides some sentinels among file blocks before sending them to the remote server. When the verifer wants to check retrivability of the data, it only asks the server for those sentinels. In order to keep those sentinels indistinguishable for the the remote server, the data owner encrypts the file after adding sentinels. In contrast to the simple one, it uses one key regardless of the size of the file. Also, unlike the simple solution that the entire file is processed, it accesses only parts of file. Therefore, the I/O operations is less. This scheme has disadvantages such that the files need to be in encrypted form so it incurs computation overhead in clients such as mobile devices and PDA.

D. Proof of Ownership

In this notion, the client proves ownership of the file outsourced by the client to server. This notion differs from POR and PDP in that POR and PDP need to embed some secret in the file before outsourcing it and the client can check with the cloud server whether the file is in there by asking for the secret and comparing it with what he has. The proof of ownership comes after the need to save some storage by duplication. The owner of the files needs to prove to the server he owns this file.

In [51], Halevi et al. introduced the proof of ownership idea. In [51], the ideas behind proving the ownership are the Collision Resistant Hash functions and Merkle Hash Tree. In [51],The owner of a file creates a Merkle Hash Tree (MHT) and sends the file to the cloud, called verifier. Once it is received by cloud, the file is divided into bits using pairwise independent hash and then the verifier creates a Merkle Hash Tree for this file. Once the prover asks for the ownership of the file, the verifier sends a challenge, which is the root and the number of leaves. The prover calculates the sibling path and returns it to verifier as proof of ownership of this file. The verifier after receiving the sibling path,checks this path against what the merkle tree has and validate the prover. However, this violate the privacy of users since their sensitive data is leaked to the remote server and this issue does not addressed by Halevi et al in [51]. Therefore, there has to be a way to prevent that remote server from accessing outsourced data and building a user profile[52].

XI. Data availability

In [53], Fawaz, et al. developed a storage architecture, figure 7 which covers security, reliability, and availability. The underlying technique of their proposed architecture uses a storage method based on RAID 10. They used three server providers and stripped the data to two servers and the party bits in the third server provider. They followed a sequential way to store the data after encrypting it and dividing the cipher into blocks. One block is in one server provider storage, the next block is in the next server provider storage and the parity

bit in the third server provider. A Parity bit can be in any server provider storage while the other in the other server provider storage. In case the two server providers collide to collect the data, each one has, the encryption will protect the data from unauthorized access. In case one server provider service is distributed, by using a parity bit and an available server provider, the service will be available. Also, it is the same in case one service provider corrupts the data. The number of service provider in this storage architecture can be any number.

In [54], a HAIL (High Availability and integrity Layer) is designed to address the threat caused by a service provider being unavailable. A HAIL distributes the data across many cloud providers to keep their service available all the time. A HAIL leverages many cloud service providers to make a solution that is reliable out of unreliable components and it is cost effective. The idea behind the HAIL is inspired by RAID, which is reliable storage made from unreliable storage. The HAIL works when there is corruption. It does not detect the corruption but it remedies it by avoiding this corruption in a subset of storage providers by using the data in the other service provider storage.

In [55], Bessani et al. proposed Depsky which uses many clouds to build a cloud-of-clouds to address two security requirements in their storage system, which are confidentiality and availability of data. They combined the byzantine quorum protocol as well as secret sharing cryptographic and erasure codes.

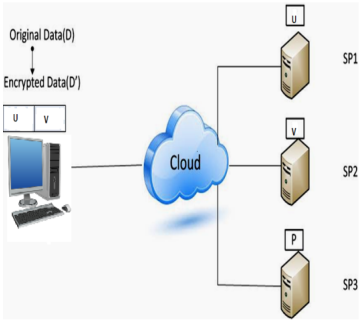


Fig. 7: The proposed parity scheme [53]

XII. Data confidentiality issues

Usually the data is encrypted before it is outsourced. The service provider gets encrypted data. Therefore, it is considered not useful or meaningless. However, the client is responsible for handling the access control policy, encrypting the data, decrypting it and managing the cryptographic keys[56]. Even this would cause a burden to the user; sharing it with others exposes it to risks. When the data is shared among many users, there has to be more flexibility in the encryption process to handle users of the group, manage the keys between users, and enforce the access control policy in order to protect the data confidentiality[57]. Sharing the data among a group of users adds more burden on the owner of the outsourced data.

In [59], the authors describe a cryptosystem in which the data owner encrypts the data by using his public key and

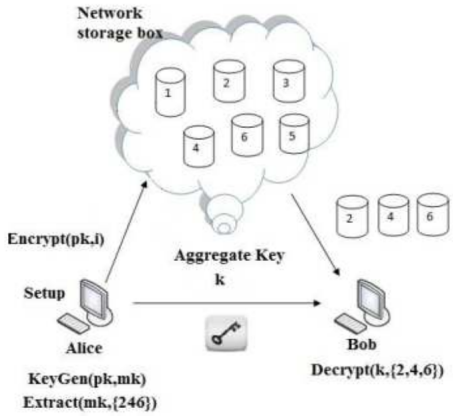


Fig. 8: Key aggregate cryptosystem for sharing data [58]

identifiers called a class on the encryption process. Also, the owner has a master key to create others secret keys for one, some classes of data, or all classes of ciphertext. Once the user gets his aggregate key, he only decrypts the class of ciphertext this key is for. It is an aggregate key where each part of it can decrypt part of the ciphertext.the whole key can decrypt the whole ciphertext. Therefore, this cryptosystem helps in sharing data among a group of users with fine grain access control and without giving them a key that can decrypt all that data. This figure8 shows the general view of this system.

A. Access control:

When data is outsourced to the cloud, which is untrusted because it is in a domain where security is not managed by the data owner, data security has to be given more attention. When more than one entity want to share data, there has to be a mechanism to restrict who can access that data. Many techniques have been discussed in the literature. Those techniques were proposed to keep data content confidential and keep unauthorized entity from accessing and disclosing the data by using access control while permitting many authorized entities to share those data. The following are some of the techniques that are in the literature.

B. Public Key Encryption

Public key encryption is used to encrypt the data by using the public key. Only the one who has the private key can decrypt this data. There are many issues that make this way hard to apply in the cloud when many people need to access those files.

In [60], Sana et el. proposed a lightweight encryption algorithm by utilizing symmetric encryption performance to encrypt files and utilizing asymmetric encryption efficient security to distribute keys. There are many disadvantages of using this method. One of them is key management issue and the need to get fine-grained access to file, such part of it. Also, this solution is not flexible and scalable because encryption and decryption is needed when a user leave the group in order

to prevent him from accessing the data. Key generation and encryption process is shown in figure 9

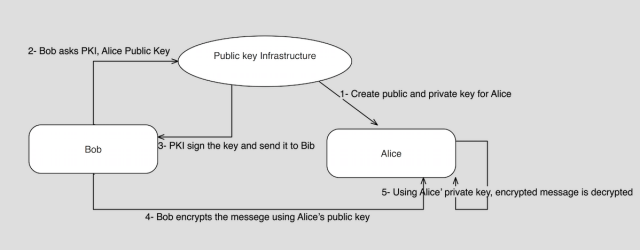


Fig. 9: Public Key Encryption

C. Identity-Based Encryption (IBE)

Shamir, in [61], has introduced identity-based encryption. The owner of data can encrypt his data by specifying the identity of the authorized entity to decrypt it based on that entity?s identity, which must match the one specified by the owner. Therefore, there is no key exchange. Encryption process is shown in figure 10

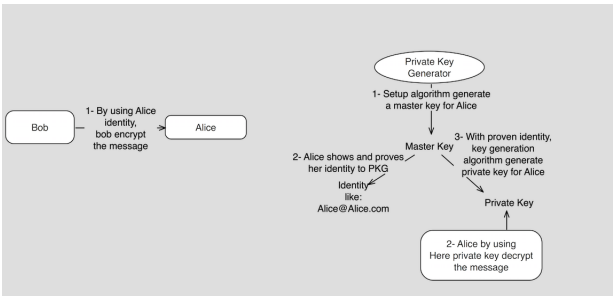


Fig. 10: Identity-Based Encryption (IBE)

D. Attribute Based Encryption (ABE)

In attribute based encryption, an identity of a user is identified by a set of attributes. This set of attributes generates the secret key. Also, it defines the access structure used for access control. This access control are using encryption to encrypt data for confidentiality and share it among group of users. It is a kind of integrating the encryption with the access control.

In [62], attribute-based encryption, know as fuzzy identity-based encryption, was proposed a few years after IBE. In this scheme, a group of attributes identify someone's identity. Data owner encrypts his data and only the one who has attributes that overlap with the attributes specified in the ciphertext can decrypt it. There are general schemes than ABE, which is based on trees. Key generation process is shown in figure 11 and encryption and decryption algorithm is shown in figure 12

1) Key Policy Attribute Based Encryption (KP-ABE): In [63], key policy attribute-based encryption was proposed. This is more general than ABE because it expresses more conditions than just matching the attributes to enforce more

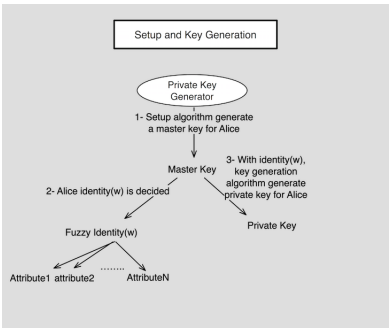


Fig. 11: Attribute Based Encryption (ABE)

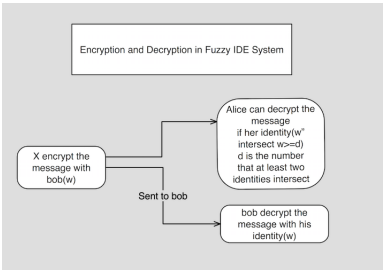


Fig. 12: Encryption\Decryption Attribute Based Encryption (ABE)

control. In this mechanism, ciphertext is linked with a set of attributes. The private key is linked to monotonic access structure. This access structure is based on a tree to specify the identity of the user. When the user?s private key has the attributes that satisfy the attribute in ciphertext, the user decrypts the ciphertext. Key generation process is shown in figure 13 and encryption and decryption algorithm is shown in figure 14. A disadvantage of this method is that the decriptor must trust the key generator to generate keys for a correct person with the right access structure. If the data needs to be re-encrypted, the new private keys have to be issued in order to keep accessing that data. Therefore, there is a need to get the policy associated with the key. Also, it does not support non-monotonic access structure which expresses negative attributes such 'not'.

In [64], Ostrovsky et al. propose a scheme that support non-monotonic access structure which supports positive and negative attributes. However, this scheme increases the size of ciphertext and key. Also, there is cost related to time needed for encryption and decryption. In KP-ABE, the size of ciphertext increases with the number of associated attributes linearly.

In [65], a scheme is proposed that results in constant size of ciphertext regardless of the number of attributes and supports non-monotonic access structure. However, the size of the key is quadratic size of number of the attributes. To overcome that disadvantage, a ciphertext policy attribute-based encryption was proposed. However, CP-ABE costs more than KP-ABE[66].

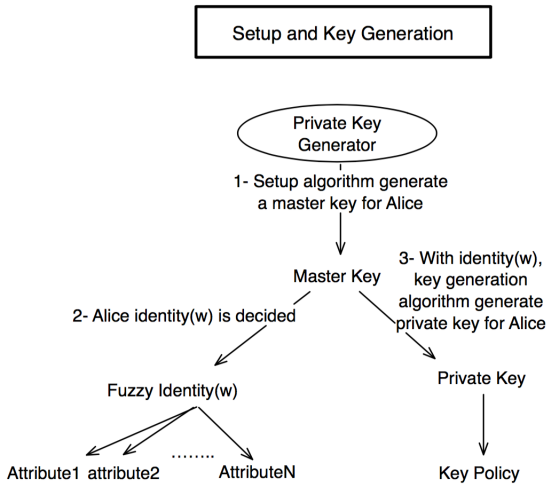


Fig. 13: Key Policy Attribute Based Encryption key Generation

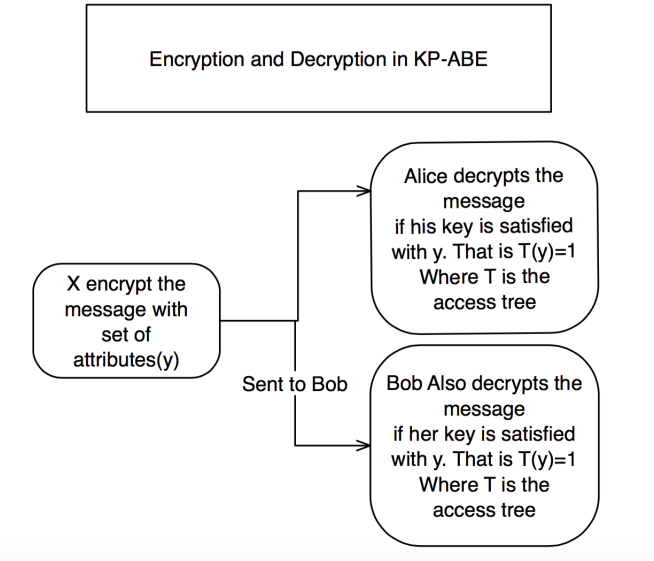


Fig. 14: KP-ABE encryption \decryption

2) Ciphertext Policy Attribute Based Encryption (CP-ABE): In [67], CP-ABE was proposed. In this scheme, the access structure, which is responsible for specifying the encryption policy, is associated with ciphertext. A private key for a user is created based on his attributes. A user can decrypt the ciphertext if the attributes in his private key satisfy the access structure in ciphertext. The benefit of making an access structure with ciphertext is that the encryptor can define the encryption policy and all already-issued private keys can not be changed unless the system is rebooted. There are four functions for the CP-ABE scheme. The four functions are as follows [67][68]. (MasterKey, PublicKey)=Setup(P): A trusted authority runs this function and it takes a security parameter(P)

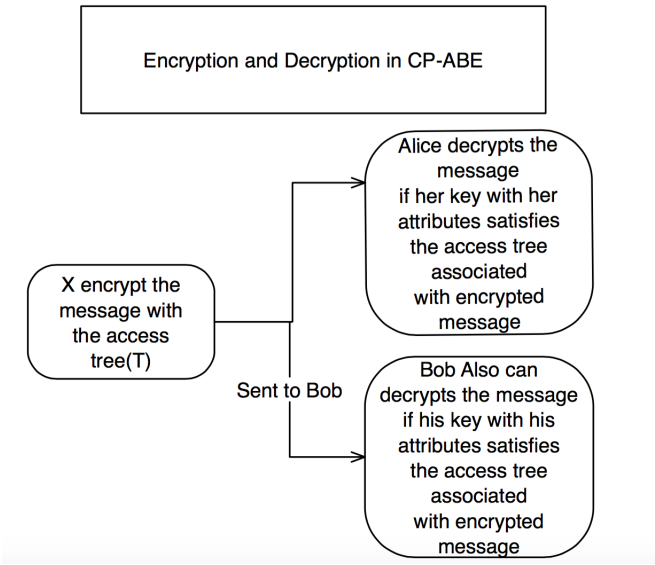


Fig. 15: KP-ABE encryption \decryption

as its input and master key (MK) and public key (PK) as its output.

SK=Key Generation(A,MK): A trusted authority runs this function and it takes a set of attributes (A) and Master Key (MK) as its input and its output is a secret key for a user associated with a set of attributes.

ciphertext (CT)=Encryption (M,MK,P): The data owner runs this function to encrypt his data. It takes a message (M), access control policy (P) and master public key (PK) as its inputs. Its output is a ciphertext under access control policy (P). Encryption algorithm is shown in figure 15

M=Decryption(ciphertext,SK) A decryptor who has the ciphertext runs this function. This ciphertext, under access policy (P) and secret key (SK), can be encrypted if and only if the access policy of the secret key overlap satisfies the access policy of the ciphertext and Its output is the original message. If it does not satisfy those conditions, the decryptor cannot get the original message. decryption algorithm is shown in figure 15.

XIII. Multi-Cloud Computing (MMC) issues

Cloud computing now is moving to multi-cloud computing because of security issues stemming from using a single cloud such data availability. This figure 16 shows how the clients connect to the clouds. Some of the issues that multi-cloud computing are data availability and security [70], Cachinet et al. said ”Services of single clouds are still subject to outage.? There is a fear among organizations that a single cloud would not fulfill their demands such as reliability and availability. Some organizations need the availability to be high and need their data to be far from being locked in. Therefore, they need a system that is always available and not under control of a single cloud provider. The notion of a multi-cloud will be become a trend in these years.

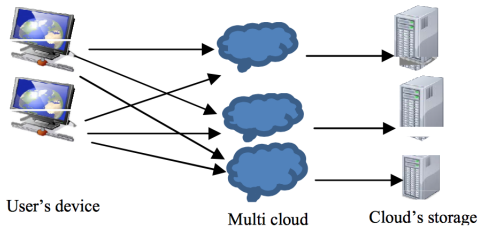


Fig. 16: Multi-cloud computing [69]

In [6], Alzain et al. have discussed many security issues in a single cloud and they are promoting the multi-cloud and its solutions to address single cloud security issues. They promised by using multi-cloud, valuable information such as credit card information and medical records could be protected from untrusted third parties and malicious insiders.

In [71], the authors said that moving from a single cloud to multi-cloud distributes trust, reliability, and security among multiple cloud providers. In addition to that, the users can avoid moving their data once they got locked in, by using another clouds to run their business.

In [72], Mahesh et al. suggests encrypting data, dividing it into chunks and storing those chunks in many cloud service providers. They insisted this would help to prevent all security issues of the cloud.

In [73], SUGANTHI et al. proposed a solution for protecting the privacy of the signer of that data from a third party auditor while auditing process. When an owner of data partions their data and sign them and distribute them to multi-clouds and share them with others, the third party could get the identity of the signer since it is needed when auditing. Therefore, they proposed this solution to prevent violating the privacy of the owner by knowing their identity by using creating homomorphic authenticators by using aggregate signatures[73]. Aggregate signature scheme is a group of signatures that are aggregated to one digital signature[74]. One Aggregate signature for n signatures of m messages that are from u users is the result of this scheme[74]. Therefore, the benefit of using it here is that the auditor will know the users how sign the messages but without knowing specifically how sign each message.

XIV. Mobile Cloud Computing

A. Limitations of mobile devices

With the advancement in mobile devices such as more processing, storage, memory, sensors and operating system capabilities, there is a limitation with regard to energy resources needed for complex computation. Some of the application in mobile devices are data-intensive or compute-intensive application. Due to battery life, the mobile device cannot run them. Therefore, the cloud computing is needed to run those complex computations. The mobile device's application augments the processing tasks to the cloud computing.

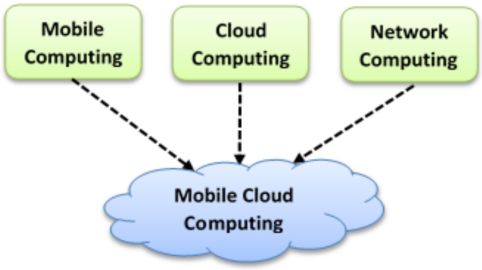


Figure 1. Evolution of Mobile Cloud Computing

Fig. 17: Mobile cloud computing [75]

B. Mobile Cloud Computing

Mobile cloud computing is using the mobile as front end and the cloud as back end for the storage and computation. In the figure 17, mobile cloud computing consists of mobile computing, cloud computing, and network.

In [76], three schemes are proposed for confidentiality and integrity of mobile device's files stored in the cloud. The first scheme is encryption based Scheme(EnS). In this scheme, the mobile device encrypts the file and gets its hash code. The encryption key is a concatenation of the password entered by a user, file name changed to bits and file size to defend brute force attack on a cloud server since the length of the password is limited. Only the file name is kept in the file and everything related to the file is deleted. When downloading the file from the cloud server, only the password is needed to decrypt the file. This process will need more processing on the mobile device side. They proved the confidentiality and integrity of the file using this scheme when it is stored in a distrusted clouds server. In order to overcome the power consumption in the first scheme, a coding based scheme is proposed. This scheme is not using encryption function since it consumes less power. The confidentiality of the file is protected by using matrix multiplication and the integrity is ensured by using hash-based message authentication code. The file is divided to many blocks and each block is divided to many chunks and each chunk in n bits. Each block represents matrix with chunks number as rows and bits as columns. a code victor matrix is created from the entered password. For confidentiality, each matrices are multiplied by the code victor matrix which result in secrecy code. For the integrity, all secrecy codes are concatenated and hashed. The result of the previous is the integrity key. The file is hashed with the integrity key which results in message authentication code. The third scheme is Sharing based Scheme(ShS) which applies X-OR operations on the file. This scheme needs less computational power. Hash-based message authentication code is used to verify the integrity of file while X-or operation is used to protect the confidentiality of the file.

In [77], Khan et al. propose a new scheme called block-based sharing scheme. This scheme overcomes all limitations of the previous schemes proposed in [76]. They use X-OR operation. First, they extend the password entered by a user

in order to be the same as block size. For example, the block size is 160 bit and the password entered by the user is 60 bits. In this case, they extend 60 bits to be 160 bits. Second, they divide a file to blocks with the same size. After that, they X-or the first block with first extended password. The second block is X-ORed with extended password after shifting each bit to the right. Therefore, each block is x-ORed with distinct password with size equal to the size of block. For integrity, they hash the concatenation of the file name, extended password and file size in order to get an integrity key. Then, they hash the file with the integrity key in order to get message authentication code. Once that done, only cipher text, message authentication code, and the hash of file name to the cloud. The hash of file name is sent for file retrieval. This scheme results in less energy consumption, memory utilization, and CPU utilization.

In [78], the authors used homomorphic encryption, multi-cloud computing and mobile. They used multiple cloud schemes for storing the data to avoid data lock in and used homomorphic encryption to run computations without down-loading the data back and forth between cloud computing and mobile to avoid the communication costs. Since encryption is expensive for the mobile devices, there are some propositions to avoid using it.

In [79], Bahrami1 et al. proposed a lightweight method for data privacy in mobile cloud computing. They used JEPG file as their case study because it is a common file in mobile. They divide the JEPG file into many splits, distribute them to many file based one predefined pattern, and scramble chunks randomly in each split file with help of psedue-random permutations with the chaos system. After that each file is sent to MCCs. For retrieval process, the split files are collected from MCCs. Each split chunks are rearranged by using the chaos system. After that, all split files is rearranged based pattern, predefined before. They used this method because it is low in computation and works effectively in the mobile. When they compared it with encrypting the JEPG in the mobile and sending it, they found their solution is more efficient. Their proposed method has two requirements: balancing computation overhead with maintaining the security and avoiding offloading the file to the mobile cloud computing for encryption by making the file is meaningless before sending it.

XV. Conclusion

Cloud computing is an emerging technology that will receive more attention in the future from industry and academia. The cost of this technology is more attractive when it is compared to building the infrastructure. However, there are many security issues coming with this technology as happens when every technology matures. Those issues include issues related to the previous issues of the internet, network issues, application issues, and storage issues. Storing data in a remote server leads to some security issues. Those issues are related to confidentiality of data from unauthorized people in remote sites, integrity of stored data in remote servers and the availability of the data when it is needed. Also, sharing data in cloud when the cloud service provider is mistrusted is an issue. However, we mentioned some techniques that protect data seen by the cloud service provider while it is shared among many users.Many studies have been conducted to discover the issues that affect confidentiality, integrity, and availability of data to find a solution for them. Those solutions will lead to more secure cloud storage, which will also lead to more acceptance from the people and the trust on the cloud will increase.

References

[1] S. Subashini and V. Kavitha, "A survey on security issues in service delivery models of cloud computing," *Journal of Network and Computer Applications*, vol. 34, no. 1, pp. 1 – 11, 2011.

[2] P. Mell and T. Grance, "The nist definition of cloud computing," 2011.

[3] M. T. Khorshed, A. S. Ali, and S. A. Wasimi, "A survey on gaps, threat remediation challenges and some thoughts for proactive attack detection in cloud computing," *Future Generation Computer Systems*, vol. 28, no. 6, pp. 833 – 851, 2012.

[4] Z. Zhou and D. Huang, "Efficient and secure data storage operations for mobile cloud computing," in *Proceedings of the 8th International Conference on Network and Service Management*. International Federation for Information Processing, 2012, pp. 37–45.

[5] K. Kumar and Y.-H. Lu, "Cloud computing for mobile users: Can offloading computation save energy?" *Computer*, no. 4, pp. 51–56, 2010.

[6] M. AlZain, E. Pardede, B. Soh, and J. Thom, "Cloud computing security: From single to multi-clouds," in *System Science (HICSS), 2012 45th Hawaii International Conference on*, Jan 2012, pp. 5490–5499.

[7] M. Sookhak, H. Talebian, E. Ahmed, A. Gani, and M. K. Khan, "A review on remote data auditing in single cloud server: Taxonomy and open issues," *Journal of Network and Computer Applications*, vol. 43, pp. 121–141, 2014.

[8] E. Aguiar, Y. Zhang, and M. Blanton, "An overview of issues and recent developments in cloud computing and storage security," in *High Performance Cloud Auditing and Applications*. Springer, 2014, pp. 3–33.

[9] I. Gul, M. Islam *et al.*, "Cloud computing security auditing," in *Next Generation Information Technology (ICNIT), 2011 The 2nd International Conference on*. IEEE, 2011, pp. 143–148.

[10] E. M. Mohamed, H. S. Abdelkader, and S. El-Etriby, "Enhanced data security model for cloud computing," in *Informatics and Systems (INFOS), 2012 8th International Conference on*. IEEE, 2012, pp. CC–12.

[11] S. Ramgovind, M. M. Eloff, and E. Smith, "The management of security in cloud computing," in *Information Security for South Africa (ISSA), 2010*. IEEE, 2010, pp. 1–7.

[12] F. Sabahi, "Cloud computing security threats and responses," in *Communication Software and Networks (ICCSN), 2011 IEEE 3rd International Conference on*. IEEE, 2011, pp. 245–249.

[13] X. Wang, B. Wang, and J. Huang, "Cloud computing and its key techniques," in *Computer Science and Automation Engineering (CSAE), 2011 IEEE International Conference on*, vol. 2. IEEE, 2011, pp. 404–410.

[14] S. Subashini and V. Kavitha, "A survey on security issues in service delivery models of cloud computing," *Journal of network and computer applications*, vol. 34, no. 1, pp. 1–11, 2011.

[15] J. Yang and Z. Chen, "Cloud computing research and security issues," in *Computational Intelligence and Software Engineering (CiSE), 2010 International Conference on*. IEEE, 2010, pp. 1–3.

[16] M. Lori, "Data security in the world of cloud computing," *Co-published by the IEEE Computer And reliability Societies*, pp. 61–64, 2009.

[17] C. Wang, K. Ren, W. Lou, and J. Li, "Toward publicly auditable secure cloud data storage services," *Network, IEEE*, vol. 24, no. 4, pp. 19–24, 2010.

[18] L. Wei, H. Zhu, Z. Cao, X. Dong, W. Jia, Y. Chen, and A. V. Vasilakos, "Security and privacy for storage and computation in cloud computing," *Information Sciences*, vol. 258, pp. 371–386, 2014.

[19] C. Wang, Q. Wang, K. Ren, and W. Lou, "Privacy-preserving public auditing for data storage security in cloud computing," in *INFOCOM, 2010 Proceedings IEEE*. Ieee, 2010, pp. 1–9.

[20] K. Yang and X. Jia, "Data storage auditing service in cloud computing: challenges, methods and opportunities," *World Wide Web*, vol. 15, no. 4, pp. 409–428, 2012.

[21] CSA, "The notorious nine cloud computing top threats in 2013," $The_Notorious_Nine_Cloud_Computing_Top_Threats_in_2013.pdf$.

[22] D. Hubbard and M. Sutton, "Top threats to cloud computing v1. 0," *Cloud Security Alliance*, 2010.

[23] W. Baker, "M," 2011 data breach investigations report,"." [Online]. Available: http://www.wired.com/images$_blogs/threatlevel/2011/04/Verizon$ $2011-DBIR_04-13-11.pdf$

[24] G. Brunette, R. Mogull *et al.*, "Security guidance for critical areas of focus in cloud computing v2. 1," *Cloud Security Alliance*, pp. 1–76, 2009.

[25] D. Catteddu, "Cloud computing: benefits, risks and recommendations for information security," in *Web Application Security*. Springer, 2010, pp. 17–17.

[26] H. Aljahdali, A. Albatli, P. Garraghan, P. Townend, L. Lau, and J. Xu, "Multi-tenancy in cloud computing," in *Service Oriented System Engineering (SOSE), 2014 IEEE 8th International Symposium on*, April 2014, pp. 344–351.

[27] T. Ristenpart, E. Tromer, H. Shacham, and S. Savage, "Hey, you, get off of my cloud: exploring information leakage in third-party compute clouds," in *Proceedings of the 16th ACM conference on Computer and communications security*. ACM, 2009, pp. 199–212.

[28] A. Aviram, S. Hu, B. Ford, and R. Gummadi, "Determinating timing channels in compute clouds," in *Proceedings of the 2010 ACM workshop on Cloud computing security workshop*. ACM, 2010, pp. 103–108.

[29] H. Hlavacs, T. Treutner, J.-P. Gelas, L. Lefevre, and A.-C. Orgerie, "Energy consumption side-channel attack at virtual machines in a cloud," in *Dependable, Autonomic and Secure Computing (DASC), 2011 IEEE Ninth International Conference on*. IEEE, 2011, pp. 605–612.

[30] K. Hashizume, D. G. Rosado, E. Fernández-Medina, and E. B. Fernandez, "An analysis of security issues for cloud computing," *Journal of Internet Services and Applications*, vol. 4, no. 1, pp. 1–13, 2013.

[31] W. A. Jansen, "Cloud hooks: Security and privacy issues in cloud computing," in *System Sciences (HICSS), 2011 44th Hawaii International Conference on*. IEEE, 2011, pp. 1–10.

[32] N. Gonzalez, C. Miers, F. Redígolo, M. Simplicio, T. Carvalho, M. Näslund, and M. Pourzandi, "A quantitative analysis of current security concerns and solutions for cloud computing," *Journal of Cloud Computing*, vol. 1, no. 1, pp. 1–18, 2012.

[33] M. H. Song, "Analysis of risks for virtualization technology," in *Applied Mechanics and Materials*, vol. 539. Trans Tech Publ, 2014, pp. 374–377.

[34] R. Bifulco, R. Canonico, G. Ventre, and V. Manetti, "Transparent migration of virtual infrastructures in large datacenters for cloud computing," in *Computers and Communications (ISCC), 2011 IEEE Symposium on*. IEEE, 2011, pp. 179–184.

[35] F. Zhang and H. Chen, "Security-preserving live migration of virtual machines in the cloud," *Journal of network and systems management*, vol. 21, no. 4, pp. 562–587, 2013.

[36] A. Corradi, M. Fanelli, and L. Foschini, "Vm consolidation: A real case based on openstack cloud," *Future Generation Computer Systems*, vol. 32, pp. 118–127, 2014.

[37] S. Fiebig, M. Siebenhaar, C. Gottron, and R. Steinmetz, "Detecting vm live migration using a hybrid external approach." in *CLOSER*, 2013, pp. 483–488.

[38] H. Wu, Y. Ding, C. Winer, and L. Yao, "Network security for virtual machine in cloud computing," in *Computer Sciences and Convergence Information Technology (ICCIT), 2010 5th International Conference on*. IEEE, 2010, pp. 18–21.

[39] S. Jin, J. Ahn, S. Cha, and J. Huh, "Architectural support for secure virtualization under a vulnerable hypervisor," in *Proceedings of the 44th Annual IEEE/ACM International Symposium on Microarchitecture*. ACM, 2011, pp. 272–283.

[40] C. Erway, A. Küpçü, C. Papamanthou, and R. Tamassia, "Dynamic provable data possession," in *Proceedings of the 16th ACM conference on Computer and communications security*. Acm, 2009, pp. 213–222.

[41] G. Ateniese, R. Burns, R. Curtmola, J. Herring, L. Kissner, Z. Peterson, and D. Song, "Provable data possession at untrusted stores," in *Proceedings of the 14th ACM conference on Computer and communications security*. Acm, 2007, pp. 598–609.

[42] G. Ateniese, R. Di Pietro, L. V. Mancini, and G. Tsudik, "Scalable and efficient provable data possession," in *Proceedings of the 4th international conference on Security and privacy in communication netowrks*. ACM, 2008, p. 9.

[43] K. Yang and X. Jia, "An efficient and secure dynamic auditing protocol for data storage in cloud computing," *Parallel and Distributed Systems, IEEE Transactions on*, vol. 24, no. 9, pp. 1717–1726, 2013.

[44] Q. Wang, C. Wang, K. Ren, W. Lou, and J. Li, "Enabling public auditability and data dynamics for storage security in cloud computing," *Parallel and Distributed Systems, IEEE Transactions on*, vol. 22, no. 5, pp. 847–859, 2011.

[45] C. Wang, Q. Wang, K. Ren, N. Cao, and W. Lou, "Toward secure and dependable storage services in cloud computing," *Services Computing, IEEE Transactions on*, vol. 5, no. 2, pp. 220–232, 2012.

[46] C. Wang, S. Chow, Q. Wang, K. Ren, and W. Lou, "Privacy-preserving public auditing for secure cloud storage," *Computers, IEEE Transactions on*, vol. 62, no. 2, pp. 362–375, Feb 2013.

[47] B. Balusamy, P. Venkatakrishna, A. Vaidhyanathan, M. Ravikumar, and N. Devi Munisamy, "Enhanced security framework for data integrity using third-party auditing in the cloud system," in *Artificial Intelligence and Evolutionary Algorithms in Engineering Systems*, ser. Advances in Intelligent Systems and Computing. Springer India, 2015, vol. 325, pp. 25–31.

[48] G. Ateniese, R. Burns, R. Curtmola, J. Herring, O. Khan, L. Kissner, Z. Peterson, and D. Song, "Remote data checking using provable data possession," *ACM Trans. Inf. Syst. Secur.*, vol. 14, no. 1, pp. 12:1–12:34, Jun. 2011.

[49] M. Sookhak, A. Gani, M. K. Khan, and R. Buyya, "Dynamic remote data auditing for securing big data storage in cloud computing," *Information Sciences*, 2015.

[50] A. Juels and B. S. Kaliski Jr, "Pors: Proofs of retrievability for large files," in *Proceedings of the 14th ACM conference on Computer and communications security*. Acm, 2007, pp. 584–597.

[51] S. Halevi, D. Harnik, B. Pinkas, and A. Shulman-Peleg, "Proofs of ownership in remote storage systems," in *Proceedings of the 18th ACM conference on Computer and communications security*. ACM, 2011, pp. 491–500.

[52] N. Kaaniche and M. Laurent, "A secure client side deduplication scheme in cloud storage environments," in *New Technologies, Mobility and Security (NTMS), 2014 6th International Conference on*, March 2014, pp. 1–7.

[53] F. S. Al-Anzi, A. A. Salman, N. K. Jacob, and J. Soni, "Towards robust, scalable and secure network storage in cloud computing," in *Digital Information and Communication Technology and it's Applications (DICTAP), 2014 Fourth International Conference on*. IEEE, 2014, pp. 51–55.

[54] K. D. Bowers, A. Juels, and A. Oprea, "Hail: a high-availability and integrity layer for cloud storage," in *Proceedings of the 16th ACM conference on Computer and communications security*. ACM, 2009, pp. 187–198.

[55] A. Bessani, M. Correia, B. Quaresma, F. André, and P. Sousa, "Depsky: dependable and secure storage in a cloud-of-clouds," *ACM Transactions on Storage (TOS)*, vol. 9, no. 4, p. 12, 2013.

[56] D. Chen, X. Li, L. Wang, S. Khan, J. Wang, K. Zeng, and C. Cai, "Fast and scalable multi-way analysis of massive neural data," *IEEE Trans. Comput*, vol. 63, 2014.

[57] A. N. Khan, M. M. Kiah, S. A. Madani, M. Ali, S. Shamshirband *et al.*, "Incremental proxy re-encryption scheme for mobile cloud computing environment," *The Journal of Supercomputing*, vol. 68, no. 2, pp. 624–651, 2014.

[58] P. S. Kumari, P. Venkateswarlu, and M. Afzal, "A key aggregate framework with adaptable offering of information in cloud," *International Journal of Research*, vol. 2, no. 3, pp. 5–10, 2015.

[59] C.-K. Chu, S. S. Chow, W.-G. Tzeng, J. Zhou, and R. H. Deng, "Key-aggregate cryptosystem for scalable data sharing in cloud storage," *Parallel and Distributed Systems, IEEE Transactions on*, vol. 25, no. 2, pp. 468–477, 2014.

[60] R. A. Sana Belguith, Abderrazak Jemai, "Enhancing data security in cloud computing using a lightweight cryptographic algorithm," *ICAS 2015 : The Eleventh International Conference on Autonomic and Autonomous Systems*, 2015.

[61] A. Shamir, "Identity-based cryptosystems and signature schemes," in *Advances in cryptology*. Springer, 1985, pp. 47–53.

[62] A. Sahai and B. Waters, "Fuzzy identity-based encryption," in *Advances in Cryptology–EUROCRYPT 2005*. Springer, 2005, pp. 457–473.

[63] V. Goyal, O. Pandey, A. Sahai, and B. Waters, "Attribute-based encryption for fine-grained access control of encrypted data," in *Proceedings of the 13th ACM conference on Computer and communications security*. Acm, 2006, pp. 89–98.

[64] R. Ostrovsky, A. Sahai, and B. Waters, "Attribute-based encryption with non-monotonic access structures," in *Proceedings of the 14th ACM conference on Computer and communications security*. ACM, 2007, pp. 195–203.

[65] N. Attrapadung, B. Libert, and E. De Panafieu, "Expressive key-policy attribute-based encryption with constant-size ciphertexts," in *Public Key Cryptography–PKC 2011*. Springer, 2011, pp. 90–108.

[66] Z. Qiao, S. Liang, S. Davis, and H. Jiang, "Survey of attribute based encryption," in *Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD), 2014 15th IEEE/ACIS International Conference on*, June 2014, pp. 1–6.

[67] J. Bethencourt, A. Sahai, and B. Waters, "Ciphertext-policy attribute-based encryption," in *Security and Privacy, 2007. SP'07. IEEE Symposium on*. IEEE, 2007, pp. 321–334.

[68] J. Hur and D. K. Noh, "Attribute-based access control with efficient revocation in data outsourcing systems," *Parallel and Distributed Systems, IEEE Transactions on*, vol. 22, no. 7, pp. 1214–1221, 2011.

[69] H. Hasan and S. Chuprat, "Secured data partitioning in multi cloud environment," in *Information and Communication Technologies (WICT), 2014 Fourth World Congress on*, 2014, pp. 146–151.

[70] C. Cachin, I. Keidar, and A. Shraer, "Trusting the cloud," *ACM SIGACT News*, vol. 40, no. 2, pp. 81–86, 2009.

[71] M. Vukolić, "The byzantine empire in the intercloud," *ACM SIGACT News*, vol. 41, no. 3, pp. 105–111, 2010.

[72] M. Shankarwar and A. Pawar, "Security and privacy in cloud computing: A survey," in *Proceedings of the 3rd International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA) 2014*, ser. Advances in Intelligent Systems and Computing. Springer International Publishing, 2015, vol. 328, pp. 1–11.

[73] J. Suganthi, J. Ananthi, and S. Archana, "Privacy preservation and public auditing for cloud data using ass in multi-cloud," in *Innovations in Information, Embedded and Communication Systems (ICIIECS), 2015 International Conference on*, 2015, pp. 1–6.

[74] D. Boneh, C. Gentry, B. Lynn, and H. Shacham, "Aggregate and verifiably encrypted signatures from bilinear maps," in *Advances in cryptologyEUROCRYPT 2003*. Springer, 2003, pp. 416–432.

[75] A. Donald and L. Arockiam, "A secure authentication scheme for mobicloud," in *Computer Communication and Informatics (ICCCI), 2015 International Conference on*, Jan 2015, pp. 1–6.

[76] W. Ren, L. Yu, R. Gao, and F. Xiong, "Lightweight and compromise resilient storage outsourcing with distributed secure accessibility in mobile cloud computing," *Tsinghua Science & Technology*, vol. 16, no. 5, pp. 520–528, 2011.

[77] A. N. Khan, M. M. Kiah, M. Ali, S. A. Madani, S. Shamshirband *et al.*, "Bss: block-based sharing scheme for secure data storage services in mobile cloud environment," *The Journal of Supercomputing*, vol. 70, no. 2, pp. 946–976, 2014.

[78] M. Louk and H. Lim, "Homomorphic encryption in mobile multi cloud computing," in *Information Networking (ICOIN), 2015 International Conference on*, Jan 2015, pp. 493–497.

[79] M. Bahrami and M. Singhal, "A light-weight permutation based method for data privacy in mobile cloud computing," in *Mobile Cloud Computing, Services, and Engineering (MobileCloud), 2015 3rd IEEE International Conference on*, March 2015, pp. 189–198.

Estimating the Parameters of Software Reliability Growth Models Using the Grey Wolf Optimization Algorithm

Alaa F. Sheta[†] and Amal Abdel-Raouf[†‡]
[†]Computers and Systems Department, Electronics Research Institute, Giza, Egypt
[‡]Computer Science Department, Southern Connecticut State University, USA

***Abstract*—In this age of technology, building quality software is essential to competing in the business market. One of the major principles required for any quality and business software product for value fulfillment is reliability. Estimating software reliability early during the software development life cycle saves time and money as it prevents spending larger sums fixing a defective software product after deployment. The Software Reliability Growth Model (SRGM) can be used to predict the number of failures that may be encountered during the software testing process. In this paper we explore the advantages of the Grey Wolf Optimization (GWO) algorithm in estimating the SRGM's parameters with the objective of minimizing the difference between the estimated and the actual number of failures of the software system. We evaluated three different software reliability growth models: the Exponential Model (EXPM), the Power Model (POWM) and the Delayed S-Shaped Model (DSSM). In addition, we used three different datasets to conduct an experimental study in order to show the effectiveness of our approach.**

***Index Terms*—Software Reliability, Reliability Growth Models, Grey Wolf Optimizer, Exponential Model, Power Model, Delayed S-Shaped Model**

I. Introduction

With the increasing importance of software systems in almost all aspects of our lives, there is a great need for the production of high quality software systems. The traditional model of software quality factors, suggested by McCall [1], consists of eleven different factors that should be considered in determining the quality of software. Subsequent models include Evans and Marciniak [2], which consists of twelve factors, and Deutsch and Willis [3], which consists of fifteen factors. All of these models incorporate reliability as one of the software quality factors. Software reliability is defined according to [4] as: "the probability, over a given period of time, that the system will correctly deliver services as expected by the user." A more precise definition of software reliability is given by [5], [6] as: "the probability of failure-free operation over a specified time, in a given environment, for a specific purpose." Unfortunately, the task of identifying and repairing software faults is costly. Moreover, the cost of finding the remaining faults increases as the number of faults decreases until the cost exceeds the benefit [7], [8]. Therefore, there is no software system that is failure-free, which is why the reliability requirements should be included in any software development contract. Software reliability is measured based on the maximum allowable rate of failure and can represent an entire system or one or more of its parts [9]–[11].

The cost of software development is always higher for more reliable systems. Consequently, the desired reliability should be determined depending on the criticalness of failure-free operation of the system. For example, the failure rate of a life-threatening system such as heart-monitor should be very low while a company website may have a higher failure rate.

In the literature, many methods are introduced to estimate and predict software reliability [12]–[20]. The proposed methods can be classified into two main categories [9]. The first category is Software Reliability Prediction Models and the second category is Software Reliability Growth Models (SRGM). Software reliability prediction models consider predicting the reliability early in the development life cycle. In the requirements, design or implementation phases, the model uses historical data and some quantitative measurements like Lines of code (LOC) and depth of nesting loops to estimate the failure rate. Examples of software reliability prediction models include the orthogonal defect classification model [21] and the constructive quality model [22], [23]. Some reliability models may be based on software architecture and others on modified adaptive testing [24], [25]. The second category, SRGM, represents how the system reliability changes over time during the testing phase and based on test data. SRGMs collect defect data and statistically correlate this data with known mathematical functions to predict software reliability [26]–[29].

Many SRGMs are proposed to represent the relationship between software reliability and time. SRGMs can be classified as either parametric or non-parametric models. The most famous parametric models are the Non-Homogeneous Poisson Process (NHPP) models used in [30]–[32]. Non-parametric models have less restricted assumptions as they can predict reliability based only on defect data [33]. Other SRGMs are introduced using Neural Networks in [14], [34], [35], using Bayesian learning in [36], [37] and using particle swarm optimization in [38], [39].

In this paper, we utilize the Grey Wolf Optimization (GWO) algorithm to predict faults during the software testing process using software faults historical data. The rest of this paper is organized as follows: In Section II, we briefly introduce

some SRGM models that we use in our study. Section III provides an overview of the GWO algorithm. Section IV shows the evaluation criterion adopted in this study. The experimental results developed for parameter estimation of software reliability are given in Section V. Finally, we provide the conclusions and future work in Section VI.

II. Reliability Growth Models (SRGM)

The inability to meet software requirements and/or deviation from the goal for which the software was developed is defined as software failure. Software reliability depends mainly on the way we handle failure. For example, detecting failure during execution and repairing it increases the reliability of the software as a function of time. This is what happens during the software testing process and before release of software to the market. Software reliability growth models (SRGMs) are the models concerned with the explanation and the description of software failures.

In the literature, many SRGMs were presented to estimate the reliability of software systems [40], [41]. Each SRGM assumes a function called $M(t)$ that measures the number of failures experienced at a given time t. The SRGM parameters are estimated based on either the failure times $t_1, t_2, \ldots$ or the times between failures $\Delta t_1, \Delta t_2, \ldots$. For a given software project, $\mu(t)$ represents the mean value function of a SRGM reflecting the expected number of failures experienced at time t. The derivative of the mean value function with respect to time, $\frac{d\mu(t)}{dt}$, is defined as the failure intensity $\lambda(t)$. In the following subsections, we briefly describe three well-known SRGM models that we use in our study.

A. Exponential Model (EXPM)

The exponential model was first provided in [5], [42]. This model is also known as the Goel-Okumoto exponential model [43] shown in Equations 1.

$$\begin{aligned} \mu(t;b) &= b_0(1-e^{-b_1 t}) \\ \lambda(t;b) &= b_0 b_1 e^{-b_1 t} \end{aligned} \tag{1}$$

The parameter b_0 is the expected total number of failures recovered at the end of the testing process (i.e. v_0). b_1 represents the rate at which the defect rate decreases (see Equation 2).

$$b_1 = \frac{\lambda_0}{v_0} \tag{2}$$

where the parameter λ_0 is the initial failure intensity and v_0 is the total failure at the end of the testing process.

B. Power Model (POWM)

The power model is also known as the Non-Homogeneous Poisson Process (NHPP) [44]. The equations that govern μ and λ are given in Equations 3.

$$\begin{aligned} \mu(t;b) &= b_0 t^{b_1} \\ \lambda(t;b) &= b_0 b_1 t e^{b_1 - 1} \end{aligned} \tag{3}$$

Many systems have adopted the NHPP model for analysis. For example in [45], author uses the NHPP to estimate software reliability for nuclear safety software. The Bayesian statistical inference (BSI) method was used to estimate the model parameters.

C. Delayed S-Shaped Model (DSSM)

This model is known as Yamada delayed S-shaped model [46], [47]. The model is a finite failure model. Yamada et al. [27] provided this model for error detection, in which the observed growth curve of the cumulative errors has an S-shape. The system equations for $\mu(t;b)$ and $\lambda(t;b)$ are given in Equation 4.

$$\begin{aligned} \mu(t;b) &= b_0(1-(1+b_1 t)e^{-b_1 t}) \\ \lambda(t;b) &= b_0 b_1^2 t^{-b_1 t} \end{aligned} \tag{4}$$

where b_0 is the expected total number of failures and b_1 represents the failure detection rate.

III. Grey Wolf Search Algorithm

The Grey Wolf Optimizer (GWO) is a meta-heuristics algorithm introduced by Mirjalili et al. [48]. The GWO is utilized to solve many optimization problems in different fields and successfully provides highly competitive results [49]–[52].

The GWO algorithm is based on the wild behavior of the grey wolves during hunting. According to the dominant hierarchy leadership order, the GWO divides the animals' population into four categories: alpha (α), beta (β), delta (δ), and omega (ω). Consequently, the optimization process, the same as the hunting, is guided by the highest rank leaders: α, β and δ respectively which represent the best three solutions in the search space. The ω wolves, the lowest in the hierarchical rank, represent the rest of the solutions that must adjust their positions to follow the other dominant wolves.

It is assumed that each candidate solution with dimension n is represented by the vector $\vec{X}$ such that the Grey wolf position vector is given as:

$$\vec{X} = \{x_1, x_2, \ldots, x_n\} \tag{5}$$

During the hunting process, the grey wolves surround the prey (i.e. solution of the problem). This surrounding behavior in GWO can be represented mathematically as follows:

$$\vec{D} = |\vec{C}.\vec{X}(t)_p - \vec{X}(t)| \tag{6}$$

$$\vec{X}(t+1) = \vec{X}_p(t) + \vec{A}.\vec{D} \tag{7}$$

where $\vec{X}(t)_p$ is the position vector of the prey, $\vec{X}(t)$ is the position vector of the Grey wolf, t is the current iteration, $\vec{A}$ and $\vec{C}$ are coefficient vectors that vary to allow the wolves to adjust their positions in the space around the prey. The coefficient vectors $\vec{A}$ and $\vec{C}$ are computed according to Equations 8.

$$\begin{aligned} \vec{A} &= 2\,\vec{a}\,\vec{r}_1 - \vec{a} \\ \vec{C} &= 2\,\vec{r}_2 \end{aligned} \tag{8}$$

It is given that the elements $\vec{a}$ are linearly decreasing from the value of 2 to the value of 0 over the search process and $\vec{r}_1, \vec{r}_2$ are random vectors selected in the domain of [0,1].

Then the GWO saves the best three solutions (alpha, beta and delta wolves) and allows the other solutions (omega wolves) to adjust their positions according to the positions of the best solutions. The following equations are used to calculate the distance between the current position and α, β, and δ, respectively (see Equations 9):

$$\begin{aligned} \vec{D}_\alpha &= |\vec{C}_1.\vec{X}_\alpha - \vec{X}| \\ \vec{D}_\beta &= |\vec{C}_2.\vec{X}_\beta - \vec{X}| \\ \vec{D}_\delta &= |\vec{C}_3.\vec{X}_\delta - \vec{X}| \end{aligned} \tag{9}$$

where $\vec{X}_\alpha, \vec{X}_\beta$ and $\vec{X}_\delta$ are the positions of the α, β and δ, respectively, $\vec{X}$ is the position of the current solution and $\vec{C}_1, \vec{C}_2$ and $\vec{C}_3$ are random vectors. Then, the final position of the current solution can be calculated as in Equation 10.

$$\begin{aligned} \vec{X}_1 &= \vec{X}_\alpha + \vec{A}_1.\vec{D}_\alpha \\ \vec{X}_2 &= \vec{X}_\beta + \vec{A}_2.\vec{D}_\beta \\ \vec{X}_3 &= \vec{X}_\delta + \vec{A}_3.\vec{D}_\delta \end{aligned} \tag{10}$$

Thus, $\vec{X}(t+1)$ can be computed as follows:

$$\vec{X}(t+1) = \frac{\vec{X}_1 + \vec{X}_2 + \vec{X}_3}{3} \tag{11}$$

where t represents the number of iterations and $\vec{A}_1, \vec{A}_2$ and $\vec{A}_3$ are random vectors that vary to allow the wolves to attack towards the prey. Finally, the hunting process ends when the grey wolves attack the prey after it stops moving. In the next sections we show how to utilize the GWO to estimate the parameters for number of SRGMs.

IV. Measure for Model Predictability

To make a comparison between different SRGMs it is important to measure the model accuracy in terms of some meaningful measurements. In our case we adopt the Goodness-of-fit criteria. These criteria are applied to measure the quality of the solution provided and determine the proximity of the estimated failures to the measured failures.

Assume we have N measurements which represent the cumulative number of failures found at time t_i where t_i is the accumulated execution time. Then $\mu(t_i, b)$ can be defined as the projected number of failure at time t_i by a model.

According to the Goodness-of-fit criterion, a curve corresponding to a selected model is fitted to all data points $t_i, \mu_i,\ i = 1, \ldots, n$; then the difference between the actual measured failures y and the estimated failures $\hat{y}$ based on the proposed model is compared and evaluated using the Variance-Accounted-For (VAF) and the Mean Magnitude of Relative Error (MMRE) [53].

$$VAF = [1 - \frac{var(y - \hat{y})}{var(y)}] \times 100\% \tag{12}$$

$$MMRE = \frac{1}{N} \sum_{i=1}^{N} \frac{|y_i - \hat{y}_i|}{y_i} \tag{13}$$

Another evaluation criterion that we use in our study is the correlation coefficient R that can be calculated using the following equation.

$$R = \frac{\sum_{i=1}^{N}(y_i - \overline{y})(\hat{y}_i - \overline{\hat{y}})}{\sqrt{\sum_{i=1}^{N}(y_i - \overline{y})^2 \sum_{i=1}^{N}(\hat{y}_i - \overline{\hat{y}})^2}} \tag{14}$$

Finally we use the mean square error as the evaluation criterion in our convergence behavior analysis as shown in the following section.

$$MSE = \frac{1}{N} \sum_{i=1}^{N} |y_i - \hat{y}_i| \tag{15}$$

V. Experiment Results

To develop our new technique for solving the problem of estimating the parameters of SRGM we used GWO MATLAB toolbox. We started by setting the number of search agents (grey wolves) and the maximum number of iterations for the experiment. From our experience we found that 30 agents and 50 iterations led to highly accepted results. The objective function gets the variables as a vector ($[x_1 x_2 ... x_n]$) and returns the objective value.

Our experiments explore the use of the GWO method to estimate the parameters of three software projects using three SRGMs. In each case, we estimate the model parameters for EXPM, POWM and DSSM models, generate the convergence curves using the GWO method and show the scattered plot.

A. Test/Debug Data 1

A real-time control application presented in [54], [55] is adopted as the first case study with a daily collected data. The real-time control application program has a size of 870 Kilo line of code (KLOC) of FORTRAN and a middle level language code. To estimate the model parameters b_0 and b_1 based on the GWO method, we needed to set up the search space. In our case, $b_0 \in [0, 500]$ and $b_1 \in [0, 1]$.

In Figure 1 (a), we show the actual and estimated accumulated failures curves for the EXPM, POWM and DSSM and the convergence behavior curves of the GWO process for the three developed models. A scattered plot of the three developed models is shown in Figure 1 (b).

Table I shows the estimated parameters for the SRGMs together with the model equations. In Table IV, we summarize the results of two evaluation criterion MMRE and VAF values for the three developed models EXPM, POWN and DSSM. In this case study, the DSSM model provided the best results in terms of VAF while the EXPM model's MMRE was the minimum in comparison to other POWM and DSSM.

B. Test/Debug Data 2

In the second case study, a real-time application [56] of a software system containing 200 modules of FORTRAN language was used to test our proposed methodology. The data consists of 111 measurements [55]. We ran the GWO to tune the parameters of EXPM, POWM and DSSM.

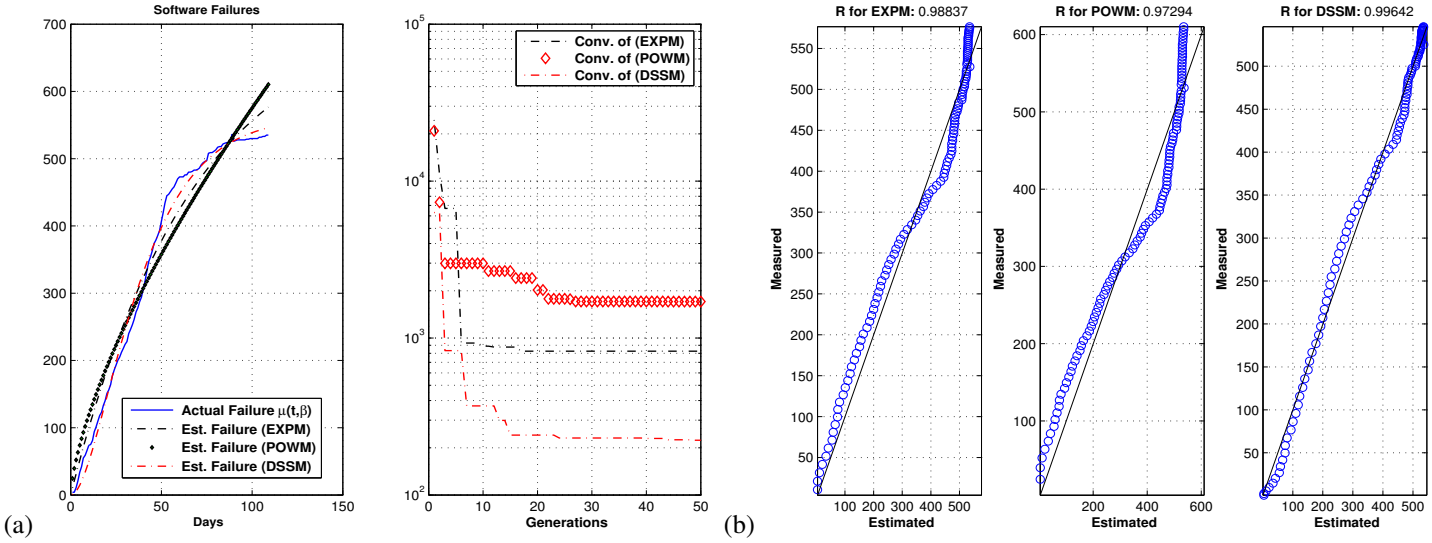


Fig. 1. (a) Actual and estimated failures and convergence curves for GWO (b) Scattered plot for the EXPM, POWM and DSSM using 109 Measurements

TABLE I
SRGMs WITH PARAMETER ESTIMATED USING GWO - 109 MEASUREMENTS

Exponential Model (EXPM)	$\mu(t;b)$ = 717.098 (1 - $e^{-0.01495539t}$)
Power Model (POWM)	$\mu(t;b)$ = 24.541 $t^{0.684974}$
Delayed S-Shaped Model (DSSM)	$\mu(t;b)$ = 562.3995 (1 - (1 + 0.04947323 t)$e^{-0.04947323t}$)

TABLE II
SRGMs WITH PARAMETER ESTIMATED USING GWO - 111 MEASUREMENTS

Exponential Model (EXPM)	$\mu(t;b)$ = 538.6468 (1 - $e^{-0.025683171t}$)
Power Model (POWM)	$\mu(t;b)$ = 30 $t^{0.625803}$
Delayed S-Shaped Model (DSSM)	$\mu(t;b)$ = 486.3256 (1 -(1 + 0.06691487 t)$e^{-0.06691487t}$)

In our case, $b_0 \in [0, 30]$ and $b_1 \in [0, 2]$. In Figure 2 (a), we show the actual and estimated accumulated failures curves for the EXPM, POWM and DSSM models and the convergence curves of the GWO process for the three developed models. A scattered plot of the three developed models is shown in Figure 2 (b).

Table II shows the estimated parameters for the SRGM models together with the model equations. The computed evaluation criterion are included in Table IV. Based on the developed experiments for this case, the results show that the DSSM model provided the best performance using the GWO tuned parameters as it has the minimum MMRE and the maximum VAF compared to other proposed models.

C. Test/Debug Data 3

In our third case study, we used a Test/Debug data set including 46 measurements as presented in [56]. We ran the GWO to find the best parameters to tune the EXPM, POWM and DSSM. In our case, $b_0 \in [0, 1000]$ and $b_1 \in [0, 1]$. In Figure 2 (a), we show the actual and estimated accumulated failures curves for the EXPM, POWM and DSSM models and the convergence curves of the GWO process for the three developed models. A scattered plot of the three developed models is shown in Figure 2 (b).

The estimated parameters for SRGMs are shown in Table III. In this case, the results show that the DSSM model was able to provide the best results in terms of MMRE while both the EXPM and the POWM models have better VAF values as shown in Table IV.

VI. CONCLUSION AND FUTURE WORK

In this paper, we proposed a GWO-based methodology to estimate the parameters of software reliability growth models (SRGMs). The estimated model parameters are used to predict the accumulated failures in a software system during the testing process. The problem is formulated for the GWO algorithm with the objective of minimizing the difference between the actual failures and the estimated accumulated failures.

Our methodology was employed to estimate the parameters of three adopted SRGMs: the exponential model, power model, and S-shaped model. Then the proposed models were applied to three real measured test/debug datasets. The results show that the proposed methodology is able to successfully estimate the parameters of SRGMs.

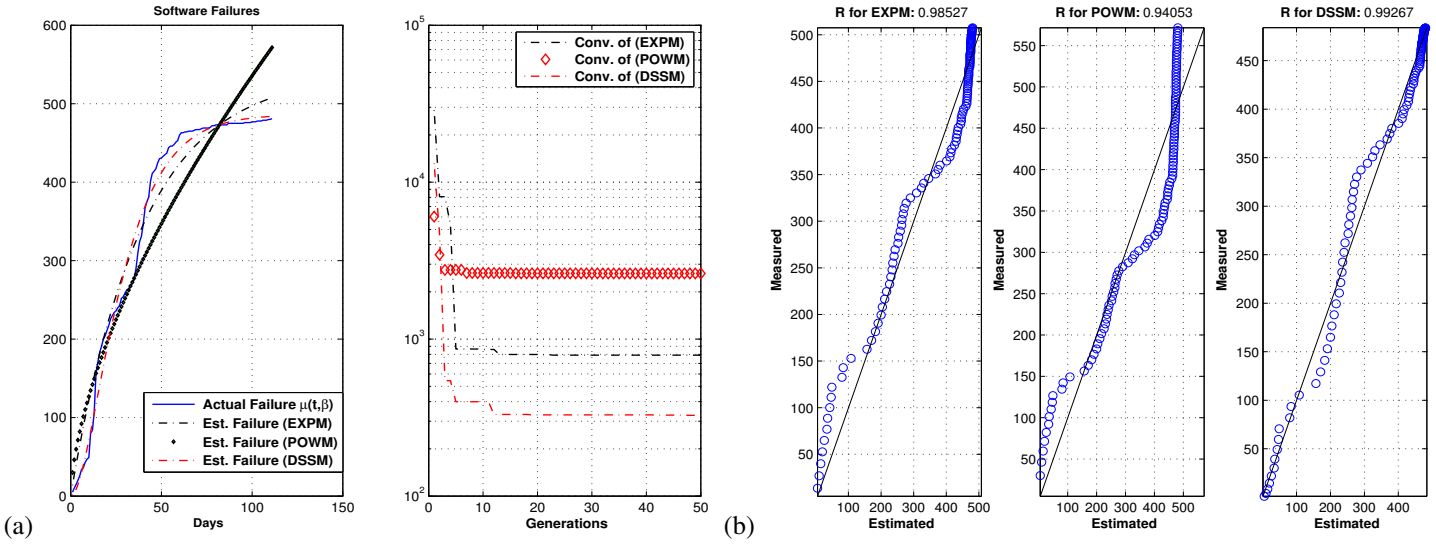


Fig. 2. (a) Actual and estimated failures and convergence curves for the GWO (b) Scattered plot for the EXPM, POWM and DSSM using 111 Measurements

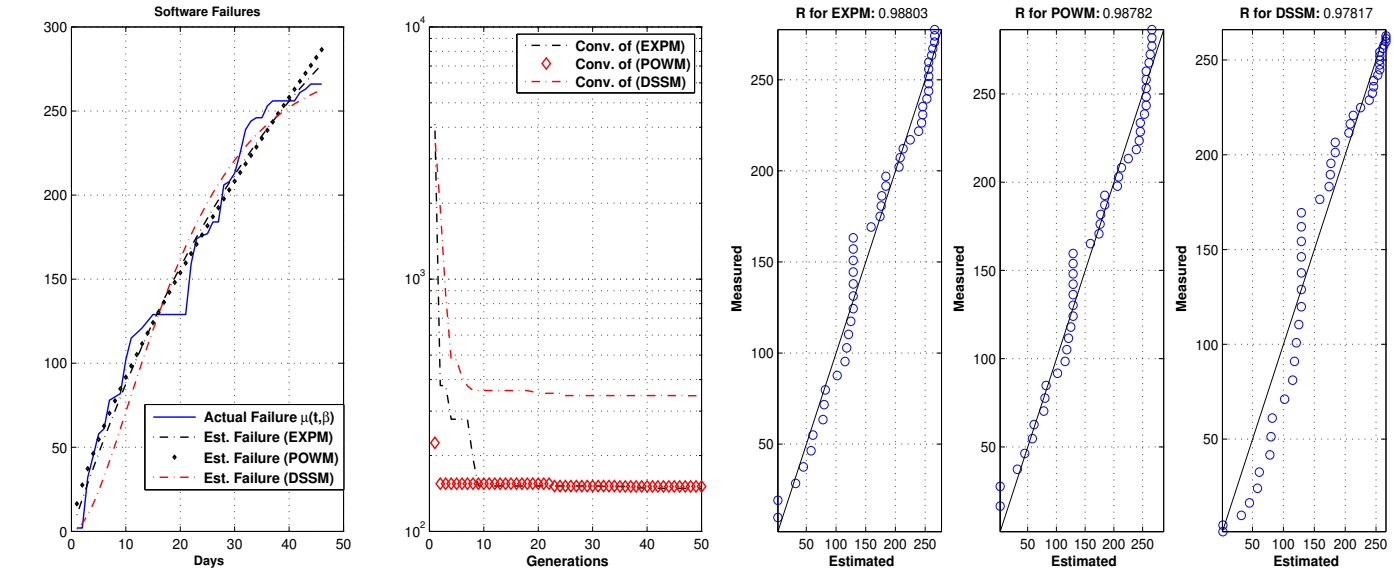


Fig. 3. (a) Actual and Estimated Accumulated failures and the convergence curves for the GWO (b) Scattered plot for the EXPM, POWM and DSSM using 46 Measurements

TABLE III
SRGMs WITH PARAMETER ESTIMATED USING GWO - 46 MEASUREMENTS

Exponential Model (EXPM)	$\mu(t;b)$ = 422.5453(1 - $e^{-0.02324815t}$)
Power Model (POWM)	$\mu(t;b)$ = 16.4506 $t^{0.746282}$
Delayed S-Shaped Model (DSSM)	$\mu(t;b)$ = 280.2617(1 - (1 + 0.09711093 t)$e^{-0.09711093t}$)

TABLE IV
EVALUATION RESULTS OF THE THREE MODELS USING GWO

	Test/Debug Data 1		Test/Debug Data 2		Test/Debug Data 3	
	MMRE	VAF%	MMRE	VAF%	MMRE	VAF%
EXPM	0.19027	97.347	0.19998	96.536	15.683	97.611
POWM	37.05	94.394	0.32297	88.343	22.736	97.566
DSSM	8.8572	99.268	0.072361	98.536	8.3919	94.701

For verification, a convergence behavior analysis was conducted. The results verify the effectiveness of the GWO algorithm to solve the problem with highly accepted performance. For future work, we plan to explore other techniques for modeling the software reliability growth based on other search algorithms in an effort to improve performance.

REFERENCES

[1] J. McCall, *Factors in Software Quality: Preliminary Handbook on Software Quality for an Acquisiton Manager*, vol. 1-3. General Electric, November 1977.

[2] M. W. Evans and J. J. Marciniak, *Software Quality Assurance and Management*. New York, USA: John Wiley and Sons, 1987.

[3] M. S. Deutsch and R. R. Willis, eds., *Software Quality Engineering, A Total Technical Management Approach, Ch.3*. Englewood Cliffs, NJ, USA: Prentice Hall, 1988.

[4] I. Sommerville, *Software Engineering: (Update) (8th Edition) (International Computer Science)*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 2006.

[5] J. Musa, "A theory of software reliability and its application," *IEEE Trans. Software Engineering*, vol. 1, pp. 312–327, 1975.

[6] J. Musa, A. Iannino, and K. Okumoto, *Software Reliability: Measurement, Prediction, Applications*. McGraw Hill, 1987.

[7] H. Pham, *Software Reliability*. Springer-Verlag, 2000.

[8] P. G. Bishop and R. Bloomfield, "Worst case reliability prediction on a prior estimate of residual defects," in *Proceedings of the 13th IEEE International Symposium on Software Reliability Engineering (ISSRE-2002)*, pp. 295–303, 2002.

[9] J. Musa, "Data analysis center for software: An information analysis center," *Western Michigan University Library, Kalamazoo, Michigan*, 1980.

[10] J. Musa, *Software Reliability Engineering: More Reliable Software, Faster and Cheaper*. Published Author House, 2004.

[11] J. Musa and L. A. Williams, "How should software reliability engineering be taught?," in *ISSRE*, p. 3, 2005.

[12] N. Karunanithi, D. Whitley, and Y. K. Malaiya, "Prediction of software reliability using connectionist models," *IEEE Trans. on Software Engineering*, vol. 18, no. 7, 1992.

[13] E. O. Costa, S. R. Vergilio, A. Pozo, and G. Souza, "Modeling software reliability growth with genetic programming," in *Proceedings of the 16th IEEE International Symposium on Software Reliability Engineering*, ISSRE '05, (Washington, DC, USA), pp. 171–180, IEEE Computer Society, 2005.

[14] S. H. Aljahdali, D. Rine, and A. Sheta, "Prediction of software reliability: A comparison between regression and neural network non-parametric models," in *Proceedings of the ACS/IEEE International Conference on Computer Systems and Applications*, AICCSA '01, (Washington, DC, USA), pp. 470–, IEEE Computer Society, 2001.

[15] R. Kumar, K. Khatter, and A. Kalia, "Measuring software reliability: A fuzzy model," *SIGSOFT Softw. Eng. Notes*, vol. 36, pp. 1–6, Nov. 2011.

[16] A. Amin, L. Grunske, and A. Colman, "An approach to software reliability prediction based on time series modeling," *J. Syst. Softw.*, vol. 86, pp. 1923–1932, July 2013.

[17] J. Wang, Z. Wu, Y. Shu, Z. Zhang, and L. Xue, "A study on software reliability prediction based on triple exponential smoothing method (wip)," in *Proceedings of the 2014 Summer Simulation Multiconference*, SummerSim '14, (San Diego, CA, USA), pp. 61:1–61:9, Society for Computer Simulation International, 2014.

[18] M. K. Bhuyan, D. P. Mohapatra, and S. Sethi, "A survey of computational intelligence approaches for software reliability prediction," *SIGSOFT Softw. Eng. Notes*, vol. 39, pp. 1–10, Mar. 2014.

[19] J. Pati and K. K. Shukla, "A hybrid technique for software reliability prediction," in *Proceedings of the 8th India Software Engineering Conference*, ISEC '15, (New York, NY, USA), pp. 139–146, ACM, 2015.

[20] M. K. Bhuyan, D. P. Mohapatra, and S. Sethi, "Measures for predicting software reliability using time recurrent neural networks with back-propagation," *SIGSOFT Softw. Eng. Notes*, vol. 40, pp. 1–8, Sept. 2015.

[21] R. Chillarege, I. S. Bhandari, J. K. Chaar, M. J. Halliday, D. S. Moebus, B. K. Ray, and M. Y. Wong, "Orthogonal defect classification-a concept for in-process measurements," *IEEE Transactions on Software Engineering*, vol. 18, pp. 943–956, Nov 1992.

[22] N. E. Fenton and M. Neil, "A critique of software defect prediction models," *IEEE Trans. Softw. Eng.*, vol. 25, pp. 675–689, Sept. 1999.

[23] F. Deissenboeck, E. Juergens, K. Lochmann, and S. Wagner, "Software quality models: Purposes, usage scenarios and requirements," in *Proceedings of the Seventh ICSE Conference on Software Quality*, WOSQ'09, (Washington, DC, USA), pp. 9–14, IEEE Computer Society, 2009.

[24] W.-L. Wang, D. Pan, and M.-H. Chen, "Architecture-based software reliability modeling," *J. Syst. Softw.*, vol. 79, pp. 132–146, Jan. 2006.

[25] H. Hu, C.-H. Jiang, K.-Y. Cai, W. E. Wong, and A. P. Mathur, "Enhancing software reliability estimates using modified adaptive testing," *Inf. Softw. Technol.*, vol. 55, pp. 288–300, Feb. 2013.

[26] S. H. Aljahdali and M. E. El-Telbany, "Genetic algorithms for optimizing ensemble of models in software reliability prediction," *Artificial Intelligence and Machine Learning (AIML)*, vol. 8, pp. 5–13, 6 2008.

[27] S. Yamada, *Software Reliability Modeling: Fundamentals and Applications*. Springer Publishing Company, Incorporated, 2013.

[28] N. R. Barraza, "A parametric empirical bayes model to predict software reliability growth," *Procedia Computer Science*, vol. 62, pp. 360 – 369, 2015. Proceedings of the 2015 International Conference on Soft Computing and Software Engineering (SCSE'15).

[29] L. K. Singh, G. Vinod, and A. K. Tripathi, "Early prediction of software reliability: A case study with a nuclear power plant system," *IEEE Computer*, vol. 49, no. 1, pp. 52–58, 2016.

[30] S. Kundu, T. K. Nayak, and S. Bose, *Statistical Models and Methods for Biomedical and Technical Systems*, ch. Are Nonhomogeneous Poisson Process Models Preferable to General-Order Statistics Models for Software Reliability Estimation?, pp. 137–152. Boston, MA: Birkhäuser Boston, 2008.

[31] P. Kapur, D. Goswami, A. Bardhan, and O. Singh, "Flexible software reliability growth model with testing effort dependent learning process," *Applied Mathematical Modelling*, vol. 32, no. 7, pp. 1298 – 1307, 2008.

[32] K.-Y. Cai, D.-B. Hu, C.-G. Bai, H. Hu, and T. Jing, "Does software reliability growth behavior follow a non-homogeneous poisson process," *Inf. Softw. Technol.*, vol. 50, pp. 1232–1247, Nov. 2008.

[33] Z. Wang, J. Wang, and X. Liang, "Non-parametric estimation for nhpp software reliability models," *Journal of Applied Statistics*, vol. 34, no. 1, pp. 107–119, 2007.

[34] S. Aljahdali, A. F. Sheta, and D. Rine, "Predicting accumulated faults in software testing process using radial basis function network models," in *Proceedings of the ISCA 17th International Conference Computers and Their Applications, April 4-6, 2002, Canterbury Hotel, San Francisco, California, USA*, pp. 26–29, 2002.

[35] H. Zeng and D. Rine, "Estimation of software defects fix effort using neural networks," in *28th International Computer Software and Applications Conference (COMPSAC 2004), Design and Assessment of Trustworthy Software-Based Systems, 27-30 September 2004, Hong Kong, China, Workshop Papers*, pp. 20–21, 2004.

[36] S. Wagner, "A bayesian network approach to assess and predict software quality using activity-based quality models," in *Proceedings of the 5th International Conference on Predictor Models in Software Engineering*, PROMISE '09, (New York, NY, USA), pp. 6:1–6:9, ACM, 2009.

[37] K. Jeet, R. Dhir, and H. Verma, "A comparative study of bayesian and fuzzy approach to assess and predict maintainability of the software using activity-based quality model," *SIGSOFT Softw. Eng. Notes*, vol. 37, pp. 1–9, May 2012.

[38] A. Sheta, "Reliability growth modeling for software fault detection using particle swarm optimization," in *Proceedings of the 2006 IEEE Congress on Evolutionary Computation (CEC2006)*, pp. 3071–3078, 2006.

[39] A. Sheta, "Parameter estimation of software reliability growth models by particle swarm optimization," *Artificial Intelligence and Machine Learning (AIML)*, vol. 7, pp. 55–61, 9 2007.

[40] M. Xie, "Software reliability models - past, present and future," *In N. Limnios and M. Nikulin (Eds). Recent Advances in Reliability Theory: Methodology, Practice and Inference*, pp. 323–340, 2002.

[41] S. Yamada, "Software reliability models and their applications: A survey," in *International Seminar on Software Reliability of Man-Machine Systems - Theories Methods and Information Systems Applications - August 17-18, Kyoto University, Kyoto, Japan*, 2000.

[42] P. B. Moranda, "Predictions of software reliability during debugging," in *Proceedings of Annual Reliability and Maintainability Symposium*, pp. 327–332, 1975.

[43] A. Geol and K. Okumoto, "Time-dependent error-detection rate model

for software reliability and other performance measures," *IEEE Trans. Reliability*, vol. 28, pp. 206–211, 1979.

[44] L. H. Crow, "Reliability for complex repairable systems," *Reliability and Biometry, SIAM*, pp. 379–410, 1974.

[45] G.-Y. PARK and S. C. JANG, "A software reliability estimation method to nuclear safety software," *Nuclear Engineering and Technology*, vol. 46, no. 1, pp. 55 – 62, 2014.

[46] S. Yamada, M. Ohba, and O. S., "S-Shaped reliability growth modeling for software error detection," *IEEE Trans. Reliability*, pp. 475–478, 1983.

[47] S. Yamada, M. Ohba, and O. S., "S-Shaped software reliability growth models and their applications," *IEEE Trans. Reliability*, pp. 289–292, 1984.

[48] S. Mirjalili, S. M. Mirjalili, and A. Lewis, "Grey wolf optimizer," *Adv. Eng. Softw.*, vol. 69, pp. 46–61, Mar. 2014.

[49] M. H. Sulaiman, Z. Mustaffa, M. R. Mohamed, and O. Aliman, "Using the gray wolf optimizer for solving optimal reactive power dispatch problem," *Appl. Soft Comput.*, vol. 32, pp. 286–292, July 2015.

[50] S. Mirjalili, "How effective is the grey wolf optimizer in training multi-layer perceptrons," *Applied Intelligence*, vol. 43, pp. 150–161, July 2015.

[51] N. Jayakumar, S. Subramanian, S. Ganesan, and E. B. Elanchezhian, "Combined heat and power dispatch by grey wolf optimization," *International Journal of Energy Sector Management*, vol. 9, no. 4, pp. 523–546, 2015.

[52] E. Emary, H. M. Zawbaa, C. Grosan, and A. E. Hassenian, *Afro-European Conference for Industrial Advancement: Proceedings of the First International Afro-European Conference for Industrial Advancement AECIA 2014*, ch. Feature Subset Selection Approach by Gray-Wolf Optimization, pp. 1–13. Cham: Springer International Publishing, 2015.

[53] M. Shin and A. L. Goel, "Empirical data modeling in software engineering using radial basis functions," *IEEE Transactions on Software Engineering*, vol. 26, no. 6, pp. 567–576, 2000.

[54] T. Minohara and Y. Tohma, "Parameter estimation of hyper-geometric distribution software reliability growth model by genetic algorithms," in *Proceedings of the 6th International Symposium on Software Reliability Engineering*, pp. 324–329, 1995.

[55] A. Sheta, "Reliability growth modeling for software fault detection using particle swarm optimization," in *2006 IEEE Congress on Evolutionary Computation, Sheraton, Vancouver Wall Centre, Vancouver, BC, Canada, July 16-21, 2006.*, pp. 10428–10435, 2006.

[56] Y. Tohman, K. Tokunaga, S. Nagase, and M. Y., "Structural approach to the estimation of the number of residual software faults based on the hyper-geometric distribution model," *IEEE Trans. on Software Engineering*, pp. 345–355, 1989.

Impact of IP Addresses Localization on the Internet Dynamics Measurement

Tounwendyam Frédéric Ouédraogo
University of Koudougou
UFR-Sciences et Technique
Avenue M. Yameogo BP 376 KDG, Burkina Faso

Tonguim Ferdinand Guinko
Université de Laval
Facult´e des Sciences et de G´enie
1045 avenue de la M´edecine, Qu´ebec
Canada

***Abstract*—Many projects have sought to measure the dynamics of the Internet by using end-to-end measurement tools. The RADAR tool has been designed in this context. It consists in periodically tracing the routes from a monitor toward a set of destinations, IP addresses chosen randomly in the Internet. However, the localization of these destinations on the topology has a significant influence on the observed dynamics.**

We study the dynamics observed when the destinations are localized at a country scale. We show that this localization may lead to observe a different dynamics. The local dynamics observed in our case is mainly a routing dynamics whereas the load-balancing dominates the entire Internet dynamics.

***Keywords*—Networks; Internet; Dynamics; Measurement; Localization**

I. Introduction

One of the main challenges in studying the Internet is the topology dynamics. However, understanding Internet dynamics remains crucial for many applications, including measurement tools, network protocols.

Recent studies have revealed important results on the measurement and characterization of the Internet dynamics [19], [13], [22], [13], [6], [5], [25]. Earlier work on the Internet dynamics has concerned mainly the measurement challenge. One of the important results has introduced the ego-centered views of the Internet dynamics and provides measurement tool and data [16].

The ego-centered view approach to study the Internet dynamics at IP-level topology consists in focusing on what a single monitor can see of the entire Internet dynamics. This manner to measure the Internet topology performs periodically end-to-end measurements from the monitor to the destinations and provides a time series of routing trees. This approach to study the dynamics has allowed obtaining important results on the characterization of the Internet dynamics [17], [19].

The set of destinations is comprised of IP addresses chosen randomly among all those which are on the Internet. The observed dynamics with the ego-centered view approach relies on where the destinations are located on the topology.

This paper addresses the issues of the impact of the destinations localization on the observed dynamics. We analyze the observed dynamics when the destinations are in a restricted area of the topology, precisely at the scale of a country.

We have performed, in the same period two kinds of measurements of the Internet dynamics, the one at country scale and the other at global scale. We found different routing topologies and dynamics. Surprisingly, we observe in the local measurement that the routing topology has more branching on the paths towards the destinations and weak dynamics mainly due to the routing changes, whereas in global measurement, the routing topology is filiform with high dynamics due to the load-balancing. The main contribution of this paper to the research community is to highlight the heterogeneity of the routing dynamics at IP level.

The rest of the paper is organized as follows. The section II presents the measurement framework and the dataset. In section III, we make an analysis comparatively of the local and global routing topologies. Section IV presents the results the investigation on the impact of the destinations localization on the observed Internet dynamics through the local and global measurements.

II. Dataset

Our work relies on the data of Internet dynamics obtained from the RADAR measurement. A RADAR measurement consists in periodical TRACETREE measurements done from a single monitor toward a set of destinations. A TRACETREE measurement is the outcome of parallel end-to-end measurements toward the destinations. Then TRACETREE provides a routing tree where the root is the monitor and the leaves are the destinations. The RADAR measurement represents regular snapshots of the routing topology at IP-level around the monitor.

There are RADAR data measurements performed from about hundred monitors mainly from PLanetLab[1] and publicly available[2]. These measurements are done with parameters to whom the relevance has been shown by these authors [16]. Among these parameters there are 3 000 destinations, 10 minutes between two consecutive TRACETREE measurements, random choice of destinations and monitors. Obviously, other parameters may be also relevant.

We performed two kinds of RADAR measurements, in the same period of time and on the same monitor but with different sets of destinations. The first set of destinations composed of

[1] https://www.planet-lab.org/
[2] http://data.complexnetworks.fr/Radar/

3 000 IP addresses chosen randomly among all the IP addresses available on the Internet[3], aims to measure the entire topology dynamics. Next, we will refer to this measurement as the global measurement. The second set of destinations also made of 3 000 IP addresses but chosen randomly among the IP addresses assigned to a country. The measurement is restricted to the country scale. We will refer to the measurement with the second set of destinations as the local measurement. We conducted the global and local measurements from several monitors located at different countries and we made similar observations.

In this paper, we present our results from a monitor located in France (Paris). The measurement has lasted five months. We obtained from this monitor 15 375 rounds of global measurement and 15 620 with the local measurement. Each round measurement lasted around 4 minutes, with 10-minute break between two consecutive rounds.

III. Snapshots of the Topology

The ego-centered views approach to measure the Internet allows to map in short time the topology and provides a snapshot of the routing topology around the monitor as routing tree. In this section, we present how the localization of the destinations influences the routing tree.

Figure 1 shows a large difference between local and global IP addresses although they use the same number of destinations. Moreover the number of destinations reached is roughly the same with local and global measurements. These observations allow saying that the local and global topologies around the same monitor are different. But notice that two rounds measurement may have the same number of IP addresses with large difference of their topologies.

The local measurement obtains approximately the same number of IP addresses. It is not the case of the global where we observe more fluctuations. This fluctuation shows some dynamics of the topology with the global measurement that are not observed with the local, see section IV.

Before going further to analyze the difference of their topologies, let us show that the distances of the destinations can explain the difference on the number of IP addresses observed with the local and global measurements. The distance D_{sd} from the monitor s to a given destination d is the path length with the extremities s and d.

$$D_{sd} = \sum_{i=1}^{k}(n_i, n_{i+1}) \text{ where } n_1 = s,\ n_{k+1} = d.$$

We compute the distance average $\overline{D} = \frac{1}{N}\sum_j D_{sd_j}$ of distances D_{sd_j} of destinations d_j.

Figure 2 shows the average distances of the local and global measurements. The global distance is larger than the local distance over time.

The difference of average distances means that the global destinations are farther from the monitor than the local destinations, therefore routes from the monitor to the global destinations gather more IP addresses. The routes towards

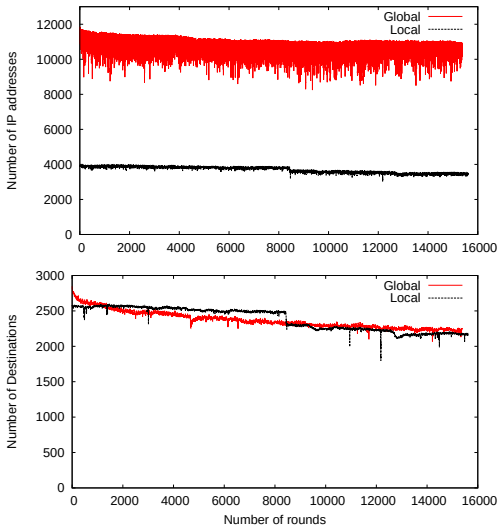


Fig. 1. Top: Evolution of the number of IP addresses seen at each round of local and global measurements. Bottom: Evolution of the number of destinations reached at each round of local and global measurements

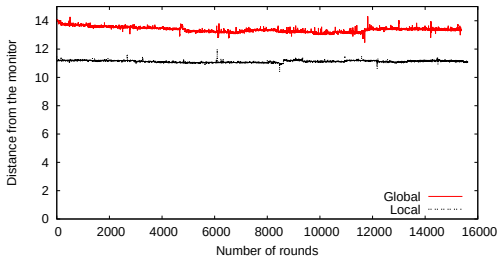


Fig. 2. Average distance of the destinations reached over time.

local destinations in the routing tree are less disjoint and they have a large part of their paths in common. Obviously, most of them are more close to the monitor. A rough estimation of the number of IP addresses induced by the difference of the route lengths (approximately 3), proves the fact that the global measurement observes more IP addresses than the local measurement.

The degree of vertices[4] is an important feature of the topology of trees and more generally complex networks. We use the degree of vertices to define a quantity that characterize the routing topology of the local and global. We denote $T = (V, E)$ a tree, V the set of vertices and E the set of edges. We define a function $\mathcal{F}_T$ to measure the filiform level of the tree T, characterized by vertices of degree 2 *i.e.* vertices having a single successor in the tree. We denote $e_{v_i^2 v_j^2}$ an edge where extremities v_1 and v_2 have degree less or equal to 2,

$$\mathcal{F}_T = \sum_{v_i, v_j \in V} e_{v_i^2 v_j^2}$$

. The more there are many edges $e_{v_i^2 v_j^2}$, the more the tree is filiform.

[3] IP addresses that reply to PING request.

[4] A degree of a vertex v is the number of edges connected to v.

Conversely, a tree with many branches (vertex with degree more than 2) is less filiform. We define the number of triples as a quantity that measures the branch-level in the tree. A triple is defined as a couple of edges having a same extremity in the tree. The more there are triples, the more there are branches in the tree.

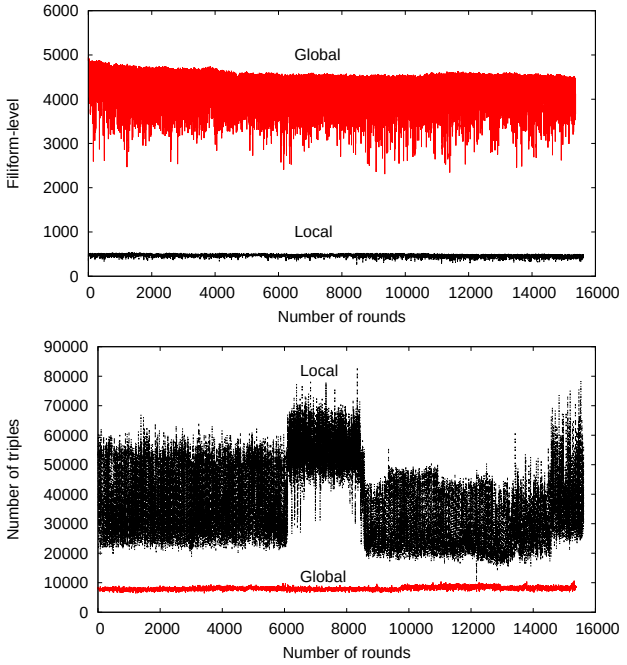


Fig. 3. Top: Filiform $\mathcal{F}_T$ of the routing trees of the local and global over time. Bottom: Number of triples in the routing trees of the local and global over time.

Figure 3 shows the evolution of the two quantities, the filiform function and the number of triples of the local and global routing trees. It seems obvious that when the number of triples in the tree is high, the filiform becomes low because they are opposed. However, we note that the routing tree of global measurement is quite different from the local routing tree. The global routing tree has a high level of filiform and has small number of triples while conversely, the local routing tree has low level of filiform and a high number of triples. We observe a clear fluctuation of the global filiform and not in the local because the scale effect flattened the variations of the local filiform small value. it is the same for the number of triples.

The high number of triples of the local routing tree means that most routers have several paths leading to the destinations and most leaves are near these routers, just few hops. Therefore, the paths towards these destinations have a large part in common that reduces the possibility to observe more IP addresses. This explains why the local measurement gets less IP addresses than global measurement.

The high level of filiform of the global routing tree shows that the routes from the monitor toward the destinations diverge rapidly. Most of the destinations have a large part of the paths isolated.

In summary, the local and global sets of destinations lead to different routing topologies around the monitor. This result is particularly important in the case of ego-centered views approach that is inherent local measurement.

IV. Dynamics characterization

There are two main properties that characterize the observed dynamics at IP-level topology. The first property is the high pace of discovery of the IP addresses. Indeed, new IP addresses appear until the end of the measurement with a more sustained pace than expected [17]. The second property concerns the dynamics pattern of the topology at IP-level. The dynamics of the IP addresses observed around the monitor reveals a parabolic shape. The dynamics pattern relies on the occurrences of the IP addresses observed during measurement [19].

A. *Pace of appearance of* IP *addresses*

This property of the dynamics has been observed in many measurements of the Internet topology. *New* IP addresses appear sustainably until the end of measurement. How many of these IP addresses existed on Internet before being observed by the measurement? Investigate on this issue is out of the scope of this paper. In this work we consider as new, all the IP addresses which appear for the first time.

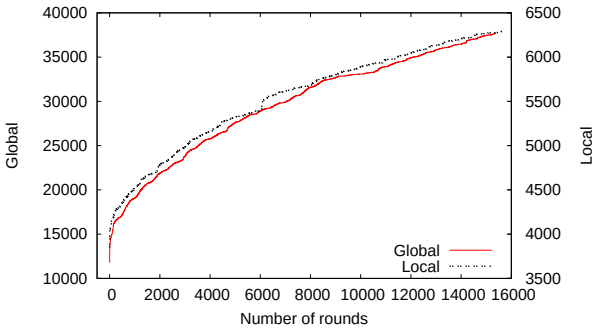


Fig. 4. Evolution of the number of new IP addresses observed during the measurement. Left y-axis: global measurement. Right y-axis: local measurement.

Figure 4 shows the number of new IP addresses of the local and global measurements. We do not observe significant difference between the local and global measurements. In both case new IP addresses are discovered with a sustained pace, except some fluctuations corresponding to rounds of measurement where new IP addresses did not appear. The local measurement observes less IP addresses than the global measurement but proportionally they discover the new IP addresses at the same pace. For example, in the last day of measurement the local measurement sees 10 IP addresses and the global measurement 108 IP addresses.

As the routes of the global destinations are longer than those of the local destinations, they are probably less stable [5], thus allows discovering more IP addresses. The local destinations are more close to the monitor and the dynamics captured is mainly near the monitor. However, the appearance

[5]A route in Internet is stable when it does not change over time. The same sequence of IP addresses is discovered between the monitor and destination.

of new IP addresses also remains sustained until the end of the measurement.

There is no significant difference between the local and global measurements regarding the pace of appearance of new IP addresses.

B. Dynamics pattern

The pattern of the dynamics around a single monitor is given by the correlation between the number of occurrences and the number of blocks of consecutive occurrences of the IP addresses. The number of occurrences of an IP address is the number of rounds in which it appears. If it was observed in all rounds, there is one block of consecutive occurrences. Otherwise the number of consecutive occurrences is the number of blocks, as illustrated in Figure 5.

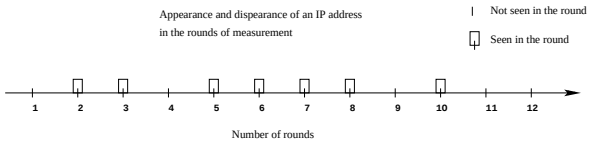


Fig. 5. Representation of the number of occurrence and the number of blocks of an IP address in 12 rounds measurement. The IP address has 7 occurrences (it appears in $2, 3, 5, 6, 7, 8$ and 10) There are 3 blocks of consecutive occurrences. The first block is rounds 2 and 3, the second block is rounds 5, 6, 7 and 8 and the last block is round 10.

Figure 6 shows the pattern produced by the correlation between the numbers of occurrences and the numbers of blocks of the IP addresses of the local and global measurements. This pattern gives a view of the dynamics produced by the appearance and disappearance of the IP addresses.

It appears a parabolic shape which limits for the points inside. The parabolic curve has two tangent lines related to the definition of blocks and occurrences. The first tangent line $(y = x)$ means that the number of blocks of an IP address cannot exceed its occurrences number, see Figure 5. The points closed to this tangent are the IP addresses which tend not to be seen in two consecutive rounds. This blinking trend of IP addresses in the measurement is due to the load-balancing routers [6] [1], [2].

The second tangent line $(y = M - x)$ (M the number of rounds) means that the number of occurrences is less than the total number of rounds. The IP addresses close to this tangent tend to be observed consecutively, without interruption throughout the measurement.

The distribution of points is not uniform. It clearly appears many clusters, in the global measurement, see Figure 6. We note that the most significant is the cluster around the line $y = x/12$ and to a lesser extent around the line $y = x/2$.

As already explained in [19], a load-balancing router spreads the traffic among c paths, each IP address belonging to any of these paths has a probability $p = 1/c$ of being seen at each round, leading to the number of occurrences equal to np approximately.

[6]A load-balancing router spreads the traffic among several paths, per-packet , per-destination

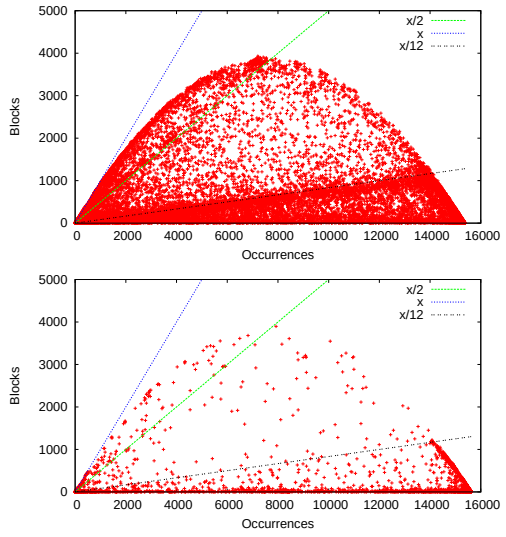


Fig. 6. Dynamics pattern. Each point represents an IP address, obtained by the number of occurrences on the x-axis and the number of blocks on the y-axis. Top: global measurement. Bottom: local measurement.

An IP address belonging to load-balanced paths can have any probability p of being seen. Therefore, the cluster around the line $y = x/12$ corresponds to IP addresses having a probability $p = 1/12$ of being seen. These IP addresses are observed after load-balancing routers that spread the traffic between 12 paths. Similarly, the cluster around the line $y = x/2$ to IP addresses observed after load-balancing routers that spread the traffic between 2 paths.

These clusters do not appear in the local dynamics, see Figure 6 (Bottom). This means that the local measurement has less load-balancing routers than the global measurement.

We are resuming the explanations of the parabola of these authors [19]. Considering a given round as the first of a consecutive blocks of occurrences of an IP address, with the probability p that this IP address was seen in this round, multiplied by the probability $1 - p$ that it was not seen in the previous round. Multiplying these two probability by n, gives the expected number of blocks $np(1 - p)$, which is the equation of parabola. The explanation is simple and formal. But in some cases, an IP address may be on the path of several load-balancing routers.

The IP addresses at the end of the parabola tend to be seen in consecutive rounds. This kind of dynamics is not due to the load-balancing but to the routing changes. The parabolic shape does not appear clearly in the local dynamics, except the beginning and the end of the parabola. There are many points close to the x-axis. These points correspond to the IP addresses observed in consecutive rounds.

The dynamics observed at IP-level shows different patterns with regard to the destinations localization. In our case, the localization of the destinations at a country scale shows that the load-balancing effect is not observed around the monitor. The local dynamics is mainly due to the routing changes.

V. Related work

The Internet dynamics measurement passes by an efficient and fast method of mapping the topology. Many works have been done to provide methods to map the Internet topology [23], [18], [3], [9], [20], [14]. Most of these methods rely on the end-to-end measurement and the pertinence of some of them has been studied. For instance in [12], [21] the authors study the relevance of vast exploration distributed of the Internet topology. Similar work [15] has addressed the relevance of the properties of Internet and other complex networks obtained by these methods of measurement.

Measuring the dynamics of Internet require fast mapping tool of the topology. Over years, important improvement have been made to get fast and an efficient dynamics measurement tool [16], [8], [20], [6], [5].

Among the early contributions, there is [16]. Their authors proposed a tool RADAR to measure efficiently around single monitor. Recent work [6], [5] studies the dynamics of end-to-end path in the Internet considering the presence of load-balancing effect on the measurement and proposed a new tool DTRACK able to measure and anticipate the path dynamics.

A relevant choice of source, destination (vantage points) is important to efficent end-to-end measurement [10], [24], [4]. Particularly, this work [10] addressed the problem of better selection of destinations to the Internet measurement and they propose an automatic generation of the hit list, an efficient set of destinations.

Our work addresses the relevance of destinations to the observed Internet dynamics. Recent contribution on the Internet dynamics has shown some characteristics. The Internet dynamics is faster than unexpected and the occurrences of IP addresses around the monitor fit to a pattern [19], [17]. On the same stream of this study, we focus on the influence of the destinations localization on the characteristics of the IP-level dynamics observed around a monitor.

VI. Conclusion

In order to study the Internet dynamics, previous work proposed an approach to measure the dynamics from a single monitor. The monitor performs TRACEROUTE-like measurements towards a set of destinations. In the end-to-end measurement, the destinations are an important parameter. In this paper, we showed the influence of the destinations localization on the observed dynamics.

We performed two kinds of measurement, local and global and showed that their routing trees are different. We also show how this difference on the routing topologies may influence the observed dynamics through the two main characteristics of the dynamics at IP-level topology: the sustained discovery of IP addresses and the parabolic shape of the dynamics pattern.

We found that the number of new discoveries of IP addresses remains sustainable whatever with local or global, but it is not the same with the dynamics pattern. The load-balancing is the main dynamics in the global measurement whereas the routing changes are the most observed in the local measurement. The load-balancing is the main dynamics in the global measurement whereas the routing changes are the most observed dynamics in the local measurement.

Many works remain to be done on the dynamics characterization and modeling. Precisely, formally identified the different types of dynamics (load-balancing and routing dynamics) is crucial for several applications. For instance, to design efficient measurement tool. Relying on the dynamics characteristics established, it will be possible to propose more realistic models of the Internet dynamics.

Acknowledgment

We would like to thank researchers Clémence Magnien and Matthieu Latapy of Complex Networks/LIP6 team for helping with providing measurement framework and for useful discussions on the Internet topology dynamics.

References

[1] B. Augustin, X. Cuvellier, F. Orgogozo, Viger, T. Friedman, M. Latapy, C. Magnien, and R. Teixeira. Traceroute anomalies : Detection and prevention in internet graphs. *Computer Networks*, 52:998–1018, 2008.

[2] B. Augustin, T. Friedman, and R. Teixeira. Multipath Tracing with Paris Traceroute. In *Proc. Workshop on End-to-End Monitoring, E2EMON*, May 2007.

[3] Robert Beverly, Arthur Berger, and Geoffrey G. Xie. Primitives for active internet topology mapping: Toward high-frequency characterization. In *Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement*, IMC '10, pages 165–171, New York, NY, USA, 2010. ACM.

[4] Mingming Chen, Meng Xu, and Ke Xu. A delay-guiding source selection method in network topology discovery. In *Communications (ICC), 2011 IEEE International Conference on*, pages 1–6, June 2011.

[5] I. Cunha, R. Teixeira, and C. Diot. Measuring and characterizing end-to-end route dynamics in the presence of load balancing. In *PAM'11 Proceedings of the 12th international conference on Passive and active measurement*, pages 235–244, 2011.

[6] I. Cunha, R. Teixeira, D. Veitch, and C. Diot. Dtrack: A system to predict and track internet path changes. *Networking, IEEE/ACM Transactions on*, 22(4):1025–1038, Aug 2014.

[7] Radar data. http://data.complexnetworks.fr/Radar/.

[8] B. Donnet, P. Raoult, T. Friedman, and M. Crovella. Efficient algorithms for large-scale topology discovery. In Derek L. Eager, Carey L. Williamson, Sem C. Borst, and John C. S. Lui, editors, *Proceedings of the International Conference on Measurements and Modeling of Computer Systems, SIGMETRICS 2005, June 6-10, 2005, Banff, Alberta, Canada*, pages 327–338. ACM, 2005.

[9] B. Eriksson, G. Dasarathy, P. Barford, and R. Nowak. Efficient network tomography for internet topology discovery. *Networking, IEEE/ACM Transactions on*, 20(3):931–943, June 2012.

[10] Xun Fan and John Heidemann. Selecting representative ip addresses for internet topology studies. In *Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement*, IMC '10, pages 411–423, New York, NY, USA, 2010. ACM.

[11] Global and local data. http://www.univ-koudougou.bf/Radarbf/.

[12] Jean-Loup Guillaume, Matthieu Latapy, and Damien Magoni. Relevance of massively distributed explorations of the internet topology: Qualitative results. *Computer Networks*, 50.

[13] Hamed Haddadi, Steve Uhlig, Andrew W. Moore, Richard Mortier, and Miguel Rio. Modeling internet topology dynamics. *Computer Communication Review*, 38(2):65–68, 2008.

[14] B. Holbert, S. Tati, S. Silvestri, T. La Porta, and A. Swami. Network topology inference with partial path information. In *Computing, Networking and Communications (ICNC), 2015 International Conference on*, pages 796–802, Feb 2015.

[15] M. Latapy and C. Magnien. Complex network measurements: Estimating the relevance of observed properties. In *INFOCOM 2008. 27th IEEE International Conference on Computer Communications, Joint Conference of the IEEE Computer and Communications Societies, 13-18 April 2008, Phoenix, AZ, USA*, pages 1660–1668. IEEE, 2008.

[16] M. Latapy, C. Magnien, and F. Ouédraogo. A radar for the internet. In *Workshops Proceedings of the 8th IEEE International Conference on Data Mining (ICDM 2008), December 15-19, 2008, Pisa, Italy*, pages 901–908. IEEE Computer Society, 2008.

[17] C. Magnien, F. Ouedraogo, G. Valadon, and M. Latapy. Fast dynamics in internet topology: Observations and first explanations. In *Proceedings of the 2009 Fourth International Conference on Internet Monitoring and Protection*, ICIMP '09, pages 137–142, Washington, DC, USA, 2009. IEEE Computer Society.

[18] P. Marchetta and A. Pescape. Drago: Detecting, quantifying and locating hidden routers in traceroute ip paths. In *Computer Communications Workshops (INFOCOM WKSHPS), 2013 IEEE Conference on*, pages 109–114, April 2013.

[19] A. Medem, C. Magnien, and F. Tarissan. Impact of power-law topology on ip-level routing dynamics: Simulation results. In *Computer Communications Workshops (INFOCOM WKSHPS), 2012 IEEE Conference on*, pages 220–225, March 2012.

[20] Jian Ni, H. Xie, S. Tatikonda, and Y.R. Yang. Efficient and dynamic routing topology inference from end-to-end measurements. *Networking, IEEE/ACM Transactions on*, 18(1):123–135, Feb 2010.

[21] F. Ouédraogo and C. Magnien. Impact of sources and destinations on the observed properties of the internet topology. *Computer Communications*, 34(5):670–679, 2011.

[22] J.-J. Pansiot. Local and dynamic analysis of internet multicast router topology. *Annales Des Tlcommunications*, 62(3-4):408–425, 2007.

[23] Y. Shavitt and E. Shir. Dimes: Let the internet measure itself. *SIGCOMM Comput. Commun. Rev.*, 35(5):71–74, October 2005.

[24] Y. Shavitt and U. Weinsberg. Quantifying the importance of vantage points distribution in internet topology measurements. In *INFOCOM 2009, IEEE*, pages 792–800, April 2009.

[25] Feng Ying and Zhao Hai. Analysis of the hierarchical characteristics of ip-level topology dynamic node. In *Intelligent Human-Machine Systems and Cybernetics (IHMSC), 2014 Sixth International Conference on*, volume 2, pages 338–341, Aug 2014.

Improving Credit Scorecard Modeling Through Applying Text Analysis

Omar Ghailan
Faculty of Computers & Info.
Cairo University, Egypt

Hoda M.O. Mokhtar
Faculty of Computers & Info.
Cairo University, Egypt

Osman Hegazy
Faculty of Computers & Info.
Cairo University, Egypt

***Abstract*—In the credit card scoring and loans management, the prediction of the applicant's future behavior is an important decision support tool and a key factor in reducing the risk of Loan Default. A lot of data mining and classification approaches have been developed for the credit scoring purpose. For the best of our knowledge, building a credit scorecard by analyzing the textual data in the application form has not been explored so far. This paper proposes a comprehensive credit scorecard model technique that improves credit scorecard modeling though employing textual data analysis. This study uses a sample of loan application forms of a financial institution providing loan services in Yemen, which represents a real-world situation of the credit scoring and loan management. The sample contains a set of Arabic textual data attributes defining the applicants. The credit scoring model based on the text mining pre-processing and logistic regression techniques is proposed and evaluated through a comparison with a group of credit scorecard modeling techniques that use only the numeric attributes in the application form. The results show that adding the textual attributes analysis achieves higher classification effectiveness and outperforms the other traditional numerical data analysis techniques.**

***Keywords*—Credit Scoring; Textual Data Analysis; Logistic Regression; Loan Default.**

I. Introduction

Credit Scoring is a decision support tool used to identify the level of risk associated with the applicants for a specified service. It is based on applying a group of statistical techniques to predict the behaviour of those applicants and assigning scores reflecting how much good or bad they are expected to be [1]. The credit scorecard models are widely used in the risk management of the banks, insurance companies, and other financial institutions aim to identify the quality or the risk of their customers. The credit scorecard model is designed to replace the old judgmental system which depends on the decision maker or the creditor to assign the risk score. The credit scoring model's purpose is to increase the efficiency, and the reliability of the judgment process [2].

The developed applications and the proposed researches in this area used several statistical techniques to build the credit scorecard models such as Support Vector Machine [3][4], Neural Networks [5][6], Logistic Regression [7][8], Genetic Programming [9], Nearest Neighbour [10], and other hybrid techniques [11][12]. Each of those techniques has its form in representing the scorecard generated from the model. Each technique has its strengths and its advantages in some of the circumstances but there is no overall best one in all the circumstances. Group of the existing credit scoring models will be discussed in section II.

The data sources used in making a credit scoring decisions includes both the applicant form details and the information collected by a credit reference agency like the public registries, internal records in the bank from previous experiences, transactions, and any other activities initiated by the applicant in the bank [13].

Beside all the work that has been done in this area, there is a drawback point related to the structure of the data used to build the statistical model. This limitation was a result of using only that data that could be represented as numeric values and neglecting the textual data regardless of its importance in purpose of simplifying the analysis calculations and the shape of the output of the model. This limitation results in decreasing the degree of efficiency of those models by neglecting some of the available valuable data that could be used to extract some features to increase the percentage of accuracy and correctness of the classification model.

In this study, we aim to improve the existing credit scorecard modeling by employing the textual data analysis. The textual analysis results in extracting a group of textual features inserted to the credit scorecard to increase its accuracy depending on statistical aspects. The resulting scorecard from the proposed methodology is compared with a group of scorecards generated from using only the numerical data analysis. Logistic Regression, Decision Tree, SVM, and Neural Network techniques are used in the comparison using a sample of 180 loan application forms collected from a financial institution providing loan services in Yemen. The results of the comparison using Error Rate, Recall, Precision, and F1-score show the improvement in the credit scorecard accuracy when employing the textual data analysis.

The organization of this paper is as follows, related works are discussed in Section II. The proposed methodology of integrating the textual data analysis in the credit scorecard model is discussed in Section III. In Section IV, experimental results are explained. Finally, section V draws the conclusions of the paper and highlights future work related to the study.

II. Related Works

A. Related Work in Credit Scoring

Recently, the credit scoring statistical techniques have been investigated widely due to increasing the interest of financial

institutions to classify their customers specially that related to the loans. Several studies have been conducted to improve the accuracy and effectiveness of the classification techniques used in building the credit scoring models.

In [3], the authors proposed a hybrid credit scoring model by integrating SVM technique with Linear Discriminant Analysis (LDA), F-score, Decision Tree, and Rough Sets as features selection pre-processing methods. The experiments using Australian training dataset extracted from UCI Repository concluded that the hybrid model increases the classification accuracy with average accuracy rate approaches 86.52% when applying SVM/LDA model.

Another Hybrid SVM-based credit scoring models was investigated in [4]. The results showed that integrating SVM and genetic algorithm techniques can enhance the feature selection task compared to decision tree and neural networks classifiers.

The authors of [5] investigated building credit scoring models using neural networks classification techniques such as multilayer perceptron and modular neural networks compared to the other traditional techniques such as logistic regression and linear discriminant analysis. The results indicated that customized neural networks model with total correct classification rate approaches 83.19% performs better than the other models have been used in the comparison.

Another study investigated applying neural networks technique in the credit scorecard modeling was represented in [6]. A comparison with other techniques such as Probit Analysis, Discriminant Analysis, and Logistic Regression was conducted to evaluate the NN model's performance using credit risk datasets collected from Egyptian banks. The results concluded that neural nets model outperformed the other techniques with accuracy rate approaches 95.52%.

The authors in [7] proposed two credit scoring models using Logistic Regression and Radial Basis Function techniques applied to training datasets collected from Jordanian banks. The results indicated that the logistic regression model outperformed the radial basis function model with average correct classification accuracy rate approaches 85.4%.

The performance of the Logistic Regression technique when dealing with the credit scoring was investigated in [8]. The authors studied two logistic regression models on a training datasets collected from a Brazilian bank. The study concluded that there is no remarkable improvement in the prediction power when using the logistic regression with state-dependent sample selection model compared to the naive logistic regression model.

Genetic programming (GP) credit scoring model was investigated in [9]. The presented experiments used a collection of Egyptian public sector banks' data sets to test the performance of the proposed model. The experimental results concluded outperforming the GP model compared to the Probit Analysis (PA) Logistic Regression model.

Building a credit scoring model using a hybrid Adaptive Neuro Fuzzy Inference System was proposed in [11]. Using training datasets collected from an international bank in Turkey, the proposed model was compared to the other commonly utilized models in this field. The experimental results concluded that the proposed hybrid model outperforms the other techniques used in the comparison such as Neural Network and Linear Discriminant Analysis models.

The authors in [14] investigated several data mining techniques to study the classification models applied to the imbalanced credit scoring data sets. The authors have explored the suitability of least square, support vector machines, gradient boosting and random forests techniques beside other classification techniques such as logistic regression, neural networks and decision trees. The experiments illustrated that the gradient boosting and random forest classifiers are the most effective techniques for the imbalanced dataset classification.

A reassigning credit scoring model (RCSM) was presented in [15]. The authors constructed a hybrid model using Case-Based Reasoning (CBR) and Artificial Neural Network (ANN) classification techniques. The experimental results concluded outperforming the proposed model with average accuracy rate approaches 82.5% compared to Classification Tree (CART), Linear Discriminant Analysis (LDA), Back Propagation Network (BPN), and Logistic Regression (LR) models.

B. Related Work in Arabic Text Categorization

Due to the increasing in the interests of extracting the information from the textual data to support the decision making process, text mining field expanded lately to increase the efficiency and accuracy of the developed models and to include some new languages that were not targeted before[16].

For Arabic textual data classification, there are many approaches that been investigated towards developing a classifying model depending on traditional classification techniques such as Decision Trees, Logistic Regression, SVM , and neural network techniques [17].

Some researchers developed specially designed models targeted to improve the Arabic text classifiers such as the rule based models generating IF-Then rules based on the Decision Trees Models and this type of models in many studies outperformed the other techniques in case of the Arabic textual data analysis [18].

According to the research in [19], the authors used an Arabic text classifier based on Support Vector Machine technique. The classifier used CHI square method to select the features, which improved the performance of the classifier with F-measure=88.

The authors in [20] compared between Naive Bayesian method and Support Vector Machine algorithm on different Arabic text. The study concluded that the SVM algorithm outperforms the Nave Bayesian model (NB) with regards to all measures used in the comparison.

A comparative study investigated three classifiers for Arabic text categorization in [21]. The results of the comparison showed that the Nave Bayesian model outperforms both the K-NN and the distance-based classifiers.

A distance-based classifier for Arabic text categorization was proposed in [22]. Authors proposed a classifier applied to features extraction for category-specific features that capture inherent category-specific properties. The results showed that

the proposed classifier is very accurate and robust with average error rate approaches 0.0744.

Another comparative study in classifying the Arabic text documents using the N-gram frequency statistics was investigated in [23], the authors compared between using the Dice's measure of similarity and Manhattan distance statistics. The study concluded that N-gram text classifier using the Dice measure outperforms the other classifier that used the Manhattan measure.

In [24], a KNN model has been applied to classify Arabic text documents. The authors concluded that using N-Gram in the document indexing outperforms the traditional single term indexing method with average accuracy 0.73 for the N-Gram and 0.66 for the single term indexing technique.

Arabic text classification using Decision Trees (C4.5), One Rule, Rule Induction (RIPPER), and Hybrid (PART) models were studied in a comparative study represented in [25]. The results indicate that PART hybrid approach outperformed the other algorithms used in the study.

In [26], the authors used the decision tree technique based on term stemming, document normalization, and term weighting. Combining Term Frequency (TF), Inverse Document Frequency (IDF), Term Frequency Inverse Document Frequency (TFIDF), and pruning infrequent terms significantly affects the classification model by reducing the dimensionality and utilizing the text mining model especially for the large datasets.

A classification system based on Decision Trees algorithm has been evaluated in [27]. The experiments was performed over self-collected datasets and concluded that the proposed hybrid approach using the embedded information gain criterion of the decision tree algorithm is a good Arabic text classifier with average classification accuracy rate approaches 93%.

In [28], a comparative study investigated SVM and Decision Tree C4.5 models using 17658 self-collected documents. The results indicated that Decision Trees Model has more accurate classification than SVM when dealing with the Arabic textual Data.

An association-rules based classifier model for the Arabic textual data has been studied in [29]. The experiments on a self-collected sample concluded that the proposed classifier features high accuracy rates.

A comparative study investigated the performance of Nave Bayes, SVM, and Decision Tree (C4.5) Classifiers when applied to self-collected Arabic text datasets in [30]. The study concluded that the Nave Bayes classifier with average accuracy rate approaches 85.25% outperformed the other models used in the study.

According to [31], a modified Neural Network Model is developed using the Singular value Decomposition representation (SVD). The data used in that NN research consists of 453 documents with 14 categories collected from Al-Hadeeth Books. The modified version of NN in this research outperformed the original artificial network model with classification accuracy rate approaches 88.33%.

III. METHODOLOGY

This section demonstrates the proposed methodology to improve the credit scorecard model by applying the text data analysis along with traditional numeric data analysis method.

A. Textual Data Pre-Processing

For the text data, the proposed method of the text analysis consists of the following steps:

- Text Parsing: This step is used to parse, stem, identify the noun groups, and identify the part-of-speech of the text fields. It is also used to remove the words included in the stop list.
- Text Filter: This step is used to filter the words extracted from the text parsing based on some pre-defined criteria. In our study, the pre-defined minimum number of documents to accept the word is equal 3. Entropy Term Weight formula 1 is used in calculating the importance weight of the extracted words. This technique gives a higher weight for the rare terms. If the term appears in only one document, it will have entropy weight =1 (the max entropy weight). If the term appears in all the documents then it will have entropy weight =0 (the lowest entropy weight) [32]. Entropy Term Weight equation is

$$G_i = 1 + \sum_{j=1}^{d_i} \frac{p_{ij} log_2(p_{ij})}{log_2(n)} \quad (1)$$

  Where: p_{ij}= the frequency that term i appears in document j divided by the frequency that term i appears in document collection
  n = number of documents in the collection
  d_i= number of documents in which term i appears
- Topics Extraction: A group of topics are constructed from the words resulting from the filtering step. The extracted topics contain the words related to each other based on their appearance in the training dataset collection. The main target of this step is to reduce the dimensionality of the features that will enter the regression model since it will not be effective if we use each extracted word as a feature in our classification model. The employed technique uses Latent Semantic Indexing (LSA) concept through Singular Value Decomposition (SVD) [33]. This helps in grouping similar words into a limited number of distinct sets. Those sets are applied as topics. Each topic will be used as a feature in our regression model. In our research, 30 text topics are collected from each text field to be used in building the credit scorecard.

B. Interactive Grouping

This step is used to eliminate the weak characteristics that need to be neglected when applying the logistic regression model since some of the used characteristics, either the textual or the numerical fields, have no influence in the final scorecard.

Each numerical field is distributed into intervals based on the similarity in its values' predictive power. Those intervals will be used as features in the final credit scorecard model.

In the developed model, the Information Value is calculated using formula 2 to determine the overall predictive power of the attribute. The predictive power increases as the ability of separating the good and bad records increases [34].

$$IV = \sum_{i=1}^{L}(Distr\ Good_i - Distr\ Bad_i) * ln(\frac{Distr\ Good_i}{Distr\ Bad_i}) \quad (2)$$

Where: L=Number of intervals (levels) in the characteristic.

In this study, the characteristics with IV less than 0.10 are eliminated because of their low prediction power.

C. Applying the Logistic Regression Model

By applying the previous pre-processing steps, each text topic or numerical interval is transformed into a column with value 0 or 1 indicating if the customer's profile contains that feature or not. The resulted 0/1 matrix entered the standard Logistic Regression model represented in 3 which serves the target of generating the credit scorecard for both the numerical attributes and the modified textual attributes [35].

$$Logit(p) = \ln(\frac{p}{1-p}) = \beta_0+\beta_1X_1+\beta_2X_2+...+\beta_nX_n \quad (3)$$

Where: p = the posterior probability of the class 0
X = the input variables
β = the estimation coefficients of the X input variables.

D. Calculating the Final Scores

The Logistic Regression model's output is tuned by applying Weight of Evidence. WOE enhances the final credit scores because Logistic Regression model considers both the features' values 0 and 1 in its processing while the credit scorecard model gives a higher consideration to having the value 1 which indicates that the customer's profile contains the extracted feature.

$$WOE = ln(\frac{Distr\ Good_i}{Distr\ Bad_i}) \quad (4)$$

$$FinalScore = WOE * Estimation\ Coefficient \quad (5)$$

Figure 1 summarizes the proposed methodology.

IV. EXPERIMENTAL RESULTS

In this section, we study the effect of applying the text analysis on the credit scorecard model's accuracy by comparing the proposed model's results with the traditional techniques that depend only on the numerical variable when building the credit scorecard.

The dataset used in the experiment consists of 179 records (divided into 150 training data + 29 test data). The dataset is self-collected from CAC Bank, a financial institution providing a group of loan services in Yemen. The dataset contains of 16 text fields and 8 numeric fields.

Four classification models were developed for the purpose of the comparison with the proposed model. Those models applied the traditional techniques of building the credit scorecard using only the numeric variables. The developed models configurations are shown in table I. The models used in the experiments are implemented using SAS Enterprise Miner tool.

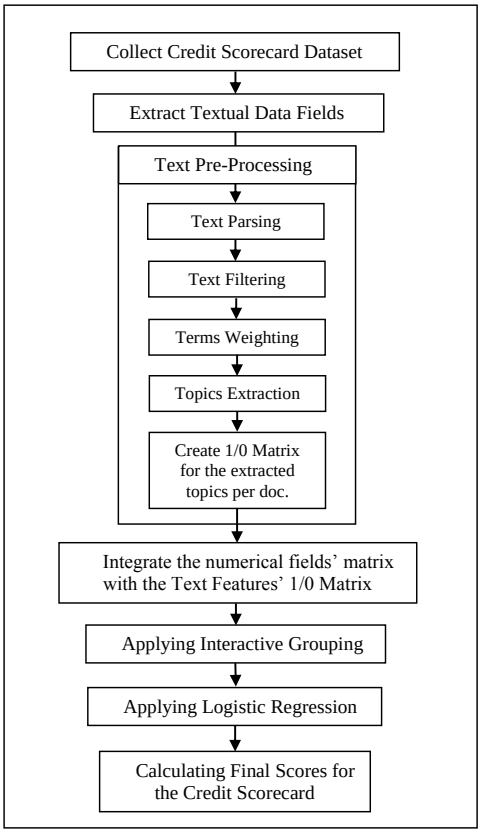


Fig. 1: The steps that make up the proposed methodology

TABLE I: Experimental Models Configurations

Model	Parameter	Value
Logistic Regression	Two-Factor Interactions	N
	Polynomial Terms	N
	Regression Type	LOGISTIC
	Link Function	LOGIT
Decision Tree	Model Type	C4.5
	Ordinal Criterion	ENTROPY
	Significance Level	0.2
	Maximum Branch	2
	Maximum Depth	6
SVM	Estimation Method	DQP
	Scale Predictors	Y
	Regularization	TUNING
	Constant value	0.1
	Kernel	LINEAR
Neural Network	Architecture	MLP
	Termination	OVERFITTING
	Maximum Iterations	8
	Number of Hidden Units	2
	Direct/Tanh/Sine	Y

The performance evaluation statistics of the traditional techniques applied in building the credit scorecard by using only the numeric variables are shown in table II.

TABLE II: Classification Accuracy for Four Numerical Data Scorecard Models

Evaluation Method	Decision Tree	SVM	Neural Network	Logistic Regression
Average Squared Error	0.1513	0.193	0.3027	0.336
Roc Index	0.888	0.833	0.467	0.733
Misclassification Rate	0.2414	0.308	0.5172	0.483
Wrong Classifications	7	8	15	14

All the performance values indicate that the decision tree model outperformed the other models with the lowest average square error of '0.1513' and lowest misclassification rate of '0.241' which is slightly better performance than SVM. Hence, the decision tree model is used in the comparison with the proposed model. The results of the experiments are shown in table III.

TABLE III: The Evaluation Measures of The Propsoed Model

Evaluation Method	Decision Tree (Credit Scorecard Model without Text Analysis)	Proposed Methodology (Credit Scorecard Model with Text Analysis)
True Positive	9	14
True Negative	13	13
False Positive	1	1
False Negative	6	1
Overall Accuracy (ACC)	75.9%	93.1%
Precision (PPV)	0.90	0.93
Recall (TPR)	0.60	0.93
F1 Score	0.72	0.93

The experimental results imply that the enhancement that implemented in the credit scorecard model after applying text analysis and adding the textual variables' extracted features to the credit scorecard is highly affecting the model's accuracy with reference to all the statistical evaluation measures used in the comparison.

Sample of the text features in the credit scorecard resulted from applying the proposed methodology to a training data set extracted from a financial institution providing loan services in Yemen is presented in table IV.

TABLE IV: Sample of the Proposed Credit Scorecard Model's Output

Attribute	Feature	Score
Activity Description	استيراد ، ماء ، مكائن تجارة ، دريلات ، كمبريوسرات	0.226
	أثاث ، صناعة الأثاث ، متخصص ، تعليمي	-0.395
	missing data	-1.004
	قماش ، بالجملة للأقمشة ، معتمد ، رجالي ، نسائي	0.483
	تجزئة ، تجار ، توزيع ، استهلاكية ، مورد	-0.660
	جسور ، مجال الطرقات ، والجسور وصيانتها ، تنفيذ مقاولات ، إنارة	2.287
	استيراد ، إستيراد الأدوية ، صيدلاني ، دواء ، مستلزم	1.120
Main Competitor	تابعة ، سعيد ، شركة ، هائل ، فاهم	-9.464
	الحاج للدواجن ، ناصر ، مهراس ، داجنة ، الزيلعي	0.458
Main Customers	شركة ، بترولیم ، نكس	-0.953
	مركز ، محافظة ، محلة ، وزارة الاشغال	1.235
Main Suppliers	تركيا ، المكلا ، ريسوت ، مؤسسة ، هائل	-0.532
	الإمارات ، سيراميك ، الجزيرة ، الشارقة ، سابتكس	0.518
Oraganization Address	جبل ، شرق ، ذمار ، ستون ، غربي	-2.012
	صناع ، مجمع ، تعز، جند ، الحوبان	2.230
	عدن ، كريتر ، فرع	-0.560
Oraganization Type	مؤسسة ، فردة ، منشأ ، منشأة	0.198
	missing data	-5.659
	حكومي ، دولة ، تابعة	5.358
	شركة ، مقفل ، مساهمة ، شركة مساهمة ، مسئولية	0.494

V. Conclusions

This paper has investigated improving the credit scorecard modelling by applying the textual data analysis for the text information filled in the forms provided by the applicants. The developed model increases the number of features in the credit scorecard by adding the textual features to the numerical features resulting from the logistic regression model after applying the pre-processing steps to the textual fields. The results of the experiments using a self-collected dataset revealed that adding the textual fields' features improves the accuracy of the credit scorecard model by increasing the correct classification rate.

Future studies should aim to apply other advanced statistical techniques; such as genetic algorithms and fuzzy discriminant analysis, integrated with the textual data analysis to build an enhanced credit scorecard. In addition to this, the plan is to collect larger dataset to increase the accuracy of the model.

ACKNOWLEDGMENT

The authors would like to thank CAC Bank's Compliance Management for availing the dataset used in the research's experiments.

REFERENCES

[1] N. Siddiqi, Credit risk scorecards: developing and implementing intelligent credit scoring, vol. 3, John Wiley & Sons, 2012.

[2] H. A. Abdou and J. Pointon, "Credit scoring, statistical techniques and evaluation criteria: A review of the literature", Intelligent Systems in Accounting, Finance and Management, vol.18, no. 2-3, pp. 59-88, Apr 2011.

[3] F. L. Chen and F. C. Li, "Combination of feature selection approaches with SVM in credit scoring", Expert Systems with Applications, vol. 37, no. 7, pp. 4902-4909, Jul 2010.

[4] C. L. Huang, M. C. Chen, and C. J. Wang, "Credit scoring with a data mining approach based on support vector machines. Expert systems with applications, vol. 33, no. 4, pp. 847-856, Nov 2007.

[5] V. S. Desai, J. N. Crook, and G. A. Overstreet, "A comparison of neural networks and linear scoring models in the credit union environment", European Journal of Operational Research, vol. 95, no. 1, pp. 24-37, Nov 1996.

[6] H. Abdou, J. Pointon, and A. El-masry, "Neural nets versus conventional techniques in credit scoring in Egyptian banking", Expert Systems with Applications, vol. 35, no. 3, pp. 1275-1292, Oct 2008.

[7] H. A. Bekhet and S. F. K. Eletter, "Credit risk assessment model for Jordanian commercial banks: Neural scoring approach", Review of Development Finance, vol. 4, no. 1, pp. 20-28, Mar 2014.

[8] F. Louzada, P. H. Ferreira-Silva, and C. A. Diniz, "On the impact of disproportional samples in credit scoring models: An application to a Brazilian bank data", Expert Systems with Applications, vol. 39, no. 9, pp. 8071-8078, Jul 2012.

[9] H. A. Abdou, "Genetic programming for credit scoring: The case of Egyptian public sector banks", Expert Systems with Applications, vol. 36, no. 9, pp. 11402-11417, Nov 2009.

[10] W. E. Henley, D. J. Hand, "A k-nearest-neighbour classifier for assessing consumer credit risk", The Statistician, pp. 77-95, Jan1996.

[11] S. Akkoc, "An empirical comparison of conventional techniques, neural networks and the three stage hybrid Adaptive Neuro Fuzzy Inference System (ANFIS) model for credit scoring analysis: The case of Turkish credit card data", European Journal of Operational Research, vol. 222, no. 1, pp. 168-178, Oct 2012.

[12] T. S. Lee, C. C. Chiu, C. J. Lu, and I. F. Chen, "Credit scoring using the hybrid neural discriminant technique", Expert Systems with applications, vol.23, no. 3, pp. 245-254, Oct 2002.

[13] L. C. Thomas, "A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers", International journal of forecasting, vol. 16, no. 2, pp. 149-172, Jun 2000.

[14] I. Brown, C. Mues, "An experimental comparison of classification algorithms for imbalanced credit scoring data sets", Expert Systems with Applications, vol. 39, no.3, pp. 3446-3453, Feb 2012.

[15] C. L. Chuang and R. H. LIN, "Constructing a reassigning credit scoring model", Expert Systems with Applications, vol. 36, no.2, pp. 1685-1694, Mar 2009.

[16] F. Thabtah, O. Gharaibeh, and H. Abdeljaber, "Comparison of rule based classification techniques for the Arabic textual data", Innovation in Information & Communication Technology (ISIICT), Fourth International Symposium on. IEEE, p. 105-111, Dec 2011.

[17] R. Feldman and J. Sanger, The text mining handbook: advanced approaches in analyzing unstructured data, Cambridge University Press, 2007.

[18] M. Kantardzic, Data mining: concepts, models, methods, and algorithms, John Wiley & Sons, 2011.

[19] A. Mohd A Mesleh, "Chi square feature extraction based SVMs Arabic language text categorization system", Journal of Computer Science, vol. 3, no. 6, pp. 430-435, 2007.

[20] S. Alsaleem, "Automated Arabic Text Categorization Using SVM and NB", Int. Arab J. e-Technol., vol. 2, no. 2, pp. 124-128, Jun 2011.

[21] R. M. Duwairi, "Arabic Text Categorization", Int. Arab J. Inf. Technol., vol. 4, no. 2, pp. 125-132, 2007.

[22] R. M. Duwairi, "A Distance-based Classifier for Arabic Text Categorization", DMIN, p. 187-192, Jun 2005.

[23] L. Khreisat, "Arabic text classification using N-gram frequency statistics a comparative study", DMIN, pp. 78-82, Jun 2006.

[24] R. Al-shalabi and R. Obeidat, "Improving KNN Arabic text classification with n-grams based document indexing", Proceedings of the Sixth International Conference on Informatics and Systems, Cairo, Egypt, pp. 108-112, Mar 2008.

[25] M. Al-diabat, "Arabic text categorization using classification rule mining", Applied Mathematical Sciences, vol. 6, no. 81, pp. 4033-4046, Mar 2012.

[26] M. K. Saad and W. Ashour, "Arabic text classification using decision trees", Proceedings of the 12th international workshop on computer science and information technologies CSIT, pp. 75-79, 2010.

[27] F. Harrag, E. El-qawasmeh, and P. Pichappan, "Improving Arabic text categorization using decision trees" Networked Digital Technologies, 2009. NDT'09. First International Conference on. IEEE, pp. 110-115, Jul 2009.

[28] S. Al-harbi, A. Almuhareb, A. Al-Thubaity, M.. S. Khorsheed, and A. Al-Rajeh, "Automatic Arabic text classification", 2008.

[29] A. El-Halees, "Arabic text classification using maximum entropy", The Islamic University Journal (Series of Natural Studies and Engineering), vol. 15, no. 1, pp. 157-167, 2007.

[30] A. H. Wahbeh and M. Al-kabi, "Comparative Assessment of the Performance of Three WEKA Text Classifiers Applied to Arabic Text", Abhath Al-Yarmouk: Basic Sci. & Eng, vol.21, no. 1, pp. 15-28, 2012.

[31] F. Harrag and E. El-qawasmah, "Neural Network for Arabic text classification", Applications of Digital Information and Web Technologies, 2009. ICADIWT'09. Second International Conference on the. IEEE, pp. 778-783, 2009.

[32] J. R. Quinlan, C4. 5: programs for machine learning, Elsevier, 2014.

[33] R. Albright, "Taming Text with the SVD", SAS Institute Inc., Cary, NC, Jan 2004.

[34] P. B. Cerrito, Introduction to data mining using SAS Enterprise Miner, SAS Publishing, 2006.

[35] D. W. Hosmer Jr and S. Lemeshow, Applied logistic regression. John Wiley & Sons, 2004.

Iterative Threshold Decoding Of High Rates Quasi-Cyclic OSMLD Codes

Karim Rkizat
Mohammed V University in Rabat, ENSIAS
Labo SIME, Team TSE
Rabat, Morocco

Anouar Yatribi
Mohammed V University in Rabat, ENSIAS
Labo SIME, Team TSE
Rabat, Morocco

Mohammed Lahmer
Moulay Ismail University
High School of Technologie
Meknes, Morocco

Mostafa Belkasmi
Mohammed V University in Rabat, ENSIAS
Labo SIME, Team TSE
Rabat, Morocco

***Abstract*—Majority logic decoding (MLD) codes are very powerful thanks to the simplicity of the decoder. Nevertheless, to find constructive families of these codes has been recognized to be a hard job. Also, the majority of known MLD codes are cyclic which are limited in the range of the rates. In this paper a new adaptation of the Iterative threshold decoding algorithm is considered, for decoding Quasi-Cyclic One Step Majority logic codes (QC-OSMLD) codes of high rates. We present the construction of QC-OSMLD codes based on Singer difference sets of rate 1/2, and codes of high rates based on Steiner triple system which allows to have a large choice of codes with different lengths and rates. The performances of this algorithm for decoding these codes on both Additive White Gaussian Noise (AWGN) channel and Rayleigh fading channel, to check its applicability in wireless environment, is investigated.**

Keywords*—Iterative threshold decoding; Quasi-Cyclic codes; OSMLD codes; Majority logic decoding; Steiner Triple System; BIBD

I. Introduction

Today LDPC codes [1] are present in most Telecom standards like DVB-S2 and WiMAX [2]. However, the decoding of these codes remain algorithmically complex and in situations such as the DVB-S2 [3] are often concatenated with codes such as Reed Solomon to improve performances. In our point of view, the MLD codes are better competitors for LDPC codes and this for several reasons. In fact, the hardware implementation of these codes is very simple and only requires AND gates. The cyclic OSMLD codes can be decoded iteratively by an extension of the Massey algorithm [4] which is less complex than the believe propagation algorithm but almost with the same performances.

In this article, the studied subject is QC-OSMLD codes which, unlike the cyclic OSMLD codes, offer a wide range of rates equivalent to that used in the standards. The first QC-OSMLD codes were constructed by L. Townsend and E. Weldone [5], but most of these codes are constructed by either computer search or hand through trial-and-error, except the construction based on Singer Difference set, which is a geometry projective method. Later, Chen Zhi and al[6] had given a mathematical formulation for the construction of QC-OSMLD codes with high rates, these codes are based on Steiner Triple system (STS) .

Iterative threshold decoding QC-OSMLD codes of rate 1/2 has proven to perform remarkably well on Additive White Gaussian Noise (AWGN) channel [7]. the purpose of this paper is to investigate the performance of iterative threshold decoding of QC-OSMLD codes of rate $\frac{n_0-1}{n_0}$ constructed from Singer Difference Set, and STS on both Rayleigh fading channel and AWGN channel, is investigated .

The organization of the paper is as follows. The first section provides the reader with a concise description of not only the OSMLD codes but also the majority logic decoding algorithm and the Quasi-Cyclic Codes. Afterwards, the second section defines the Singer Difference Set, and it presents the constructed codes based on this algorithm. Section 3 is about the construction of QC-OSMLD codes of rate of the form $\frac{n_0-1}{n_0}$ based on STS, starting with a description of Balanced Incomplete Block Design (BIBD), then presenting the STS construction, and eventually, presenting the different constructed codes. Section 4 introduces the encoding method after describing the iterative threshold decoding algorithm and explaining the modification made for the Rayleigh fading channel. Finally, the last part presents the simulations results and analyses the ITD algorithm for decoding the constructed codes on both Rayleigh fading channel and AWGN channel.

II. Quasi-cyclic OSMLD Codes

A. OSMLD Codes

Consider an (n, k) linear code C with parity-check matrix H. The row space of H is an (n, n-k) code, denoted by $C^{\perp}$, which is the dual code of C or the null space of C. For any vector v in C and any vector w in $C^{\perp}$, the inner product of v and w is zero [8]. Now let consider that a codeword vector in C is transmitted over a binary symmetric channel. Taking into consideration that e(e_1, e_2,..., e_n) and r(r_1, r_2,..., r_n) are the error vector and the received vector, respectively. Then r = v + e. The construction of the below linear sum of the received

vector for any vector w in the dual code $C^{\perp}$:

$$A = \sum_{p=1}^{n} r_p w_p \quad (1)$$

Which is called a parity-check sum. Using the fact that <w,v>=0, the following relationship between the parity-check sum A and error digits in e is obtained:

$$A = \sum_{p=1}^{n} e_p w_p \quad (2)$$

Suppose that there exist J vectors in the dual code $C^{\perp}$, which have the following properties:

1) The j^{th} component of each vector w_i is a 1.
2) For i≠j there is at most one vector whose i^{th} component is a 1.

These J vectors are said to be orthogonal on the j^{th} digit position. They are called orthogonal vectors. Now, let us form J parity-check sums from these J orthogonal vectors,
For each i in 1,.., J $A_i = \sum_{p \neq 1} e_p + e_j$ the error digit e_j is checked by all the check sums above. Because of the second property of the orthogonal vectors, any error digit other than e_j is checked by at most one check sum. These J check sums are said to be orthogonal on the error digit e_j. If all the error digits in the sum A_i are zero for i≠j, the value of A_i is equal to e_j. Based on this fact, the parity-check sums orthogonal on e_i can be used to estimate e_i, or to decode the received digit r_i.

B. Majority logic decoding principle

The error digit e_j is decoded as 1 if at least one-half of the check sums orthogonal on e_j, are equal to 1; otherwise, e_j is decoded as 0 like majority rule [8]. When C is a cyclic code, each e_i can be decoded simply by cyclically permuting the received word r into the buffer store.

Example 1:
Let us consider the (7,3) code, which is the short code in difference set codes class. This code is specified by the perfect difference set P=0, 2, 3 of order 21. From this perfect set, the following three check sums orthogonal on e_7 could be formed:
$A_1 = e_4 + e_5 + e_7$
$A_2 = e_2 + e_6 + e_7$
$A_3 = e_1 + e_3 + e_7$
If a simple error e=(000001) occurs, then $A_1 = A_2 = A_3 = 1$. If a double error occurs; for example, e_7=1 and one value of e_1, ..., e_6 is equal to 1, then two values of A_i are 1. So we can say that :
- e_7=1 if only and if at least 2 values of A_i are 1
- e_7=0, otherwise

C. Quasi-cyclic Codes

A code is said to be quasi-cyclic if every cyclic shift of a codeword by p positions results in another codeword [9]. Therefore, a QC codes are a generalization of cyclic codes with p = 1. A QC code (mn_0, mk_0) with a minimum distance d based on difference set can be specified with k_0 disjoints difference sets $\{D_1,D_2,...,D_{k_0}\}$ such that $D_i(d_{i0},d_{i1},d_{i2},...,d_{i(S-1)})$ of order S, chosen from the set {0,1,2,..., mk_0} [5]. The parity check matrix H in the systematic form of such code is completely defined as follows:

$$H = [P_1 P_2 ... P_{k_0} I_{n-k}] \quad (3)$$

The circulant matrix P_i is deducted from the difference set D_i; the elements of D_i can specify the position in the matrix header P_i with one, while d_{ij} represents one in the position j, the others rows are obtained by a cyclic shift of the header. Where I represents the identity matrix.

The majority logic decoding algorithm for QC codes is the same as cyclic codes. However, there is a little bit difference between them. Hence, in cyclic codes each error digit e_i can be decoded by cyclically permuting the received word r, but in QC codes in systematic form, shift is done cyclically by one position of each (n-k) bits simultaneously.

Example 2:
Let consider the QC code C(6,3,3). This code is of the rate 1/2 and based on the Singer difference set DS{0,1} of order 2.

The parity check matrix H in systematic form is [P I_3]

$$\begin{pmatrix} 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 \end{pmatrix}$$

The parity-check sum orthogonal on e_3 is obtained from the parity check matrix H :
$A_1 = e_2 + e_3 + e_5$
$A_2 = e_1 + e_3 + e_6$

III. Construction based on Singer Difference Set

A. Signer Difference Set

A difference set [10] of order S and modulo m ≥ S(S - 1)+1 is defined as a collection of S integer specified from the set {0,1,...,m-1} such that no two of the S(S - 1) ordered differences modulo m are identical. If m=S(S - 1)+1, then for any non-zero integer n < m, there is exactly one pair of elements in the difference set such that their difference is congruent to n modulo m. Such a set is called a perfect difference set.

Singer [11] has demonstrated how to construct such sets when $S = p^u + 1$ and p is prime. Points and lines in the projective geometry PG(m - 1, q) form a difference set of parameters [q,m]. The construction of Singer difference sets with parameters [q,m] is straightforward if a primitive polynomial of degree m in F_q is known. The following is the basis of Singer algorithm:

1) Choose a primitive polynomial of degree m in F_q

$$f(x) = x^m + \sum_{i=1}^{m} a_i x^{m-i} \quad (4)$$

2) Choose the start value λ_0 = 0, λ_1, λ_2,..., λ_{m-1}

3) Calculate the recurrence relation

$$\lambda_n = -\sum_{i=1}^{m} a_i \lambda_{n-i} \quad (5)$$

4) The set of integers { $0 \leq i < \frac{q^m-1}{q-1} : \lambda_i = 0$ } is a Pefect Difference Set

Example 3:

1) Taking primitive polynomial $x^4 + x^3 + 2$ of degree 4 on F_3
2) Choosing the start sequence λ_0 = 0, λ_1 = 0, λ_2 = 1
3) Then, calculating the recurrence relation as follows: $\lambda_n = 2\lambda_{n-1} + \lambda_{n-4}$ yields the sequence 1000121220111222202021120102100221202200200 0 . . .
4) The Positions where $\lambda_i = 0$ are {1,2,3,9,17,19,24,26,29,30,35,38,39} that form the Perfect Difference Set.

B. Construction based on Singer Difference Set

The Singer construction allows to have codes with rate $\frac{1}{2}$. It's possible to construct a single perfect difference set of order S = p^u + 1 and modulo m = p^{2u} + p^u + 1 where p is a prime number, and u is a positive integer. As a results, it's always possible to construct codes with minimum distance p^u + 2 and length 2(p^{2u} + p^u + 1).

All constructed codes are OSMLD and completely orthogonalizable since the orthogonal parity-check equations number J is always equal to d_{min}-1, where d is the minimum distance of the code.

In [5], Townsend and Weldon has constructed a rate $\frac{1}{2}$ codes of small length up to n = 366. With the help of Magma [12], many perfect difference sets have been constructed, which allows to construct a large number of QC-OSMLD codes of large lengths up to n = 2.10^9.

Table 1 shows a part of constructed codes. The parameters in this table are :

P^u : P is prime, and u an integer

m : is the modulo, m = p^{2u} + p^u + 1

(n , k) : n is the length and k is the dimension of the code

d : is the minimum distance of the code

Difference Sets : Represent the constructed Singer difference set

Density : is the density of the parity check matrix H

LDPC : specifies if the code is Low Density Parity Check (LDPC) code, for that it's obligatory to have density $\leq \log_2(n)$

From this table, it's clear that the majority of QC-OSMLD codes are LDPC codes which allow us to decode them with LDPC decoder like Sum-Product, Belief Propagation ...

IV. Construction based on Steiner Triple System

Historically, Smith [13] presented in 1968 an application of incomplete block designs to the construction of several families of error-correcting codes which may be decoded using a relatively simple majority logic decoding procedure. However, he didn't give any explicit construction for such designs. Special cases of these codes are equivalent to the Self-orthogonal Quasi-cyclic codes based on Perfect Difference Sets discussed by Townsend and Weldon [5] (1967).

Chen Zhi and al stated an [6] explicit constructions of many classes of difference families considered as base blocks for Steiner designs. He presented a construction of infinite optimal self-orthogonal quasi-cyclic codes with high rates.

This section describes briefly the different construction methods of QC-OSMLD codes based on block design. Also, a part of constructed codes generated automatically by Matlab programmes is represented.

A. Balanced Incomplete Block Design

A Balanced Incomplete Block design (BIBD) [10] is a pair (V, B) where V is a set and B is a collection of b k-subsets of V (blocks) such that each element of V is contained in exactly r blocks, and any 2-subset of V is contained in exactly λ blocks. The numbers v, b, r, k, and λ are said parameters of the BIBD.

Trivial necessary conditions for the existence of a BIBD(v, b, r, k, λ) are :

1) $vr = bk$
2) $r(k-1) = \lambda(v-1)$

The incidence matrix of a BIBD (V, B) with parameters v, b, r, k, λ is a $v \times b$ matrix $A = (a_{ij})$, in which $a_{ij} = 1$ when the i^{th} element of V occurs in the j^{th} block of B, and $a_{ij} = 0$ otherwise.

B. Construction methods

1) QC-OSMLD codes based on STS ($v = 6t + 1$): This construction [6][14] is applicable for $k = 3$ and v is a power of a prime of the form $v \equiv 1 (mod\, 6)$. Considering for each v, $GF(v = p^e)$, the Galois field of order v. Let ω be a primitive root. Then the STS difference family with parameters $v = 6t + 1$, $b = 6t^2 + t$, $r = 3t$, $k = 3$, $\lambda = 1$, is described by base blocks typically given by the form :

$$(\omega^0, \omega^{2t}, \omega^{4t}),\ (\omega^i, \omega^{2t+i}, \omega^{4t+i}),\ (\omega^{t-1}, \omega, \omega^{5t-1})$$

Another construction proposed by Rosa [15][10], for which the knowledge of a primitive root is not required, may be applicable for these designs using Skolem Sequences.

From [14] the block designs given above, an infinite class of optimal QC-OSMLD $[(t+1)(6t+1), t(6t+1), 4]$ codes with the basic block length $n_0 = (t+1)p^{e-1}$ is specified.

TABLE I: Rate $\frac{1}{2}$ QC-OSMLD codes based on Perfect Difference sets

P^u	m	(n,k)	d	Difference Sets	Density	$\log_2(n)$	LDPC
1	3	(6,3)	3	0 1	66,66	2.5849	No
2	7	(14,7)	4	0 1 3	28 .57	3,8073	No
3	13	(26,13)	5	0 1 3 9	19.23	4,7004	No
2^2	21	(42,21)	6	0 1 4 14 16	14.28	5,3923	No
5	31	(62,31)	7	0 1 3 10 14 26	11.29	5,9541	No
7	57	(114,57)	9	0 1 6 15 22 26 45 55	8.33	6,5849	No
2^3	73	(146,73)	10	0 1 12 20 26 30 33 35 57	7.89	6,8328	No
3^2	91	(182,91)	11	0 1 37 39 51 58 66 69 82 86	6.04	7,5077	Yes
11	133	(266,133)	13	0 1 3 17 21 58 65 73 100 105 111 124	4.88	8,0552	Yes
13	183	366,183)	15	0 1 3 24 41 52 57 66 70 96 102 149 164 176	4.09	8,5156	Yes
2^4	273	(546,273)	18	0 1 22 33 83 122 135 141 145 159 175 200 226 229 231 238 246	3,09	9,09	Yes
17	307	(614,307)	19	0 1 3 30 37 50 55 76 98 117 129 133 157 189 199 222 293 299	2,75	9,26	Yes
19	381	(762,381)	21	0 1 3 13 28 51 65 82 86 104 112 145 201 212 217 241 261 307 339 375	2,26	9,57	Yes
23	553	(1106,553)	25	0 1 3 14 31 60 64 109 146 151 185 265 286 313 321 337 357 375 454 460 479 486 501 544	1,77	10,11	Yes
29	871	(1742,871)	31	0 1 3 23 30 41 88 97 132 165 169 186 201 211 235 306 319 345 425 431 542 547 561 592 604 620 668 719 811 819	1,66	10,76	Yes
31	993	(1986,993)	33	0 1 3 13 101 127 154 169 204 210 226 235 259 289 297 317 356 434 474 478 495 538 570 584 589 607 618 654 749 756 801 920	1,37	10,95	Yes
37	1407	(2841,1407)	39	0 1 3 25 32 82 99 208 313 410 453 479 487 557 621 649 709 736 742 782 827 837 848 890 895 899 913 951 1040 1088 1123 1142 1172 1213 1252 1272 1288 1395	1,24	11,47	Yes
47	2257	(4514,2257)	49	0 1 3 131 138 143 296 377 381 457 566 590 690 712 773 802 891 905 973 979 996 1030 1039 1050 1065 1075 1083 1102 1123 1238 1270 1337 1387 1434 1528 1541 1590 1606 1636 1757 1788 1816 1858 1914 1978 2033 2144 2219	1,06	12,14	Yes
97	9507	(19014,9507)	99	0 1 3 37 52 191 308 332 433 914 919 984 1093 1155 1231 1238 1600 1678 1723 1732 1755 1773 1826 1930 1938 2099 2116 2141 2457 2712 2859 3058 3187 3466 3524 3655 3675 3748 4139 4145 4183 4297 4301 4518 4528 4600 4720 4777 4964 5043 5054 5176 5268 5329 5356 5496 5526 5601 5617 5851 6151 6173 6491 6539 6759 6778 6792 6878 7021 7163 7226 7290 7490 7650 7747 7860 7941 8028 8056 8154 8304 8339 8370 8438 8450 8505 8534 8574 8797 9005 9048 9094 9107 9133 9154 9270 9326 9400	0,49	14,21	Yes
181	32943	(65886,32943)	183	0 1 129 145 211 306 460 514 547 748 771 800 894 1044 1101 1152 1277 1553 1798 1833 1840 1888 1924 2118 2381 2431 2564 2601 2613 3054 3308 3669 4369 4507 4620 4807 4839 5136 5342 5452 5623 5798 5808 5914 6488 6577 6798 6816 7063 7590 7745 7894 7935 7993 7995 8365 9166 9234 9572 9836 10220 10263 10355 10692 10764 10895 11081 11272 11376 11598 11645 12078 12215 12453 12498 12536 12807 12973 13250 13296 13384 13423 13858 13935 14408 14494 14603 14818 14892 15318 15397 15478 15625 15797 16219 16454 16607 17068 17141 17200 17211 17330 17696 17722 18264 18291 18433 18659 18715 18795 18958 19607 19714 19879 20145 20324 20523 20585 21192 21349 21370 21373 21728 22555 22586 22815 22929 23208 23376 23535 23550 23894 24074 24326 24490 24518 24802 24808 24904 24926 25681 25822 25839 26204 26421 26440 26474 26518 26538 26543 26658 26966 27006 27071 27363 28337 28404 28504 28697 28895 28971 29246 29883 29897 29958 30097 30106 30110 30322 30352 30473 30771 31030 31192 31380 31582 32046 32445 32676 32739 32747 32832	0,27	16,00	Yes

2) QC-OSMLD codes based on STS ($v = 6t + 3$): The construction of such designs using the Extended Skolem Sequences is proposed. A Skolem sequence of order n is a sequence $S = (s_1, s_2, ..., s_{2n})$ of $2n$ integers satisfying the conditions :

1) for every $k \in \{1, 2, ..., n\}$ there exists exactly two elements $s_i, s_j \in S$ such that $s_i = s_j = k$
2) if $s_i = s_j = k$ with $i < j$, then $j - i = k$.

Skolem sequences are also written as collections of ordered pairs $\{(a_i, b_i) : 1 \leq i \leq n,\ b_i - a_i = i\}$ with $\cup_{i=1}^{n}\{a_i, b_i\} = \{1, 2, ..., 2n\}$

Example 4:
A Skolem sequence of order 5 : $S = (1, 1, 3, 4, 5, 3, 2, 4, 2, 5)$ or, equivalently, the collection $\{(1, 2), (7, 9), (3, 6), (4, 8), (5, 10)\}$.

An extended Skolem sequence of order n is a sequence $ES = (s_1, s_2, ..., s_{2n})$ of $2n+1$ integers satisfying conditions 1 and 2 of the previous definition and :

3) there is exactly one $s_i \in ES$ such that $s_i = 0$.

The element $s_i = 0$ is called the hook or zero of the sequence.

3) Construction By A. Rosa: Suppose $\{1, ..., 3n+1\}\backslash\{2n+1\}$ can be partitioned into m triples $\{a, b, c\}$ such that $a+b = c$ or $a + b + c \equiv 0(mod\, 6n + 3)$. (This problem is called the second Heffter difference problem). Then the set of all triples $\{0, a, a+b\}$, together with "short block" $\{0, 2n+1, 4n+2\}$, is a $(6n + 3, 3, 1)$ cyclic partial difference family; the base blocks for a $STS(6n + 3)$. Heffter's second difference problem is solved using extended Skolem sequences of order n with a hook in the nth position. From such a sequence, the pairs (b_r, a_r) is constructed such that $b_r - a_r = r$, for $1 \leq r \leq n$. Then the set of all triples $(r, a_r + n, b_r + n)$ is taken, for $1 \leq r \leq n$.

Below an explicit construction for the required Skolem sequences (as ordered pairs) :

$$n = 4s\ :\ \begin{cases} (r, 4s - r + 1) & r = 1, ..., s-1 \\ (s + r - 1, 3s - r) & r = 1, ..., s-1 \\ (4s + r + 1, 8s - r + 1) & r = 1, ..., s-1 \\ (5s + r + 1, 7s - r + 1) & r = 1, ..., s-1 \\ ((2s-1, 2s), (3s, 5s+1)), & (3s+1, 7s+1), (6s+1, 8s+1) \end{cases}$$

$$n = 4s+1,\ (n>5)\ :\ \begin{cases}(r, 4s-r+2) & r=1,...,2s\\ (5s+r, 7s-r+3) & r=1,...,s\\ (4s+r+2, 8s-r+3) & r=1,...,s-2\\ (2s+1, 6s+2), (6s+1, 8s+4), & (7s+3, 7s+4)\end{cases}$$

$$n = 4s+2,\ (n>2)\ :\ \begin{cases}(r, 4s-r+3) & r=1,...,2s\\ (4s+r+4, 8s-r+4) & r=1,...,s-1\\ (5s+r+3, 7s-r+3) & r=1,...,s-2\\ (2s+1, 6s+3), (2s+2, 6s+2), & (4s+4, 6s+4)\\ (7s+3, 7s+4), & (8s+4, 8s+6)\end{cases}$$

$$n = 4s-1\ :\ \begin{cases}(r, 4s-r) & r=1,...,2s-4\\ (4s+r+1, 8s-r) & r=1,...,s-2\\ (5s+r, 7s-r-1) & r=1,...,s-2\\ (2s, 6s-1), (5s, 7s+1), & (4s+1, 6s), (7s-1, 7s)\end{cases}$$

When $n = 2$, take the sequence :

$$\{(1,2),(4,6)\}$$

When $n = 5$, take :

$$(1,5),\ (2,7),\ (3,4),\ (8,10),\ (9,12)$$

When $n = 1$, the sequence does not exist. The construction above gives $STS(6n+3)$ with parameters $(v, b, r, k, \lambda) = (6t+3, (3t+1)(2t+1), 3t+1, 3, 1)$.

4) QC-OSMLD codes based on STS ($v = 12t+7$): This construction [10] is available for v a prime power in the form $v = 12t+7$. Let ω be a primitive root of the Galois field $GF(12t+7 = p^e)$. Then, the base blocks of a design with parameters $v = 12t+7$, $b = (2t+1)(12t+7)$, $r = 3(2t+1)$, $k = 3$, $\lambda = 1$ are given in the form $(\omega^{2i}, \omega^{2i+2t}, \omega^{4t+i})$

5) QC-OSMLD codes based on STS ($v = 12t+1$): This construction [6] is applicable for v a prime power in the form $v = p^e = 12t+1$. Let ω be the primitive root of $GF(p^e)$ such that $\omega^{4t} - 1 = \omega^q$ where q is odd. Then, the base blocks of a design with parameters $(v = 12t+1, b = t(12t+1), r = 3(12t+1), k = 4, \lambda = 1)$ are given in the form : $(0, \omega^0, \omega^{4t}, \omega^{8t})$, $(0, \omega^{2i}, \omega^{2i+4t}, \omega)$ $(0, \omega^{2t-2}, \omega^{6t-2}, \omega^{10t-2})$ such that $i = 0, ..., t-1$. Block designs [6] given above specify an infinite class of optimal QC-OSMLD $(n, k, d_{min}) = ((t+1)(12t+1), t(12t+1), 5)$ codes with basic block length $n_0 = (t+1)p^{e-1}$.

6) QC-OSMLD codes based on STS ($v = 20t+1$): This construction [6] is applicable for v a prime power in the form $v = p^e = 20t+1$. Let ω be the primitive root of $GF(p^e)$ such that $\omega^{4t} + 1 = \omega^q$ where q is odd. Then the base blocks of a design with parameters $(v = 20t+1, b = t(20t+1), r = 5t, k = 5, \lambda = 1)$ are given in the form : $(\omega^{2i}, \omega^{4t+2i}, \omega^{8t+2i}, \omega^{12t+2i}, \omega^{16t+2i})$ such that $i = 0, ..., t-1$. Block designs [6] given above specify an infinite class of optimal QC-OSMLD $(n, k, d_{min}) = ((t+1)(20t+1), t(20t+1), 6)$ codes with basic block length $n_0 = (t+1)p^{e-1}$.

C. New constructed codes

The following tables represent a small part of many constructed QC-OSMLD codes based on the methods described above. The length n, the dimension k, the minimum distance d, and the rate, and also the base blocks which represent the headers of the circulant matrix of the parity-check matrix H, are represented. Due to the significant number of base blocks in high rate, aren't represented in the tables.

TABLE II: QC-OSMLD codes based on STS $v = 12t+7$

n	k	d	Rate	Base blocks
76	57	4	0.75	1 4 16
186	155	4	0.833	1 19 20 9 16 29
344	301	4	0.87	1 4 41 9 12 25 10 36 38
804	737	4	0.92	1 19 26 4 9 52 16 36 37 7 10 64 14 40 55
4564	4401	4	0.96	-

TABLE III: QC-OSMLD codes based on STS $v = 6t+3$

n	k	d	Rate	Base blocks
50	35	4	0.7	1 3 4 2 6 8
91	70	4	0.77	1 4 6 2 5 8 3 8 11
144	117	4	0.81	1 5 6 2 7 10 3 8 12 4 11 13
206	176	4	0.85	1 6 10 2 7 12 3 8 9 4 13 15 5 14 17
286	247	4	0.86	1 7 12 2 8 11 3 9 15 4 10 14 5 14 16 6 16 17
851	782	4	0.919	-
1000	925	4	0.925	-
16800	16485	4	0.981	-

TABLE IV: QC-OSMLD codes based on STS $v = 6t + 1$

n	k	d	Rate	Base blocks
39	26	4	0.667	1 3 9 2 5 6
76	57	4	0.75	1 7 11 2 3 14 4 6 9
125	100	4	0.80	1 6 11 2 12 22 4 19 24 8 13 23
186	155	4	0.833	1 5 25 3 13 15 8 9 14 11 24 27 2 10 19
344	301	4	0.875	1 6 36 3 18 22 9 11 23 26 27 33 13 35 38 19 28 39 14 30 41
1649	1552	4	0.941	-
2071	1962	4	0.947	-
64898	64021	4	0.99	-

TABLE V: QC-OSMLD codes based on STS $v = 12t + 1$

n	k	d	Rate	Base blocks
75	50	5	0.666	0 1 6 11 0 4 19 24
148	111	5	0.75	0 1 10 26 0 3 4 30 0 9 12 16
245	196	5	0.80	0 1 25 37 0 9 29 38 0 16 23 32 0 32 43 46
366	305	5	0.833	-
511	438	5	0.86	-
873	776	5	0.89	-
1331	1210	5	0.91	-
10470	10121	5	0.966	-

TABLE VI: QC-OSMLD codes based on STS $v = 20t + 1$

n	k	d	Rate	Base blocks
244	183	6	0.75	1 9 20 34 58 4 14 19 36 49 13 15 16 22 56
405	324	6	0.80	1 7 19 49 52 4 28 34 46 76 16 22 31 55 61 1 7 43 58 64
606	505	6	0.83	-
847	726	6	0.86	-
1810	1629	6	0.90	-
4215	3934	6	0.93	-
6859	6498	6	0.95	-

V. Iterative Treshold Decoding

A. Encoding

In the case of QC codes, the encoding can be achieved with simple shift registers while the complexity is linear [9][16]. Because the quasi-cyclic code is not in systematic form, an additional k-stage register is required by this encoder to store the information symbols of the next block until encoding is completed. This difficulty can be avoided by using an equivalent systematic code. In this case the codes constructed are in systematic form.

QC codes could be encoded by either generator matrix or polynomial multiplication. These codes are defined by the parity check matrix H. The generator matrix G is obtained by the following transformation :

$$H = [PI_{n-k}] \Leftrightarrow G = [I_k P^T] \tag{6}$$

The encoding algorithm consists of multiplying the message i by the generator matrix G to get the codeword v.

$$v = i * G \tag{7}$$

The Quasi-cyclic codes have a polynomial form. Consider C(n,k) a systematic quasi-cyclic code with rate $\frac{1}{2}$, and let P which defines the code be the circulant matrix. The information vector to be encoded is denoted by i, then

$$v = iG = [i \ iP] \tag{8}$$

Let i(x) and p(x) represent the information vector and the header of the circulant matrix P in polynomial form, respectively. Obviously, the remaining rows of P are:

$$(xp(x), x^2p(x), ..., x^{k-1}p(x)) \ mod \ x^k - 1 \tag{9}$$

The algebra of polynomials modulo x^m -1 is equivalent to the algebra of m x m circulant matrices; besides, the polynomial product i(x)p(x) mod x^k-1 is similar to multiplying the vector i by the circulant matrix P. Hence,

$$v(x) = [i(x), i(x)p(x)] \tag{10}$$

Example 5: Let consider the same code as in the example 2. Now, to transmit the message i=101.
i = 101 then i(x) = $1+x^2$
And p = 101 then p(x) = $1+x^2$
Then the codeword is:
v(x) = [i(x),i(x)*p(x)]=[$1+x^2$,$(1+x^2)*(1+x^2)$]=[$1+x^2$,1+x]
$\Rightarrow$v = 101110.
In the case of QC codes with rate of the form $\frac{n_0-1}{n_0}$, the encoding like codes with rate 1/2 can be realised by either generator matrix or polynomial multiplication [9].

These codes are defined by the parity check matrix H of the form :

$$H = [P_1 P_2 ... P_{k_0} I_{n-k}] \tag{11}$$

The generator matrix G is obtained by the following transformation :

$$H = [P_1 P_2 ... P_{k_0} I_{n-k}] \Leftrightarrow G = \begin{bmatrix} & P_1^T & \\ & P_2^T & \\ I_k & \vdots & \\ & \vdots & \\ & P_{k_0}^T & \end{bmatrix}$$

In the case of codes with rate $\frac{n_0-1}{n_0}$, there are many circulant matrix (P_1 P_2 ... P_{k_0}), then for encoding the information vector 'i', it must be divided into k_0 subgroup (i_1 i_2 ... i_{k_0}) then based on the equation (10), the following equation is obtained :

$$v(x) = [i(x), \sum_{j=1}^{k_0} i_j(x)p_j(x)] \quad (12)$$

To clear this, let us consider the following example.

Example 6:

Let consider the OSMLD QC(15,10) code, with the minimum distance is d=3. The disjoint difference sets of order S=2 which define the parity matrix H of this code are $D_1\{0,1\}$ and $D_2\{0,2\}$.

The parity check matrix H in systematic form is [P_1 P_2 I_5]

$$\begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

After transformation of the parity check matrix H, the generator matrix G is obtained

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 \end{pmatrix}$$

Now transmitting the message i=1001001011. To know the codeword v corresponding to the message i, it's possible to use the classical method by using the generator matrix G v = i*G = 111011010110101. For using the polynomial method, the message must be divided into k_0=2 vectors
i_1=10010 $\Rightarrow$ $i_1(x)=1+x^3$
i_2=01011 $\Rightarrow$ $i_2(x)=x+x^3+x^4$
From the generator matrix G, headers of the two circulant matrix P_1^T and P_2^T is obtained:
c_1 = 10001 $\Rightarrow$ $c_1(x) = 1+x^4$
c_2 = 10010 $\Rightarrow$ $c_2(x) = 1+x^3$
Then, the equation (9) must be calculated:
$v(x)=[i(x),(i_1(x)*c_1(x))+(i_2(x)*c_2(x))]$
$=[1+x^3+x^6+x^8+x^9,((1+x^3)*(1+x^4))+((x+x^3+x^4)*(1+x^3))]$
$v(x)=[1+x^3+x^6+x^8+x^9,(1+x^2+x^3+x^4)+(x^2+x^3)]$
$=[1+x^3+x^6+x^8+x^9,1+x^4]$ Then, the codeword to transmit is :
v = 100100101110001 Which is the same as the codeword obtained by using generator matrix G.

B. ITD

Threshold decoding is simply the logical extension to soft decisions of majority decoding described above. In Massey's original work [17], he considered two different variations of the decoding algorithm. Considering here the method which uses the Bi equations that are obtained from A_i by a simple transformation [18].
Thanks to its speed and simplicity, the Majority Logic (ML) decoding of Quasi-Cyclic codes is significant. Therefore, it is worth investigating which Quasi-Cyclic codes can be decoded using ML decoder. Majority logic decoding is well described in [8,19]. It consists of cyclic shift register, XOR matrix, majority gate and XOR for correcting the codeword bit under decoding.

The ITD algorithm which is based on SISO extension of Massey threshold decoding algorithm [17] was developed mainly for decoding Parallel Concatenated Block Codes [20] and product codes constructed from OSMLD Codes [21], later it was enhanced for decoding not only Generalized Parallel Concatenated OSMLD Codes [22] but also OSMLD block codes [4]. ITD algorithm has an improvement in error correcting related to standard ML decoding.

Considering the transmission of the codeword $C(c_1,c_2,...,c_n)$ over an Additive White Gaussian Noise channel (AWGN), using BPSK modulation. The soft decision, which is the Log Likelihood Ratio (LLR), on the j^{th} bit of the received word $R(r_1,r_2,...,r_n)$ can be calculated as follows:

$$LLR_j = ln\left[\frac{p(c_j = 1/R)}{p(c_j = 0/R)}\right] \quad (13)$$

The hard decision vector corresponding to the received vector R is denoted by $H(h_1,h_2,...,h_n)$. Where c_j is the j^{th} bit of the transmitted codeword. For a code with J orthogonal parity check equations; the equation (13) can be expressed as:

$$LLRj = ln\left[\frac{p(c_j = 1/\{B_i\})}{p(c_j = 0/B_i)}\right] \quad (14)$$

Where B_i, for i in {1, ..., J}, are obtained from the orthogonal parity check equations on c_j bit, as follows:

$B_0 = h_j$ and each B_i with i in {1,...,J}, is calculated by eliminating the term h_j from the i^{th} orthogonal parity check equation. By applying BAYES rule, (14) becomes:

$$LLR_j = ln\left[\frac{p(\{B_i\}/c_j = 1)}{p(\{B_i\}/c_j = 0)} \times \frac{p(c_j = 1)}{p(c_j = 0)}\right] \quad (15)$$

Since the parity check equations are orthogonal on the j^{th} position, so the individual probabilities $P(B_i/c_j$ = 1 or 0) are all independent and (15) can be written as:

$$LLR_j = \sum_{i=0}^{J} ln\left[\frac{p(\{B_i\}/c_j = 1)}{p(\{B_i\}/c_j = 0)}\right] + ln\left[\frac{p(c_j = 1)}{p(c_j = 0)}\right] \quad (16)$$

Assume that the transmitted symbols are equally likely to be +1 or -1, and thus the last term in (16) is null. As a result, the equation (16) becomes:

$$LLR_j = \sum_{i=1}^{J} ln\left[\frac{p(\{B_i\}/c_j = 1)}{p(\{B_i\}/c_j = 0)}\right] + ln\left[\frac{p(\{B_0\}/c_j = 1)}{p(\{B_0\}/c_j = 0)}\right] \quad (17)$$

According to [18], (17) can be expressed as:

$$LLR_j \simeq (1-2B_0)w_0 + \sum_{i=1}^{J}(1-2B_i)w_i \qquad (18)$$

Where the value of (1-2B_i) is equal to +1 or -1, and w_i is a weighting term proportional to the reliability of the i^{th} parity check equation. then showing that:

$$(1-2B_0)w_0 = 4\frac{E_s}{N_0}r_j \qquad (19)$$

Where E_s is the energy per symbol, and N_s is the noise spectral density.

$$w_i = ln\left[\frac{1+\prod_{k=1,k\neq j}^{k=n_j}\tanh(\frac{L_{ik}}{2})}{1-\prod_{k=1,k\neq j}^{k=n_j}\tanh(\frac{L_{ik}}{2})}\right] \qquad (20)$$

Where n_i is the total number of terms in the i^{th} orthogonal parity check equation without c_j, *ik* represents the k^{th} element of the i^{th} parity check equation and with:

$$L_{ik} = 4\frac{E_s}{N_0} \mid r_{ik} \mid \qquad (21)$$

Thus the soft output can be split into two terms, namely into a normalized version of the soft input r_j and an extrinsic information L_{E_j} representing the estimates made by the orthogonal bits on the current bit c_j. Hence, (18) becomes

$$LLR_j = 4\frac{E_s}{N_0}r_j + L_{E_j} \qquad (22)$$

Using the following notation:

$$L_c = 4\frac{E_s}{N_0} \qquad (23)$$

Which is called the reliability value of the channel.

The algorithmic structure of the SISO threshold decoding can be summarized as follows:

For each j = 1,...,n

- Compute the terms B_i and w_i, i ∈ {1,..,J}
- Calculate B_0 and w_0
- Compute the extrinsic information L_{E_j}
- The Soft-output is obtained by:

$$LLR_j = L_c r_j + L_{E_j}$$

Iterative decoding process (see Figure 1) can be described as follows:

In the first iteration, the decoder only uses the channel output as input and generates extrinsic information for each symbol. In subsequent iterations, a combination of extrinsic information and channel output is used as input.

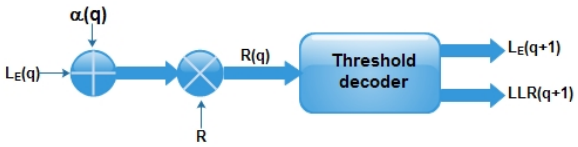


Fig. 1: Scheme of iterative threshold decoder

As shown in Figure 1, the soft input and the soft output of the q^{th} decoder is achieved through the following equations:

$$R(q) = R + \alpha(q)L_E(q) \qquad (24)$$

$$LLR(q+1) = L_c R(q) + L_E(q+1) \qquad (25)$$

Where R(q) represents lines (or columns) of the received data, and L_E(q) is the extrinsic information computed by the previous component decoder. In the proposed procedure, a fixed value 1/J is used for the parameter α(q) and this for all iterations. The value chosen for α(q) reacts as an average of all J estimators which contribute in the computation of L_{E_j}.

C. Modification of Rayleigh fading channel

In the channel model, each received bit rj can be expressed as :

$$r_j = a_j\widehat{c}_j + n_j \qquad (26)$$

In this representation, $\widehat{c}_j$ is a BPSK symbol associated to the transmitted bit cj, and nj is an AWGN. The Rayleigh variable aj is generated as:

$$a_j = \sqrt{x_j^2 + y_j^2} \qquad (27)$$

where x_j and y_j are zero mean statistically independent Gaussian random variables each having a variance σ^2. Considering the power normalized to one as

$$E[a_j^2] = \sigma^2 = 1 \qquad (28)$$

Which gives a variance of 0.5 for Gaussien variables.
The main matter in determining the required modification for ITD algorithm is the availability of channel side information on the Rayleigh fading channel. The threshold decoding algorithm has to be modified slightly by changing equation (23) which defines the reliability value of the channel by

$$L_c = 4\frac{E_s}{N_0}a_j \qquad (29)$$

With this modification, it's possible to use the same decoder structure which was described in Figure 1.

VI. Simulation Results and Analysis

This section considers simulation results and analysis for some decoding QC-OSMLD codes of rate $\frac{n_0-1}{n_0}$ and 1/2 with the Iterative Threshold Decoding algorithm. Some of our simulations are over AWGN channel , whereas others are over Rayleigh Fading channel; however, both of them are with modulation BPSK. Due to computational limitations, a minimal residual error of 200 have been used. In the simulations over Rayleigh fading channel, assuming an accurate fade estimate

at the receiving and an independent Rayleigh distribution of the fades.

The performance improves with each iteration in all simulation results presented. The following results represent the performance of decoding an QC-OSMLD code with the ITD algorithm and comparison with classic Threshold decoding algorithm.

A. AWGN

The Figure 2 depicts the performance of QC-OSMLD code (366,183,15). The improvement is great for the first iterations and is negligible after the 6^{th} iteration. Figure 3 presents a

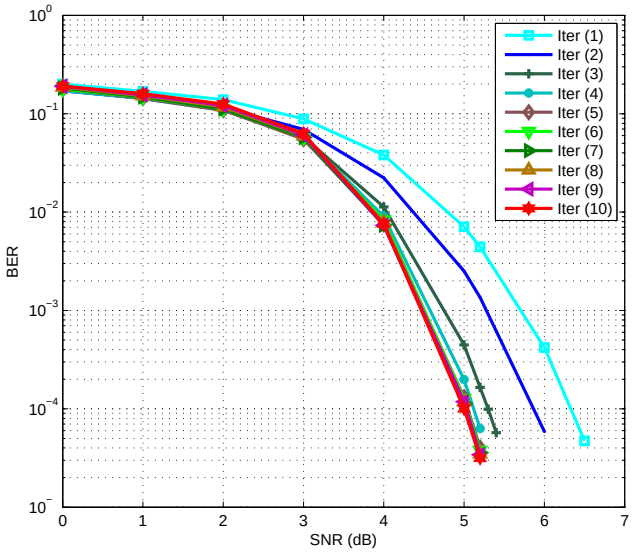


Fig. 2: The performance of the ITD algorithm with 10 iterations for decoding an QC-OSMLD code (366,183,15) over AWGN channel.

comparison between three QC-OSMLD codes (366,183,15) (1106,553,25)and the(4514,2257,49). Observing that smaller codes length has best performance at low SNR, whereas at SNR>5 for the code (4514,2257,49) the performance improves quickly from 10^{-1} to 10^{-5} between SNR=5 and SNR=6. The next comparison is between the code (182,91,11) decoded with ITD 10 iterations and the code LDPC WiMax(192,96,10), which is from [23], decoded with the Belief propagation decoder. These two codes are of the same rate $\frac{1}{2}$, and they have nearly the same dimension and minimum distance.

The Figure 4 shows that the LDPC WiMax code (192,96,10) decoded with BP algorithm outperforms the OSMLD-QC code (182,91,11) decoded with the ITD algorithm. However, the first decoder requires more iterations; on the other hand, the second one is less complex.

B. Rayleigh

The curves in the Figure 5 show the achieved bit error rates for the QC-OSMLD code (366,183,115) over Rayleigh fading channel. The number of iterations used is a 10, such that there is no significantly more to gain by more iterations.

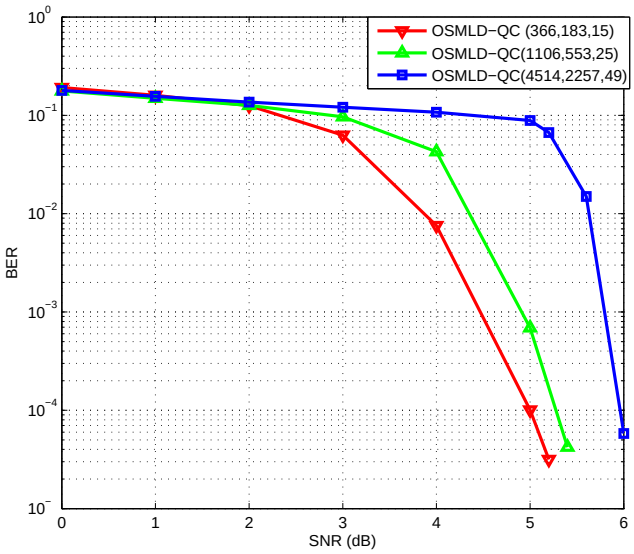


Fig. 3: Comparison between the performance of the QC-OSMLD codes (366,183,15), (1106,553,25) and (4514,2257,49) decoded with ITD 10 iteration, over AWGN channel

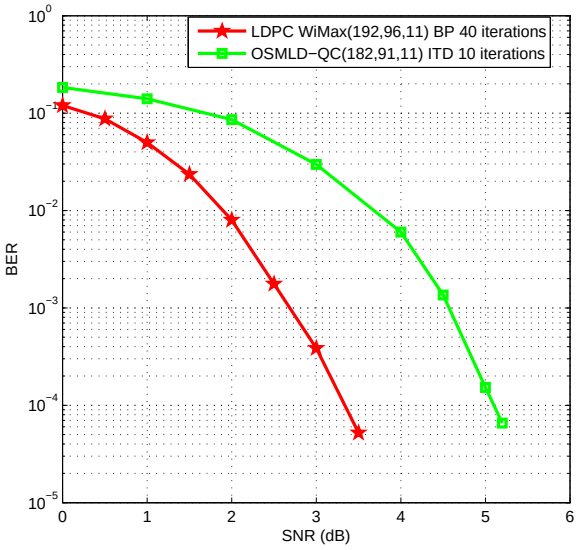


Fig. 4: Comparison between the performance of the code QC-OSMLD (182,91,11)decoded with ITD 10 iterations and and the code LDPC WiMAX(192,96,10)decoded with BP 40 iterations BP with 40 iterations

Observing that the performance increases with each iteration, and the improvement is negligible after the 7^{th} iteration .

Figure 6 presents a comparison between three QC-OSMLD codes (366,183,15) (1106,553,25)and the(4514,2257,49) over Rayleigh fading channel. Observing that the same behaviour as the AWGN channel, but in high SNR.

The Figure 7 shows the performance of decoding the QC-

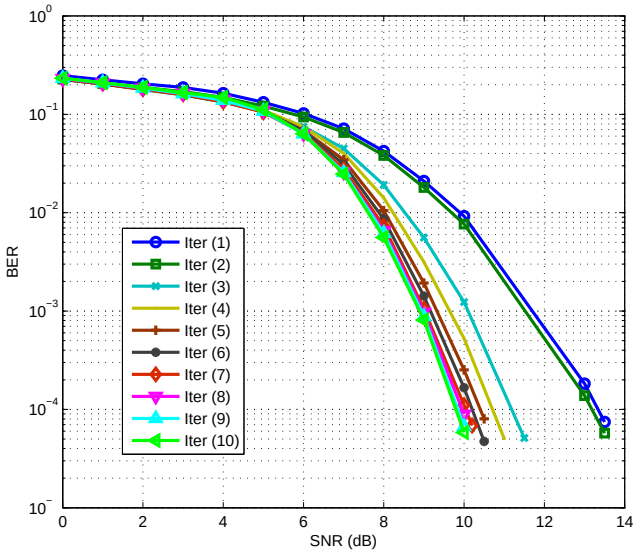


Fig. 5: The performance of the ITD algorithm with 10 iterations for decoding an QC-OSMLD code (366,183,15) over Rayleigh fading channel.

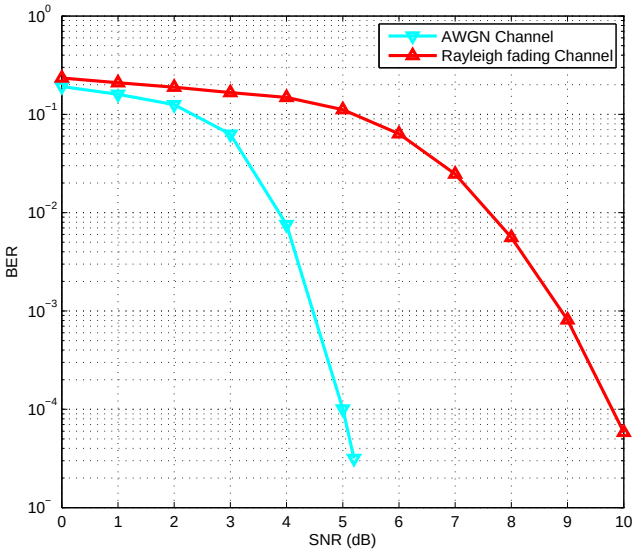


Fig. 7: Comparison between the performance of the ITD algorithm with 10 iterations for decoding an QC-OSMLD code (366,183,15) on AWGN channel and over Rayleigh fading channel

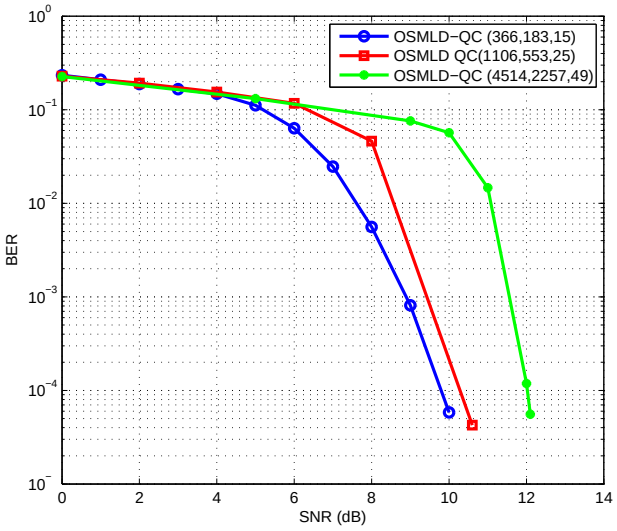


Fig. 6: Comparison between the performance of the QC-OSMLD codes (366,183,15), (1106,553,25) and (4514,2257,49) decoded with ITD 10 iterations, over Rayleigh fading channel

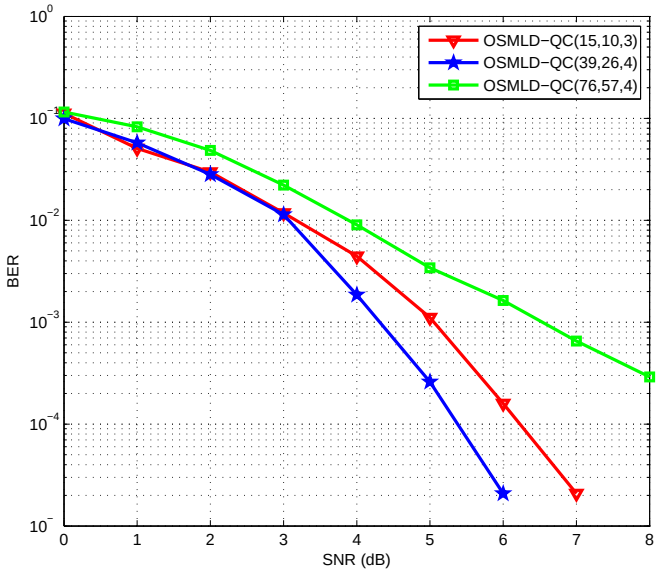


Fig. 8: Comparison between the performance of the QC-OSMLD codes (15,10,3), (39,26,4) of rate 2/3 and (76,57,4) of rate 3/4, decoded with ITD 10 iterations, over AWGN channel

OSMLD code (366,183,15) on both Rayleigh and AWGN channels. As observed in the other simulations, the performance of this code in an independent Rayleigh channel is worse than that for the AWGN channel by approximately 5 dB. It is worth mentioning that the number of iterations needed is about the same for the both channels.

The figure 8 shows a comparison between the performance of decoding the QC-OSMLD codes of different rates. To simplify, two scenarios have been opted. In the first case, there are two codes (39, 26, 4) and (15, 10, 3) which have the same rate 2/3. From the graph above, it's clear that as the code length increases, the performances rises, as well, which results in a gain of 1db at 10^{-5}. In the second case, two codes with different rates has been compared (76,57,4) of rate 3/4 and (39,26,4) of rate 2/3 holding the same minimum distance which is 4. As a result, even if the code length is large, the code of rate 3/4 is outperformed by the code of rate 2/3 with difference of 3db at 10^{-5}.

VII. Conclusion

In this paper, the construction of a class of QC-OSMLD codes based on Steiner triple system, and another class based on Singer difference sets has been investigated. The encoding methods has been presented for those codes. Also, the performances of decoding these codes with the ITD algorithm over AWGN channel and also over fading channel has been shown. The decoding algorithm used for AWGN channel is unchanged, and only the channel reliability factor needs to be redefined. The simulations results show that the constructed codes perform well when decoded with ITD algorithm. It is interesting to apply this iterative decoding algorithm on other channels models like Rice or Nakagami. Also as extension of this work we plan to investigate the performance of decoding rate $\frac{1}{n_0}$ QC-OSMLD codes with an adaptation of our ITD algorithm.

References

[1] R. G. Gallager,*"Low-Density Parity-Check Codes"*, Cambridge, MA: MIT Press, 1963.

[2] 802.16E-2005,802.16/COR1 IEEE Standard for Local and metropolitan area networks Part 16: Air Interface for Fixed and Mobile Broadband Wireless Access Systems Amendment for Physical and Medium Access Control Layers for Combined Fixed and Mobile Operation in Licensed Bands, 2/2006.

[3] European telecommunications standards institude(etsi), *"Digital video broadcasting (dvb) second generation framing structure for broadband satellite applications"*, en 302 307 v1.1.1. URL: www.dvb.org.

[4] M. Lahmer and M. Belkasmi, *"Iterative Threshold Decoding of One Step Majority Logic Decodable block Codes"*,ISSPIT Conf, December 15-18, 2007, pp. 668 - 673 Cairo, Egypt.

[5] L. Townsend and E. Weldon, *"Self-Orthogonal Quasi-cyclic Codes"*,IEEE on Information Theory, vol. IT-13, No 2, pp. 183-195, April 1967.

[6] C. Zhi, F. Pingzhy and J. Fan, *"On Optimal Self-Orthogonal Quasi-Cyclic Codes"*, in Communications, 1990. ICC '90, Including Supercomm Technical Sessions. SUPERCOMM/ICC '90. Conference Record., IEEE International Conference on , vol., no., pp.1256-1260 vol.3, 16-19 April 1990

[7] K. Rkizat M. Lahmer and M. Belkasmi, *"Iterative Threshold Decoding of Quasi-Cyclic One Step Majority Logic Decodable Codes"*,WICT'15 Conf, December 14-16, 2015, Marrakesh, Morocco.

[8] S. Lin and D.J. Costello, *"Error Control Coding: Fundamentals and Applications"*, Prentice-Hall, Englewood Cliffs, NJ, 1983.

[9] W. W. Peterson and E. J. Weldon, *"Error-Correcting Codes"*, 2nd ed. Cambridge, MA: MIT Press, 1972.

[10] C. J. Colbourn, J. H. Dinitz, *"Handbook of Combinatorial Designs"*, Second Edition, Chapman and Hall/CRC, 2007. ISBN-13 978-1584885061.

[11] J. Singer, *"A theorem in finite projective geometry and some applications to number theory"*, AIMS Trans., vol. 43, pp. 377-385, 1938.

[12] J. Cannon and W. Bosmaresolution, *"Handbook of Magma functions"*, Version 2.10 Sydney, May 2003.

[13] K. J. C. Smith, *"An Application of Incomplete Block Designs to the Construction of Error-correcting Codes"*, University of North Carolina. Department of Statistics, 1968 - 42 pages.

[14] B. Vasic and O. Milenkovic, *"Combinatorial Constructions of Low-Density Parity-Check Codes for Iterative Decoding"*, IEEE Transactions on Information Theory, Vol 50, No 6. June 2004.

[15] A. Rosa, *"Poznámka o cyklických Steinerových systémoch trojíc"*,Mat. Fyz. Časopis 16 (1966), 285-290.

[16] C.L. Chen, W.W. Peterson E.J. Weldon Jr., *"Some results on quasi-cyclic codes"*,.

[17] J.L Massey, *"Threshold Decoding"*,Cambridge, Mass., M.I.T. Press, 1963.

[18] C. Clark and B. Cain, *"Error-Correction Coding for digital communications"*,Plenum Press, 1981.

[19] L.D. Rudolph, *"A Class of Majority Logic Decodable Codes"*, IEEE Trans. Inf. Theory, vol. IT-13, pp. 305-307, May 1967.

[20] M. Belkasmi, M. Lahmer, and M. Benchrifa, *"Iterative Threshold Decoding of Parallel Concatenated Block Codes"*,Turbo Coding 2006 Conf., 4-7 April 2006, Munich, Springer.

[21] M. Belkasmi, M. Lahmer, and F. Ayoub, *"Iterative Threshold Decoding of Product Codes Constructed from Majority Logic Decodable Codes"*, ICCTA.06 Conf., pp.2376-2381, 24-28 April 2006, Damascus Syrie.

[22] F. Ayoub, M. Belkasmi, I. Chana, *"Iterative Decoding of Generalized Parallel Concatenated OSMLD Codes"*, Applied Mathematical Sciences journal, Vol.4,no.41, pp.2021-2038, 2010.

[23] H. Michael and S. Stefan, *"Database of Channel Codes and ML Simulation Results"*,University of Kaiserslautern, 2015. URL: www.uni-kl.de/channel-codes;

Multilingual Artificial Text Extraction and Script Identification from Video Images

Akhtar Jamil*, Azra Batool†, Zumra Malik†, Ali Mirza† and Imran Siddiqi†
*Yildiz Technical University, Istanbul, Turkey
†Bahria University, Islamabad, Pakistan

***Abstract*—This work presents a system for extraction and script identification of multilingual artificial text appearing in video images. As opposed to most of the existing text extraction systems which target textual occurrences in a particular script or language, we have proposed a generic multilingual text extraction system that relies on a combination of unsupervised and supervised techniques. The unsupervised approach is based on application of image analysis techniques which exploit the contrast, alignment and geometrical properties of text and identify candidate text regions in an image. Potential text regions are then validated by an Artificial Neural Network (ANN) using a set of features computed from Gray Level Co-occurrence Matrices (GLCM). The script of the extracted text is finally identified using texture features based on Local Binary Patterns (LBP). The proposed system was evaluated on video images containing textual occurrences in five different languages including English, Urdu, Hindi, Chinese and Arabic. The promising results of the experimental evaluations validate the effectiveness of the proposed system for text extraction and script identification.**

Keywords*—*Multilingual Text Detection; Video Images; Script Recognition; Artificial Neural Networks; Local Binary Patterns.

I. Introduction

Over the recent years, there has been a remarkable growth in the amount of multimedia data in the form of images, videos and audios. With the advancements in image/video capture hardware and the increase in the number of online image and video databases, digital multimedia content is likely to increase manifolds in the days to come. With this has increased the need to have efficient indexing and retrieval mechanisms allowing users rapid access to the content they are interested in. Among different types of multimedia data, the focus of our research interest lies on videos.

In addition to the visual content, videos comprise audio, text and other objects. The audio and visual information in the video could be effectively employed for development of semantic indexing and retrieval systems [1] and has been an attractive research area for over two decades now [2], [3]. In some cases, especially on the video sharing portals, users manually assign tags to videos allowing their retrieval. This retrieval, however, does not take into account the actual content of the video and is based on matching of tags only. In addition to the content of the video, a very powerful component, which could serve as an effective index, is the textual information in the video.

Text embedded in videos provides important, short and relevant information about the visual content. Examples of text occurrences include names of persons, sports scores, important dates, scene locations, movie credits, and stock rates etc. These embedded instances of text can be extracted and used as an effective index for retrieval from large video archival systems. As a result, development of automatic systems which could extract text from videos or images has been an attractive area of research in image analysis and pattern classification. Despite significant research on this problem, detection of textual information remains a challenging problem due to complex backgrounds, different font sizes and orientations and low contrast and resolution.

It is interesting to note that most of the research on this subject has focused on detecting text in a particular script. Properties of text in a particular script are exploited to detect its occurrences. Recently, there has been the trend of having multilingual text in videos especially the news channels where news tickers are flashed in multiple (generally two different) languages. It would be interesting to develop a generic system that could extract textual occurrences in videos or images irrespective of any language or script and this, in fact, is the subject of our study. The text detection module is generally integrated with text recognition (OCR) module to convert the occurrences of text in the image into text. For a detection system that works on a single script, the output of detector can directly be fed to the OCR module. In case of a multilingual detection system, however, the script of the detected text also needs to be identification so that it could be fed to the respective OCR system. This script identification has also been addressed in our work.

This work extends our previous contributions on text detection and extraction from video images [4], [5], [6].The main contribution of this research includes development of a generic text detection system in a multi-script environment which is not tuned to detect text in a particular language. The proposed approach is a combination of unsupervised and supervised techniques. In the first step, an unsupervised approach exploits the visual properties of text to segment candidate text regions using image analysis techniques. These candidate textual regions are validated by an Artificial Neural Network which is trained to differentiate between text and non-text blocks on the basis of a set of features extracted using the Gray Level Co-occurrence Matrices (GLCM). The developed system also identifies the script of the detected text using texture based features computed from the Local

Binary Patterns (LBP). The system evaluated on images with textual occurrences in five different languages (Urdu, English, Arabic, Chinese and Hindi) reports promising results on text detection as well as script recognition.

We first discuss the recent advancements in video text detection and extraction followed by the proposed methodology in Section III. Section IV describes the experimental evaluations conducted to validate the proposed methodology along with an analysis of the results realized. Finally, we conclude the paper with some ending remarks.

II. Background

Considering the applications it offers, detection of textual content from images and videos has been a highly researched area over the last decade. Text appearing in videos/images is generally classified into two categories, artificial text and scene text. Artificial text, also known as caption or superimposed text, is the text embedded and laid over the videos during the editing process to provide additional information related to its content such as news captions, sports scores, stock rates, etc. Scene text, on the contrary, is the text which appears naturally in the scene and is captured by the camera as a part of scene. Examples of scene text include text appearing on sign boards, billboards, names on shirts and vehicles etc. [7]. Detection and recognition of each category of text offers different types of applications. Scene text generally finds applications in robot navigation, license plate recognition and navigation of intelligent vehicles etc. Artificial text, which in general, is correlated with the content, is preferred for semantic indexing and retrieval of videos. Sample images containing occurrences of scene and artificial text are illustrated in Figure 1.

Fig. 1: Examples of (a) Scene text (b) Artificial text

In general, textual content based indexing and retrieval systems rely on four major steps namely text detection, localizaion, extraction and recognition. Text detection includes classification of a given region of interest as text or non-text region. Candidate text regions are fed to the localizer which finds the boundaries of text at character, word or line level depending upon the application. The localized text regions are then segmented from the background by the text extraction module. Finally, the extracted text regions could be fed to a recognition engine for conversion to text and subsequent indexing. Our research is aimed at extraction of text and subsequent identification of its script hence recognition of text is beyond the scope of our discussion.

Detection of text from images and videos has received notable research attention in the recent years. Traditionally, these methods are categorized into two broad classes, unsupervised and supervised techniques. The un-supervised approaches are based on image analysis techniques and use segmentation methods to differentiate text from other parts of the image. Supervised approaches for text detection employ machine learning algorithms to find text regions in an image. Traditionally, the supervised methods consist of two steps, training and classification. During training, features extracted from text and non-text regions are fed to a classifier to make it learn to differentiate between the two classes. During the classification phase, features extracted from the region in question are evaluated on the trained classifier which outputs the likelihood of the region as being text or non-text.

The unsupervised approaches for text detection mostly exploit the statistical and temporal features of text and, in general, work well in relatively less complex images. However, these methods may produce more false positive in complex scenes. The techniques used in this class of methods are further classified into gradient, connected component, texture and color clustering based methods.

Gradient-based methods [4], [8], [9], [10], [11], [12], [13] use edge information to segment the video images. They assume that there is high contrast between text and its background. Generally, an edge filter (e.g. Sobel or Canny operator) is applied for text detection, which is usually followed by some morphological processing to merge the desired edges to determine text lines [10], [14].

Texture based methods [15], [16], [17], [18] assume that text appearing in video frames has a unique texture that differentiates it from other objects in the image. Since the textural properties vary with font style and size, a generic texture filter for varying scenarios is hard to devise [1]. In addition, the computational complexity of these methods is also high as they require an exhaustive scan of whole image for text detection and localization.

Connected component based methods [19], [20], [21] either use region growing or splitting approach in order to group text pixels into clusters until all regions in the input image are identified. These methods are widely used for text localization due to their simple implementation. However, since these methods mainly rely on the contrast between text and background, they produce false alarms in case of low resolution images.

Color based methods [22], [23], [24], [25] use color information to cluster image content into text and non-text regions. These methods perform well for images with high

resolution and simple backgrounds. However, these assumption may not be true in many real world scenarios where text may appear in various colors and can be superimposed on complex backgrounds. In addition, due to compression, images may suffer from color bleedings affecting the performance of color based methods.

In supervised approaches for text detection, a learning machine is first trained on a set of features extracted from both text and non-text samples. Generally, these features are extracted by scanning the image with a small window which are then fed to the classifier. Calssifiers like support vector machine (SVM) and artificial neural networks (ANN) have been extensively applied for this purpose [26], [27], [28], [29], [30], [31]. In some cases, coarse-to-fine algorithms have also been evaluated where the candidate text pixels are first identified and then valiated by a classifer [32], [33].

With few exceptions, most of the text detection methods reported in the literature target text in a particular script. The literature is very rich when it comes to detect text in any of the languages based on the Latin alphabet (English, French, and German etc.). Detection of caption text in Chinese has also witnessed a significant research attention. For most of the other scripts, the research is either in its early days or is non-existent. In our proposed system, we aim to develop a generic text detection system that is not tuned to detect text in any particular script and works on multilingual text as detailed in the following section.

III. Proposed Methodology

This section presents in detail the proposed methodology for text detection and script identification. As discussed earlier, the target application of such text detection systems is indexing and retrieval of videos. The general architecture of such a system is illustrated in Figure 2. Textual information extracted from videos is fed to an Optical Character Recognition (OCR) system to convert it into text. The focus of our research, however, is on the first part, i.e. detection and extraction of text and identification of the script of the detected text.

The proposed system can be divided into three main modules. An unsupervised approach is first used to detect potential text regions. These text regions are validated through a supervised approach that employs an artificial neural network as classifier. Finally, the script of the extracted text is recognized using texture based features. Each of these modules is discussed in the following sub-sections.

A. Text Detection

For detection of potential text regions in the image, the image is first converted to grayscale [5]. A sequence of image analysis techniques is then applied to the image as discussed in the following.

1) Gradient Computation: Edges are a common feature of text in all scripts. Different scripts have different proportions of horizontal, vertical and diagonal edges corresponding to text strokes in each of these directions. In our study, we consider text in Urdu, English, Chinese, Arabic and Hindi, an example

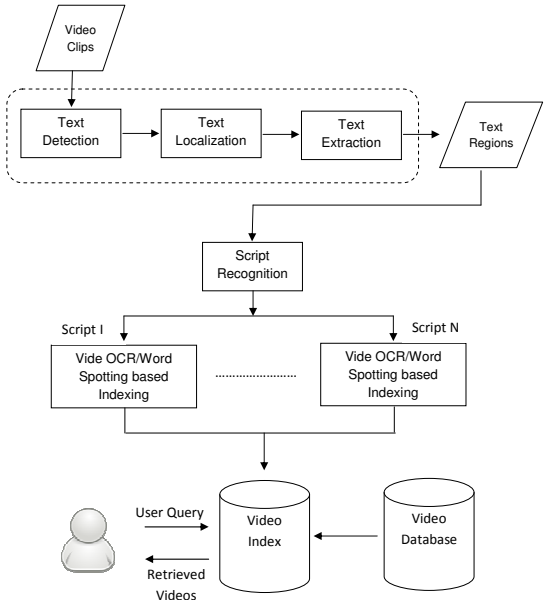


Fig. 2: General framework of a video indexing and retrieval system

of each being shown in Figure 3. It can be seen that in all of these texts, a reasonable proportion of strokes are vertical.

Fig. 3: Samples of text in (a) Urdu (b) Arabic (c) Chinese (d) Hindi (e) English

In our implementation, vertical edges are computed using the first derivative (gradient) by convolution of the image with the respective Sobel mask.

Figure 4 illustrates two images and their respective (vertical) gradient images. It should be noted that objects other than text may also respond to the gradient operator. Hence, the gradient image, in addition to text strokes may also contain many unwanted edges which are removed in the subsequent steps.

2) Mean gradient: The textual content in images occurs in clusters hence a number of studies consider enhancing the magnitude of image gradients in the text regions while suppressing it in the non-text areas. Generally this is achieved by scanning the gradient image with a small window and performing some operations [10], [8]. Authors in [10] exploit this idea using accumulated gradients where the gradient values in a predefined sliding window are accumulated. Shivakumara [8] employed the difference of the maximum and minimum values of pixels in a fixed neighborhood to calculate the value of central pixel in each window. In our study, we slide a horizontal

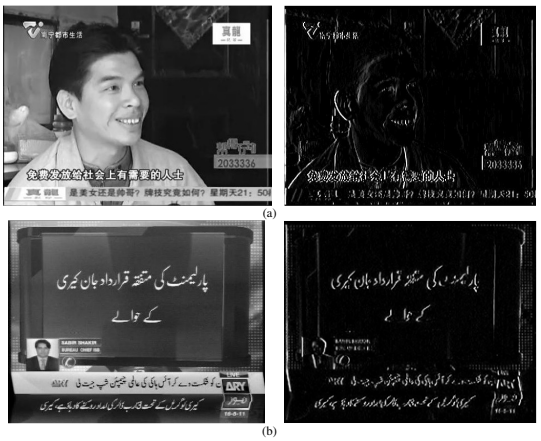


Fig. 4: Vertical gradient images (a) Chinese text (b) Urdu text

window of size $1 \times s$ on the gradient image and replace each pixel with the average of the gradient magnitude in the window [4]. The motivation behind this operation is that edges in text regions appear in clusters. Hence, computing the average gradient in windows over text regions is likely to maintain high values. On the other hand, isolated gradients in the non-text regions, when replaced by the mean of neighboring pixels, are suppressed [4]. Equation 1 summarizes the average gradient operation, s being the size of averaging window which is empirically fixed to 31 in our study.

$$Avg(x,y) = \frac{1}{s}\left[\sum_{j=-s/2}^{s/2} G(x+j,y)\right] \tag{1}$$

The averaged gradient image is binarized to have text or text-like regions as white pixels on black background. Binarization threshold is computed using Otsu's global thresholding algorithm. As a result of binarization, gradients with weak magnitude are removed (become a part of background) and text-like regions are retrained which are merged together by applying morphological operations on the binarized image.

3) Morphological Processing: In order to combine the binarized gradients into larger components, we apply horizontal run-length smoothing algorithm (RLSA). As a result of this, components in the proximity of one another are merged together while the isolated components remain separated. It can be seen from Figure 5 that most of the textual content is merged into large components which correspond to words or groups of words.

4) Foreground Density Filter: Applying the horizontal RLSA to the binarized averaged gradients joins most of the textual elements into larger components. The image, however, still contains non-text components which need to be addressed. Exploiting the same idea that text components appear in clusters, we next employ a density filter on the image using a rectangular sliding window. The window is moved in the top-bottom, left-right fashion and for each position of window the density of foreground (likely text) pixels is computed as.

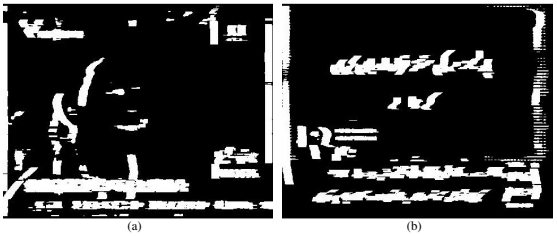


Fig. 5: Application of RLSA to averaged-gradient images (a) Chinese (b) Urdu

$$\text{Foreground Density} = \frac{\text{Number of white pixels}}{\text{Total pixels}} \tag{2}$$

The foreground pixel density is compared to a pre-defined density threshold. If the pixel density at a given window position is greater than the threshold, the central pixel is assigned a value 1, else it is considered a non-text pixel and is assigned a 0.

$$h(x,y) = \left\{ \begin{array}{ll} 1 & \text{if density(x,y)} > t \\ 0 & \text{otherwise} \end{array} \right\} \tag{3}$$

Where t is the density threshold set to 0.8 while the window size is fixed to 10×10 pixels. As evident from Figure 6. The density filter, although effective, does not suppress all

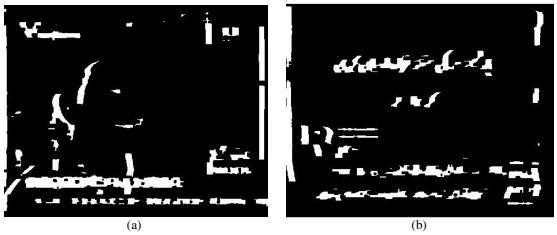


Fig. 6: Images after application of foreground density filter (a) Chinese (b) Urdu

the unwanted non-text regions. We, therefore, apply some geometrical constraints on the detected components to further reduce the false alarms.

5) Geometrical Constraints : With the realistic assumptions that size of the text on the image is large enough to be read by the audience, traditional geometrical constraints are applied to the localized bounding boxes. Another important property, as discussed earlier, is that text components are likely to occur in groups and not in isolation. Similarly, since we target horizontally aligned text, constraints can be applied to the aspect ratio of such text. Components satisfying the empirically determiend thresholds on aspect ratio, minimum height and minimum width are kept as potential text regions while the remaining components are discarded. Figure 7 illustrates the components retained as text after application of geometrical constraints on the two example images used as reference in our description.

Fig. 7: Images after application of geometrical constraints (a) Chinese (b) Urdu

After having discussed the detection of potential text regions using an unsupervised approach, we present the validation mechanism of these detected text rectangles in the next section.

B. Text Validation

The output of the text detector mostly comprises valid text regions. However, some other objects, which exhibit text like properties, are also falsely detected as text regions. The objective of validation step is to take as input each text block localized by the detector and validate it using a supervised approach. This module comprises two phases, training and validation, each of these is discussed in the following.

1) Training : A unique property of text in any script is its texture which can be exploited to distinguish it from other objects or complex backgrounds. Texture information can be captured using a variety of measures. In our implementaiton, we compute a set of features from the Gray Level Co-occurrence Matrices (GLCM) of text and non-text blocks to represent the texture. These features are then used to train a classifier, an artificial neural network in our case, to learn to discriminate text and non-text regions.

Training of the classifier requires samples of text and non-text blocks. We have used a training data set which comprises video images containing textual occurrences; 30 images for each script making a total of 150 images. The text rectangles in each image are manually extracted while rest of the image is considered as non-text region. For each text and non-text rectangle, we divide it into small blocks of 30×50. This gives a large number of text and non-text blocks which constitutes our training data. Some examples of text and non-text blocks can be seen in Figure 8.

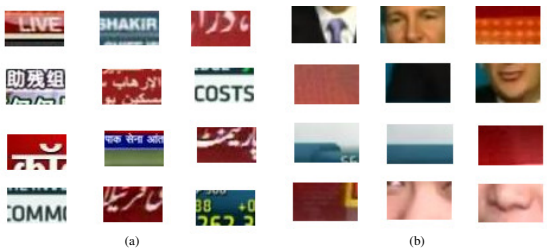


Fig. 8: Blocks used to train the neural network (a) Text blocks (b) Non-text blocks

Each block (text or non-text) is converted to grayscale and a GLCM is computed for each block. The GLCM considers the relationship among two neighboring pixels and determines how frequently different combinations of gray levels co-occur for a given direction and distance. The size of GLCM matrix is the same as the number of gray levels in the image. It is therefore a common practice to quantize the gray levels to have a smaller GLCM. In our implementation, we quantize each block to 64 gray levels and compute the GLCMs using four displacement vectors (offsets). These offsets include (0,1), (1,-1), (0,-1) and (-1,-1) and correspond to four directions $0°$, $45°$, $90°$, $135°$.

Once the GLCMs are computed, several statistics can be computed from each GLCM and could serve as features to characterize the underlying texture of the input image (block). In our study, we compute the contrast, correlation, homogeneity, entropy and energy of each GLCM and use them as features to characterize each block. These statistics are summarized in Table I. These five statistics are computed for each of the four GLCMs $(0°, 45°, 90°, 135°)$ for each training block. Finally, the average of each feature for the four directions is computed giving a 5 dimensional feature vector [34]. These features are fed to a feed forward artificial neural network. In our implementation, we use a neural network with 5 neurons in the input layer (corresponding to five features), 20 neurons in the hidden layer (chosen experimentally) and two neurons in the output layer, each neuron with a sigmoid activation function. The network is trained on 396 text blocks and 938 non-text blocks using back propagation algorithm.

SNo.	Feature	Computational Details
1.	Contrast	$\sum_{i,j=0}^{N-1} P_{i,j} = (i,j)^2$
2.	Correlation	$\sum_{i,j=0}^{N-1} P_{i,j} \left[\frac{(i-\mu_i)(j-\mu_j)}{(\sqrt{(\sigma_i^2)(\sigma_j^2)})} \right]$
3.	Homogenity	$\sum_{i,j=0}^{N-1} P_{i,j} = 0 \frac{P_{i,j}}{i+(i-j)^2}$
4.	Entropy	$\sum_{i,j=0}^{N-1} P_{i,j}(-InP_{i,j})$
5.	Energy	$\sum_{i=0}^{N-1} \sum_{j=0}^{N-1} [P(i,j)]^2$

TABLE I: Summary of GLCM based features

2) Validation of Text regions: The trained neural network is employed to validate the candidate text regions produced by the detection module. Each detected rectangle is divided into blocks which are fed to the network for classification. If more than 60% of the blocks in a detected rectangle are classified as text, the rectangle is retained as a valid text region. Otherwise, it is considered a false positive and is discarded. This validation step is intended to remove the false alarms and improve the overall precision of the system. A relaxed threshold of 60% is used so that valid text regions are not eliminated during this step and recall of the system is not compromised. The final text rectangles are then separated from the background using the text extraction module discussed in the following.

(a) (b) (c) (d) (e)

Fig. 9: Text detection and extraction examples in five different languages

C. Text Extraction

Text extraction is the step where the text components are segmented from the background. This step is straight forward if the background is homogenous but can pose difficulties on complex backgrounds. A number of global and local thresholding algorithms have been propsoed to segment text from the background both in scanned document images and video frames [35], [36], [37], [38], [14]. In our implementaiton, we employ the Wolf's algorithm [14] which has been specifically developed for segmentation of video text from the background and is known to work better than many of the binarization algorithms. Examples of text extracted using Wolf's binarization [14] can be seen in Figure 9.

This concludes our discussion on text detection which comprised detection of potential text regions, validation of these regions and segmentation of text from the background. We now present the script identification in the next section.

D. Script Identification

Script identification is aimed at identifying the script of the text detected by the detection module. Literature on script identification of video text is relatively limited as most of the text detection systems have been designed to operate on text in a known language. The existing literature on this subject is mostly on document images only and script identification from text in videos has been a less investigated area. In case of printed and handwritten document images, features at page, paragraph, line and word level have been explored for identification of script [39], [40], [41]. Among recent video text script identification methods, supervised [42] as well as unsupervised [43] techniques have been employed.

For detection of multi-script text, the objective is to find the common properties of text in different scripts and exploit these properties to allow its detection. In script recognition, the objective is to exploit the variations between different scripts. In our study, we consider text in each script as a different texture and employ Local Binary Patterns (LBP) to capture the texture information. The histograms of LBPs computed from texts in different scripts are used to train a neural network which then classifies a given text as being one of the script classes.

1) Local Binary Patterns: Local Binary Patterns, introduced by Ojala [44], [45] for texture classification, have been effectively applied to wide variety of texture classification problems [46], [47], [48], [49]. The original LBP feature [44], [45] considers for each pixel V_0 a set of neighboring pixels. The pixel values of all the neighbors are compared with the value at central pixel. If the value of a neighboring pixel is less than the central pixel, the neighbor is assigned a value of 0, otherwise, it is assigned a 1. The resulting string of 0s and 1s is considered a binary number. The computation of LBP for a reference pixel is illustrated in Figure 10.

In a later study [50], the authors proposed extensions to the original LBP operator to take into account neighborhoods of different sizes. The generalized LBP is represented using the notation (P, R), where P represents the number of neighboring pixels while R is the distance of the neighboring pixels from the central pixel. In addition, based on the number of transitions between 0s and 1s, uniform and non-uniform binary patterns were introduced. LBP codes for which the number of transitions is less than or equal to 2 are considered uniform while those with more than 2 transitions are considered non-uniform [50].

To generate an LBP based descriptor of texture, the LBP is computed for each pixel in the image and the histogram of LBP is used as feature to characterize texture. In our implementation, we compute the $(16, 2)$ LBP from the grayscale images of text blocks with dark text on bright background. For 16 neighboring points, this gives a 243 dimensional feature vector characterizing the texture of each script.

2) Training and Classification: An artificial neural network is used as classifier to recognize the script. The neural network is trained using the same training set that was used to train the network for text validation. Text rectangles from a total of 150 images, with 30 images per script are used as training data. The LBP histogram is computed from each image and the extracted histograms are fed to the network for training. The network comprises 243 neurons in the input layer (same as dimension of the feature vector/histogram), 200 neurons in the hidden layer and 5 neurons in the output layer (corresponding to 5 scripts). For recognition, the LBP histogram is determined from the detected text rectangle and is fed to the network which classifies it as being English, Arabic, Urdu, Hindi or Chinese text.

IV. Experiments and Results

All experiments are carried out on the multi-lingual artificial text database developed at Image Processing Center (IPC) - a research facility at National University of Sciences and Technology (NUST), Pakistan. The database comprises a total of 500 video frames extracted from different news channels, sports videos, talk shows etc. These images contain occurrences of artificial text in five different languages namely English, Arabic, Urdu, Chinese and Hindi with 100 images of each category as the major text of the image. A subset of this data set (images with Urdu text) has been published as [51]. The resolution of the images varies from a minimum of 320x240 to a maximum of 720x576 pixels. Out of the 100 images of each category, 30 images are used as training data (for training the ANN for text region validation and script identification) while 70 are used for testing. The ground truth data for the images was generated by labeling the text occurrences and storing the coordinates of each text rectangle.

Several evaluation metrics have been proposed to evaluate the performance of text localization systems [52], [53]. In our system, we have employed the area based precision and recall measures. Let A_E be the estimated text area given by the system and A_T be the ground truth text area, then the precision P and recall R are defined as:

$$P = \frac{A_E \cap A_T}{A_E} \tag{4}$$

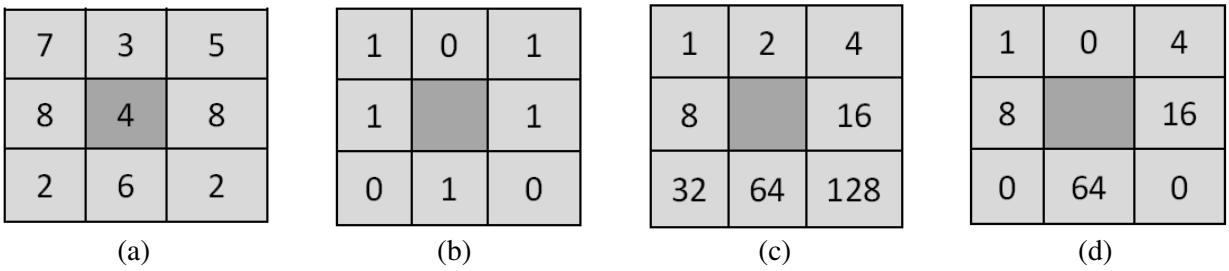


Fig. 10: Calculation of LBP (a): Pixel values (b): Binary codes (c): Weight assignment (d): Decimal number

$$R = \frac{A_E \cap A_T}{A_T} \tag{5}$$

The same idea can be extended to N images to compute the overall precision and recall values. For script recognition experiments, we report the confusion matrix and the overall correct classification rate of the system.

A. Text Detection Results

The text detection module first identifies potential text regions using an unsupervised approach. These candidate text rectangles are then validated by a supervised approach to find the final set of text regions. Detection results, in terms of precision and recall, for both of these are summarized in Table II and Table III respectively. Using the unsupervised detection scheme, an overall precision of 59% and a recall of 89% is achieved. It is interesting to note that the results are consistent across text in different languages demonstrating the generality of the system.

Language	Precision	Recall	F-measure
Urdu	0.58	0.84	0.69
English	0.61	0.92	0.73
Arabic	0.59	0.89	0.71
Chinese	0.60	0.87	0.71
Hindi	0.58	0.94	0.72
Total	**0.59**	**0.89**	**0.71**

TABLE II: Precision and recall of text detection (unsupervised)

It can be seen from Table II that precision values are lower than that of recall values. There are mainly two reasons for this. The first reason is that the system parameters are tuned to achieve high recall and, low values of precision at the detection step are acceptable. The next step of text validation is aimed to reject the false alarms and improve the precision of the system. Since validation cannot detect the text regions which are missed by detection, the recall cannot be improved by the validation step and hence high values of recall are desired at the detection step. The second reason is that we are using an area based metric to compute precision and recall where area represents the number of pixels. Figure 11 illustrates an example of the ground truth text region and the text region detected by the system. Although the system has detected the text but since all three text regions are merged in one big rectangle (having background pixels in the detected region), this results in a low precision.

Language	Precision	Recall	F-measure
Urdu	0.65	0.80	0.72
English	0.68	0.88	0.77
Arabic	0.66	0.85	0.74
Chinese	0.66	0.83	0.73
Hindi	0.60	0.87	0.71
Total	**0.65**	**0.85**	**0.74**

TABLE III: Precision and recall after text validation

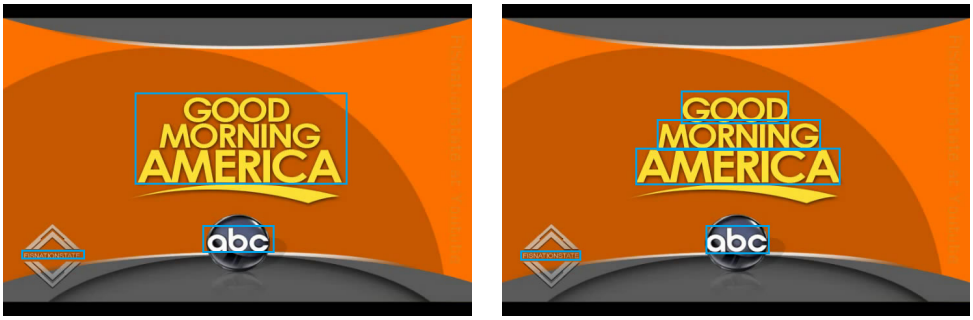
It should be noted that the idea of having a validation step after detection is to enhance the precision of the system by rejecting the regions falsely detected as text. Although precision values in Table III are better than those in Table II, there is a slight decrease in the recall values. This is because while false alarms are reduced by the validation step, some text regions are also eliminated. Overall, however, increased values of F-measure reflect the usefulness of this validation step.

B. Script Identification Results

Script identification is aimed at identifying the script of the text extracted from the images. From the view point of application, script identification module should be fed the output of text detector. However, since the text detection does not extract all the text rectangles, script recognition experiments are carried out on manually extracted text blocks. This allows evaluation of script recognition on all the text blocks in our dataset. Out of a total of 1,448 text blocks, the script of 1,291 blocks was correctly recognized making it a classification rate of 89%. The detailed confusion matrix is illustrated in Table IV where it can be observed that the performance of script identification is more or less consistent across text in different scripts.

	Arabic	**English**	**Urdu**	**Hindi**	**Chinese**	**Total**
Arabic	192	9	11	3	12	227
English	4	349	3	4	20	380
Urdu	2	14	202	2	5	225
Hindi	1	11	2	266	16	296
Chinese	0	20	7	11	282	320

TABLE IV: Script Recognition - Confusion matrix



(a) (b)

Fig. 11: (a) Detected text region (b)Ground truth text region

For script identification, we have used the histogram of local binary patterns using $(16, 2)$ neighborhood ($LBP_{(16,2)}$). By varying the neighborhood size, we study the variation in the classification rate as illustrated in Figure 12. Neighborhoods of (8,1), (8,2), (8,3), (16,1), (16,2) and (16,3) have been considered in out experiments. It can be observed from Figure 12 that the script recognition rates are not very sensitive to the neighborhood size with neighborhoods of 16 pixels naturally performing better than those of 8 pixels.

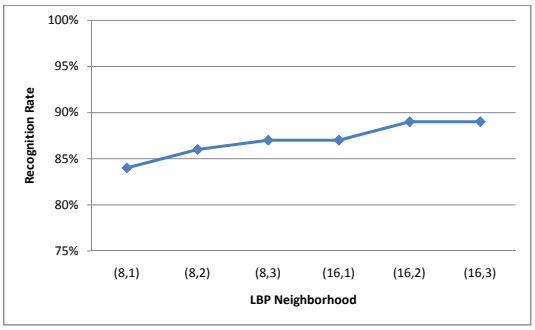


Fig. 12: Script identification rates as a function of different neighborhoods of LBP

We also performed a comparative analysis of the proposed system with well-known existing systems in the literature. The comparison can be carried out for text detection as well as script recognition. Text detection, however, has been evaluated by different metrics in different studies hence a meaningful comparison may not possible. We, therefore, present a comparison of the performance of different script recognition systems in Table V. It can be seen from the table that the database employed, the number of scripts and the number of images in each study is different making it difficult to perform a direct comparison of recognition rates. A maximum of 10 different scripts have been considered in [54] realizing a recognition rate of 91%. The system, however, has been evaluated on 100 test images only. The recognition system in [43] reports a correct classification rate of around 96% on 770 test images which indeed is very promising. Our proposed LBP based technique realizes a recognition rate of 89% on 500 test images in 5 different scripts. These results are comparable with most of the studies and we look to improve them further by introduction of other texture based features to complement the LBP features.

V. Conclusion

This work presented a system for detection of multilingual artificial textual content from video images, an important component for text based indexing and retrieval of videos. Script recognition was also considered in our study. Most of the state-of-the-art approaches for text detection target a single script/language. We have presented a generic text detection system that is not tuned on one particular type of text. The detection is implemented using a combination of unsupervised and supervised techniques. The unsupervised approach relies on image analysis techniques including edge information, morphological processing and geometrical heuristics to detect potential text regions in an image. These candidate text regions are then validated by an artificial neural network that is trained on text and non-text blocks using a set of texture features computed from Gray Level Co-occurrence Matrices (GLCMs). The proposed methodology evaluated on images containing textual occurrences in five different languages (Urdu, Arabic, Hindi, English and Chinese) realized promising results.

We also presented a script recognition module that takes text blocks as input and recognizes the script of the text. Each script is viewed as a different texture and the texture information is captured by computing the histogram of Local Binary Patterns. Recognition is carried out by an artificial neural network trained on text blocks from the five scripts considered in our study. The main idea of this module is to identify the script of the text rectangles detected in the images so that these rectangles can be further processed by their respective recognition engines.

The proposed system which presently targets extraction of text from images and recognition of the script of detected text can be extended to a complete video indexing and retrieval system. This will require either integration of recognition engines (for each of the scripts) or a word spotting based technique allowing indexing of videos on the extracted textual content. The video OCR itself is a challenging problem due to low resolution and complex backgrounds as opposed to document OCRs. Another interesting aspect which could be exploited is the temporal redundancy of text in videos. The

Study	Scripts	Languages	Data set	Overall Recognition Rate
[43]	6	English, Chinese, Japanese, Korean, Arabic and Tamil	770 images	95.71%
[40]	4	Chinese, Japanese, Korean and Roman	3200-3500 characters each	96.95% at character level and 99.85% at block level
[39]	4	English, Urdu, Hindi and Kannada	400 images	97%
[54]	10	Arabic, Cyrillic, Greek, Hebrew, Japanese, Roman, Bengali, Thai, Korean and Chinese	100 images	91%
[55]	3	English, Tamil and Chinese	500 images	51.6%
[42]	3	English, Hindi and Bengali	896 images	87.5%
Proposed method	**5**	**English, Urdu, Hindi, Chinese and Arabic**	**500 images**	**89%**

TABLE V: A comparison of script recognition systems

present system works on static images and does not take into account the redundancy that exists across multiple frames in a video. Integrating the detection results of multiple frames could serve to enhance to overall accuracy of the system. It is expected that the ideas put forward in this research would be helpful to researchers working on video retrieval systems in general and text extraction in particular.

REFERENCES

[1] K. Jung, K. In Kim, and A. K. Jain, "Text information extraction in images and video: a survey," *Pattern recognition*, vol. 37, no. 5, pp. 977–997, 2004.

[2] J. Wu, A. Narasimhalu, B. Mehtre, C. Lam, and Y. Gao, "Core: a content-based retrieval engine for multimedia information systems," *Multimedia Systems*, vol. 3, no. 1, pp. 25–41, 1995.

[3] S.-F. Chang and H. Zhang, "Content-processing for video browsing, retrieval, and editing," *Multimedia Systems*, vol. 7, no. 4, pp. 255–255, 1999.

[4] A. Jamil, I. Siddiqi, F. Arif, and A. Raza, "Edge-based features for localization of artificial urdu text in video images," in *Proc. of International Conference on Document Analysis and Recognition*, 2011, pp. 1120–1124.

[5] A. Raza, A. Abidi, and I. Siddiqi, "Multilingual artificial text detection and extraction from still images," in *Proc. of Document Recognition and Retrieval, IS&T/SPIE Electronic Imaging*, 2013, pp. 86 580V–86 580V.

[6] A. Raza, I. Siddiqi, C. Djeddi, and A. Ennaji, "Multilingual artificial text detection using a cascade of transforms," in *Proc. of the 2013 12th International Conference on Document Analysis and Recognition*, 2013, pp. 309–313.

[7] S. Antani, D. Crandall, A. Narasimhamurthy, V. Mariano, and R. Kasturi, "Evaluation of methods for detection and localization of text in video," *Proc. of the IAPR workshop on Document Analysis Systems*, pp. 506–514, 2000.

[8] P. Shivakumara, T. Q. Phan, and C. L. Tan, "A laplacian approach to multi-oriented text detection in video," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 2, pp. 412–419, 2011.

[9] T. B. Chen, D. Ghosh, and S. Ranganath, "Video-text extraction and recognition," in *Proc. of IEEE Region 10 Conference (TENCON)*, vol. 1, 2004, pp. 319–322.

[10] C. Wolf, J.-M. Jolion, and F. Chassaing, "Text localization, enhancement and binarization in multimedia documents," in *Proc. of the International Conference on Pattern Recognition*, 2002, pp. 1037–1040.

[11] L. Minhua and B. Meng, "A mixed edge based text detection method by applying image complexity analysis," in *Porc. of the 10th World Conference on Intelligent Control and Automation*, 2012, pp. 4809–814.

[12] P. Dubey, "Edge based text detection for multi-purpose application," in *Porc. of the 8th International Conference on Signal Processing*, 2006.

[13] A. Ikica and P. Peer, "An improved edge profile based method for text detection in images of natural scenes," in *Proc. of International Conference on Computer as a Tool (EUROCON)*, 2011.

[14] C. Wolf and J.-M. Jolion, "Extraction and recognition of artificial text in multimedia documents," *Pattern Analysis and Applications*, vol. 6, no. 4, pp. 309–326, 2004.

[15] C. Zhu, W. Wang, and Q. Ning, "Text detection in images using texture feature from strokes," in *Advances in Multimedia Information Processing*, ser. Lecture Notes in Computer Science, 2006, pp. 295–301.

[16] V. Wu, R. Manmatha, and E. Riseman, "Textfinder: an automatic system to detect and recognize text in images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 11, pp. 1224–1229, 1999.

[17] Z. Li, G. Liu, X. Qian, D. Guo, and H. Jiang, "Effective and efficient video text extraction using key text points," *IET Image Processing*, vol. 5, no. 8, pp. 671–683, 2011.

[18] X. Qian and G. Liu, "Text detection, localization and segmentation in compressed videos," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2006, pp. 385–388.

[19] W. Fan, J. Sun, Y. Katsuyama, Y. Hotta, and S. Naoi, "Text detection in images based on grayscale decomposition and stroke extraction," in *Proc. of the Chinese Conference on Pattern Recognition*, 2009.

[20] A. Srivastav and J. Kumar, "Text detection in scene images using stroke width and nearest-neighbor constraints," in *Proc. of the IEEE Region 10 Conference (TENCON)*, 2008.

[21] M. Kumar, Y. C. Kim, and G.-S. Lee, "Text detection using multilayer separation in real scene images," in *Proc. of the 10th International Conference on Computer and Information Technology*, 2010, pp. 1413–1417.

[22] J. Yi, Y. Peng, and J. Xiao, "Color-based clustering for text detection and extraction in image," in *Proc. of the 15th International Conference on Multimedia*, 2007, pp. 847–850.

[23] D. Lopresti and J. Zhoum, "Extracting text from www images," in *Proc. of the 4th International Conference of Document Analysis and Recognition*, 1997, pp. 248–252.

[24] C. Thillou and B. Gosselin, "Combination of binarization and character segmentation using colour information," in *Proc. of the 4th IEEE International Symposium on Signal Processing and Information Technology*, 2004, pp. 107–110.

[25] Y. Liu, S. Goto, and T. Ikenaga, "A robust algorithm for text detection in color images," in *Proc. of 8th International Conference on Document Analysis and Recognition*, 2005, pp. 339–403.

[26] R. Wang, W. Jin, and L. Wu, "A novel video caption detection approach using multi-frame integration," in *Proc. of International Conference on Pattern Recognition*, 2004, pp. 449–452.

[27] K. I. Kim, K. Jung, and J.-H. Kim, "Texture-based approach for text detection in images using support vector machines and continuously

adaptive mean shift algorithm," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 12, pp. 1631–1639, 2003.

[28] X. Liu, H. Fu, and Y. Jia, "Gaussian mixture modeling and learning of neighboring characters for multilingual text extraction in images," *Pattern Recognition*, vol. 41, no. 2, pp. 484–493, 2008.

[29] J. Ye, L.-L. Huang, and X. Hao, "Neural network based text detection in videos using local binary patterns," in *Proc. of the Chinese Conference on Pattern Recognition*, 2009.

[30] J. Yu and Y. Wang, "Apply som to video artificial text area detection," in *Proc. of the 4th International Conference on Internet Computing for Science and Engineering (ICICSE)*, 2009, pp. 137–141.

[31] C. Shin, K. Kim, M. Park, and H. J. Kim, "Support vector machine-based text detection in digital video," in *Proc. of IEEE Signal Processing Society Workshop Neural Networks for Signal Processing*, vol. 2, 2000, pp. 634–641.

[32] M. Anthimopoulos, B. Gatos, and I. Pratikakis, "A hybrid system for text detection in video frames," in *Proc. of the 8th IAPR International Workshop on Document Analysis Systems*, 2008, pp. 286–292.

[33] G. Miao, Q. Huang, S. Jiang, and W. Gao, "Coarse-to-fine video text detection," in *Proc. of IEEE International Conference on Multimedia and Expo*, 2008, pp. 569–572.

[34] T. A. Pham, "Optimization of texture feature extraction algorithm," Ph.D. dissertation, MSc Thesis, Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, 2010.

[35] W. Niblack, *An introduction to digital image processing*. Strandberg Publishing Company, 1985.

[36] J. Sauvola and M. Pietikäinen, "Adaptive document image binarization," *Pattern recognition*, vol. 33, no. 2, pp. 225–236, 2000.

[37] C. Y. Graham Leedham, K. Takru, J. H. N. Tan, and L. Mian, "Comparison of some thresholding algorithms for text/background segmentation in difficult document images," in *Proc. of the 7th International conference on document analysis and recognition*, vol. 2, 2003, pp. 859–864.

[38] M. Feng and Y.-P. Tan, "Contrast adaptive binarization of low quality document images." *IEICE Electronic Express*, vol. 1, no. 16, pp. 501–506, 2004.

[39] B. V. Dhandra, P. Nagabhushan, M. Hangarge, R. Hegadi, and V. S. Malemath, "Script identification based on morphological reconstruction in document images," in *Proc. of 18th International Conference on Pattern Recognition*, vol. 2, 2006, pp. 950–953.

[40] S. Chanda, U. Pal, K. Franke, and F. Kimura, "Script identification: A han and roman script perspective," in *Proc. of the 20th International Conference on Pattern Recognition*, 2010, pp. 2708–2711.

[41] D. Ghosh, T. Dube, and A. P. Shivaprasad, "Script recognition:a review," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 12, pp. 2142–2161, 2010.

[42] N. Sharma, S. Chanda, U. Pal, and M. Blumenstein, "Word-wise script identification from video frames," in *Proc. of the 12th International Conference on Document Analysis and Recognition*, 2013, pp. 867–871.

[43] D. Zhao, P. Shivakumara, S. Lu, and C. Tan, "New spatial-gradient-features for video script identification," in *Proc. of the 10th IAPR International Workshop on Document Analysis Systems*, 2012, pp. 38–42.

[44] T. Ojala, M. Pietikainen, and D. Harwood, "Performance evaluation of texture measures with classification based on kullback discrimination of distributions," in *Proc. of the 12th IAPR International Conference on Pattern Recognition, Conference A: Computer Vision and Image Processing*, 1994, pp. 582–585.

[45] ——, "A comparative study of texture measures with classification based on featured distributions," *Pattern recognition*, vol. 29, no. 1, pp. 51–59, 1996.

[46] D. Huang, C. Shan, M. Ardabilian, Y. Wang, and L. Chen, "Local binary patterns and its application to facial image analysis: a survey," *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. 41, no. 6, pp. 765–781, 2011.

[47] D. Bertolini, L. S. Oliveira, E. Justino, and R. Sabourin, "Texture-based descriptors for writer identification and verification," *Expert Systems with Applications*, vol. 40, no. 6, pp. 2069–2080, 2013.

[48] X. Wang, T. X. Han, and S. Yan, "An hog-lbp human detector with partial occlusion handling," in *Proc. of the IEEE 12th International Conference on Computer Vision*, 2009, pp. 32–39.

[49] I. Siddiqi, C. Djeddi, A. Raza, and L. Souici-meslati, "Automatic analysis of handwriting for gender classification," *Pattern Analysis and Applications*, 2014.

[50] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971–987, 2002.

[51] I. Siddiqi and A. Raza, "A database of artificial urdu text in video images with semi-automatic text line labeling scheme," in *Proc. of the 4th International Conference on Advances in Multimedia*, 2012, pp. 75–81.

[52] S. M. Lucas, A. Panaretos, L. Sosa, A. Tang, S. Wong, R. Young, K. Ashida, H. Nagai, M. Okamoto, H. Yamamoto, H. Miyao, J. Zhu, W. Ou, C. Wolf, J.-M. Jolion, L. Todoran, M. Worring, and X. Lin, "Icdar 2003 robust reading competitions: entries, results, and future directions," *International Journal of Document Analysis and Recognition*, vol. 7, pp. 105–122, 2005.

[53] C. Wolf and J.-M. Jolion, "Object count/area graphs for the evaluation of object detection and segmentation algorithms," *International Journal on Document Analysis and Recognition*, vol. 8, no. 4, pp. 280–296, 2006.

[54] L. Li and C. L. Tan, "Script identification of camera-based images," in *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*. IEEE, 2008, pp. 1–4.

[55] T. Q. Phan, P. Shivakumara, Z. Ding, S. Lu, and C. L. Tan, "Video script identification based on text lines," in *Document Analysis and Recognition (ICDAR), 2011 International Conference on*. IEEE, 2011, pp. 1240–1244.

Exploring the Potential of Mobile Crowdsourcing in the Sharing of Information on Items Prices

Hazleen Aris and Marina Md Din
College of Computer Science and Information Technology
Universiti Tenaga Nasional,
Kajang, Malaysia
Email: hazleen@uniten.edu.my

***Abstract*—This article presents the result of a survey performed to identify the potential of using mobile crowdsourcing as means to exchange information on the prices of household items at local stores from the consumers point of view. The potential was identified from four perspectives; mobile devices capability, internet usage pattern, supporting infrastructure and readiness towards information sharing. Survey questionnaires comprising 18 quantitative questions were distributed to 138 respondents in the forms of hardcopy and online softcopy over a one month period in May 2014. Collected data were analysed using descriptive statistics and correlation analysis methods. Findings from the analyses showed that the potential of using mobile crowdsourcing in sharing information of item prices is high as seen from the perspectives of the mobile devices capability and supporting infrastructure. Internet usage pattern of the consumers as well as their attitude towards information sharing are also in support of the potential. To the best of our knowledge, this is the first study that gathered statistical data on the potential of using mobile crowdsourcing for the sharing of information on items prices. Potential is usually assumed based on informal observation on the prevalent of mobile devices and their widespread use, and are not supported by empirical data. It is of value to the broader research communities who are currently engaged in mobile crowdsourcing research for consumers benefits.**

***Index Terms*—Mobile crowdsourcing; Price comparison; Crowdsourcing potential; Crowdsourcing survey**

I. Introduction

Shopping, in particular groceries and household items shopping, is already part and parcel of modern living. Supermarkets and grocery stores are now the main sources where people obtain their necessities from. In Malaysia for example, it was reported that the average monthly expenditure per household in 2010 was RM2190 per month, which is equivalent to RM26,280 a year. In the UK, households spent an average of £489 a week in 2012, amounting to £23,472 a year. Thus, it can be seen that household expenditures contributed significantly to the overall expenses. Any increment or reduction in household expenses will therefore significantly affect the costs of living. In relation to this, the ever increasing price of goods has become a phenomenon that is near to the heart of many people. This phenomenon is unstoppable and nothing much can be done to curb the price from keep on hiking. It is a global issue and the increase everywhere. In Malaysia, an increase of 12.1% in five years time between 2004/2005 and 2009/2010 was recorded from the national census done every five year by the Department of Statistics Malaysia [1]. A report on consumer expenditure survey by US Bureau of Labor Statistics [31] also showed an increase of 3.5% in consumer spending in 2012, and this was the second consecutive year that expenditures increased. In Singapore, the consumer price index (CPI) for general household rose by 4.6% for the full year of 2012 and by 2.4% for the full year of 2013 [3]. These are just some examples that we had gathered and it is believed that the situations in other countries are about the same.

Given the facts and figures above, it seems like the increase is unpreventable and we just have to 'live' with this increase in the cost of living. We should therefore live with it 'wisely' by taking appropriate measures to alleviate its effects. In [8], the following three approaches that can help in moderating the impact of price hike were discussed.

- By raising the standard of living of the citizens.
- By reducing the prices of household items, followed by close monitoring and surveillance.
- By providing the means to facilitate consumers to perform selective purchase of household items.

In the article, it was concluded and justified that the third approach was seen as the most practical in the sense that it requires minimal intervention from the authority and hence, can be almost immediately implemented if the technology is available. It is also in line with the urge for the consumers to play bigger role in determining the prices of items [36]. Through this approach, the power of consumers is leveraged by encouraging them to be more selective in purchasing the household items. Being selective means, rather than just buying an item the moment they see it, consumers should spend some extra effort to compare the prices of that item at the nearby stores to see which store can give them the lowest price. This extra effort, however, has to be negligible in order to make this approach favourable. Our quick survey into the market proved the feasibility of this approach. Table I shows the prices of a selection of common household items at three different stores. As can be seen from the table, the difference in items prices can be significant when the prices are totalled up.

Traditional way of comparing prices of items from one store to another requires the customer to visit each store, which is time consuming and troublesome. Thus, the main challenge is to reduce significantly the time and effort needed to perform

TABLE I: Prices of selected items at three different stores

Item	Price (RM)		
	Store A	Store B	Store C
Rambutan Thai White Rice (10kg)	30.99	26.99	27.99
Jacobs Cream Cracker (750g)	10.49	10.99	13.99
Ayam A1 Fried Chicken Crispy (850g)	7.99	10.35	10.50
Naturel Olive Oil Extra Virgin (2kg)	18.99	14.99	19.50
Drypers Soft New Born (64 pcs)	31.99	31.90	31.59
Milo Protomalt Soft Pack (1kg)	16.49	16.75	17.99
Nutriplus Fresh Egg Organic Selenium M Size (10 pcs)	5.49	5.29	4.99
Total	122.43	117.26	126.55

the comparison. Consumers should be able to perform the prices comparison without leaving the comfort of their house. Mobile crowdsourcing was identified as a viable solution to address the challenge [9], thanks to the widespread availability and accessibility of mobile devices nowadays. Through mobile crowdsourcing, consumers from different locations can contribute information on prices of items at the local stores that they visited, to a mobile application that will compile and process the information for use by other consumers. Existing literature shows that theoretically, mobile crowdsourcing has a good potential to be used in information gathering of this kind. However, its potential from the practical point of view needs to be explored too. This research therefore aims at exploring the potential of using mobile crowdsourcing in the sharing of information on prices of items from the practical point of view, that is, from the perspective of the consumers. Section II provides the theoretical background of the research that leads to the justification on the need for this research. Section III explains about the method applied, which was the survey. Section IV presents the results and analyses performed on the results. Section V discusses the findings followed by section VI on threats to validity. Section VII concludes the article.

II. Background

Crowdsourcing is one of the many new concepts that have emerged resulting from the advancement of the information and communication technology (ICT). Crowdsourcing was first introduced in 2006 to mean the act of taking a job traditionally performed by a designated agent, usually an employee, and outsourcing it to an undefined, generally large group of people, i.e. the crowd, in the form of open call [25]. The open calls are usually made through the internet. In crowdsourcing, at least three components exist; organiser, task and a group of solvers. Organiser is the person or company who initiates an assignment or a job. A task is the assignment being offered to be solved and a group of solvers is the participating crowd who is going to solve the given task. The following are the characteristics of crowdsourcing [17].

- Clearly defined crowd
- Task with clear goal
- Clear recompense received (by the crowd)
- Clearly defined crowdsourcer (organiser)
- Clear compensation received (by the crowdsourcer)
- Online assignment and participation of tasks
- Uses open call
- Uses internet

The term mobile crowdsourcing is used in [33] to refer to a particular form of crowdsourcing where the task is made available through mobile devices and the solutions are also submitted through mobile devices. This adds the ninth characteristic of mobile crowdsourcing, which is, uses mobile devices [9].

A. Benefits of crowdsourcing

The philosophy behind crowdsourcing is the involvement of the community to assist in solving clearly defined problems. Usually, these are the kinds of task that could not have been accomplished by the organiser, most probably due to the expensive cost, time constraint, too routine or requirement of rare knowledge and skills [38]. Crowdsourcing has been used in many domains in order to harness the power and wisdom of the crowd, such as in business and marketing [13][38], sociology [39][24][30], medicine and health [18][39][10][23], environmental sciences [22][21][19], as well as research and development [37][4]. These existing crowdsourcing applications can be generally viewed as commercial crowdsourcing and non-profit crowdsourcing [29].

In commercial crowdsourcing, getting a task done at a cheaper price is amongst the reasons for crowdsourcing. With the increasing availability and affordability of mobile devices such as smart phones, mobile crowdsourcing has gained increasing popularity amongst production-based companies and many have even benefitted from it. Through crowdsourcing, a problem can be explored and solved quickly through decomposition of tasks. The task to be done is decomposed and outsourced by the organiser to the crowd, which the latter submit the best solutions back to the organiser [26]. Crowdsourcing is cheaper because payment is made only for the chosen solution and may even be substituted with other kind of compensation like a small token or prize [16]. Crowdsourcing enables the products to get into the market faster too [38]. Through crowdsourcing, an organisation can tap into a wider range of talents that might be present within its own boundary. Furthermore, selected crowd may have a degree of expertise that is not available within the organisation, who can work to solve more complex issues or tasks. By interacting with the crowd, organisations can gain insight into customers or potential customers preferences that encourage tapping into the intelligence of the crowd in order to address unique, large and critical problems [30]. In addition to these, working with an external group of people can be a source of personal satisfaction too.

Non-profit crowdsourcing is performed by unpaid volunteers for public good [6]. In the non-profit sector, benefits of crowdsourcing are mainly seen in healthcare [27][28] and disaster management [14][20]. In Haiti, in year 2010, the informal sources such as news reports, discussion rooms and Twitter alerted people about cholera outbreak two weeks before the health ministry issued its report. This had managed to give early warning to the public and they could take prevention.

A similar project, HealthMap Application [11], analyses data that come from the crowd and from the authority such as the ministry of health. This really benefits the community because they are able to detect outbreaks and provide disease surveillance in real-time and hence, precaution and prevention can be prepared earlier. Crowdsourcing can also save lives. Often in our daily lives we heard news of crimes. Application that has ability to report and alert community of any crimes happening in the neighbourhood for example, could help individuals to avoid a particular area for safety reason. Furthermore, rescue mission can reach to the location in a shorter period of time if the function of Help button integrated with the application is activated when someone is under attack [2].

B. Benefits of information sharing

Information sharing referred to the exchange of data between various organisations, people and technologies. Information sharing is often interchangeably used with knowledge sharing. As far as our review is concerned, there was no clear distinction between these two and this research does not intend to distinguish between the two. The term information sharing will be used throughout this article to refer to the exchange of data between two entities. People have been sharing information since time immemorial and it is done almost naturally using whatever means that the current technologies permit. The advancement of internet and mobile devices has brought information sharing to an unprecedented pace. People share just about anything through social media now including prices of items, if they find them extremely cheap or expensive.

During the primitive era, natural resources such as smoke and sound were used to share information with others. The advent of ICT has changed the way people share information in many ways. It is the spirit of sharing the information with others, be it a good news or otherwise, that has driven people to take advantage of just about all technologies available as long as the information is conveyed. When paper was invented, newspapers were used to disseminate news and other information to the mass. When radio and television were invented, we had the news being broadcast on air. With the diffusion of computers and internet into our daily lives, we have witnessed the birth of so many information portals, which increases the accessibility of information. The latest development in ICT, with the arrival of smartphones and other mobile devices, adds mobility and ubiquity to the existing accessibility where information sharing can be done almost instantly at just about anywhere on earth.

The benefits brought about by information sharing are numerous. A total of 27 benefits of information sharing was listed in [12], which are categorised into four categories; technical, organisational, intra-organisational and environmental. Due to these benefits, information sharing has noticeably been researched in various domains. Supply chain management is an example of such domains. Information sharing in supply chain management is further divided into three types; supply-chain-wide information sharing, downstream information sharing and upstream information sharing [21]. It was found that the value of information sharing is higher for the upstream firms than for downstream firms under all information sharing types. However, when looking at individual firm, the value of information sharing is higher under downstream information sharing than upstream information sharing. Nevertheless, information sharing research is not limited to manufacturing and supply only. Its application was also being discussed for elementary school [15] and sports [5]. In the field of education, information sharing through social networking sites was being discussed as a way of marketing an institution [7].

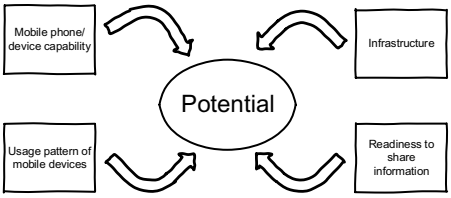


Fig. 1: Four perspectives crowdsourcing potential

Findings from the literature above theoretically confirms the benefits and potential of information sharing. It is expected that the sharing of information on items prices will also gain similar benefits and potential, and this is what we intend to discover from this study. The objective of this research is to examine the potential of using mobile crowdsourcing in the sharing of information on prices of items and the scope of the study is amongst Malaysian consumers. In order to achieve this objective, the following research questions were formulated and addressed by the research.

- What is the current state of mobile devices capabilities?
- What is the pattern of mobile devices amongst the users?
- What is the current state of the existing mobile infrastructure provided?
- What is the users tendency towards sharing information of items prices?

In order to answer the above questions, a survey was performed to identify the potential use of mobile crowdsourcing in the sharing of information on prices of items. The potential was determined from the following four perspectives; capability of mobile devices, pattern of mobile devices usage, supporting infrastructure and readiness to share information with others as shown in Figure 1.

III. Method

Personal opinion survey [36] was used to obtain the required information from the consumers. Questions in the questionnaire were close-ended and analyses performed were quantitative, as described in more detail below.

A. Questionnaire Design

After undergoing a number of revisions resulting from a series of discussions with the research team members, the initial version of the questionnaire comprises a total of 19 questions. All questions were close-ended and the choices of answer were either nominal or ordinal. Prior to its actual

distribution, a pilot test was performed on the initial version of the questionnaire involving ten participants. The purpose of the pilot test was to identify problems with regard to the clarity of the questions asked. During the pilot test, each participant answered the questionnaire in the presence of a research team member. Any issues raised during the questionnaire answering session were recorded by the researcher to be discussed later. As a result of the pilot test, a number of potential problems with regard to the clarity of the questions that might affect the understanding of the participants were identified and rectified. The following is the list of improvements made to the questionnaire.

- Highlight the word main in bold font in the question that asked about the main breadwinner because one participant missed it
- Include elaboration, in brackets, of two terms used, household items and network problem because some participants were unsure about their meanings or scopes. What was meant by household item and what kind of network problem that was referred to.
- Ambiguity of the Likert scales used in some of the questions, which was resolved by using the proposed scales in [34].
- Removal of the last question, question 19, as it was deemed similar to another question, question 18, by some respondents. Although after relooking into the questions, we still thought that each question would provide different information, we removed the question in respect of the feedback provided by the participants. This was the purpose of the pilot test anyway.

Eventually, the final version of the questionnaire contained 18 questions and was divided into two sections, Section A and Section B. Section A comprised a total of seven questions (Q1 to Q7) that asked about demographics information and information pertaining to household income and expenditure. Questionas in this section collected the following information from the participants.

- Gender (Q1)
- Marital status (Q2)
- Income earner-ship status (Q3)
- Portion (Q4)
- Record keeping of monthly expenditure (Q5)
- Monthly income (Q6) and
- Average monthly expenditure (Q7)

Section B of the questionnaire consisted of 11 questions that were arranged according to the four perspectives that determine the potential of mobile crowdsourcing in information sharing of prices of items shown in Figure 1. Details of the questions are as follows.

- Mobile device capability
  - availability (possession) (Q8)
  - capability to access the internet (Q9)
  - capability to take photo (Q10)
- Pattern of mobile device usage
  - mobile device availability during shopping (Q11)
  - mobile device accessibility during shopping (Q12)
  - hours spent accessing the internet and mobile applications (Q13)
- Supporting infrastructure
  - type of mobile plan subscription (Q14)
  - network availability with regard to internet accessibility (Q15)
- Readiness to share information with others
  - photo sharing through mobile device (Q16)
  - current practice in information sharing (Q17)
  - willingness to share information (Q18)

Two versions of the questionnaire were prepared, the hardcopy version and the online version. In the hardcopy version, all of the 18 questions from both sections were designed in such a way that they could all fit into one page, so that potential respondents would not be daunted by the number of questions that they had to answer. A one page questionnaire would be more attractive for them to participate. The online version of the questionnaire was created using Surveyshare.com, an online tool for creating and managing online questionnaires. Surveyshare.com was chosen because of our familiarity with the tool from past experience. Other tools would do just fine. Online questionnaire has advantages over the hardcopy questionnaire with features such as compulsory fields and skipping pattern that were able to avoid human errors. It is important to mention here that the coding system assigned to each question in the questionnaire (Q1, Q2 and et cetera) does not, in any way, relate to the positions of the questions in the questionnaire. The codes were assigned only after the questions were arranged according to four identified perspectives. The actual order of the questions differ in order to adhere to the qualities of good question in [32]. Furthermore, the order of questions in hardcopy questionnaire and online questionnaire also differs slightly because the hardcopy questionnaire has to take into consideration the space constraint. However, the number of questions remain the same.

The goal question metric method [35] was used to design and later analyse the results of the questionnaire. In this method, metrics is defined for each question that explains how result from each question will be analysed towards achieving the objective of the survey. It ensures the exclusion of irrelevant and redundant questions right from the beginning. Table II shows each question in section B of the questionnaire and the respective metrics used to evaluate them. Demographics questions from section A of the questionnaire were not included.

B. Procedure

The hardcopy questionnaires were distributed to the visitors who came to our booths in two research exhibitions that we participated, the UNITEN Research Exhibition 2014 (UNIREX 2014) and the International Invention and Technology Exhibition 2014 (ITEX14). The questionnaire answering sessions went smoothly with hardly any question asked, despite answering them at the booth in the presence of the

TABLE II: GQM table for the questions in the questionnaire

Goal	Purpose	Identify
	Issue	the potential of
	Object	mobile crowdsourcing in soliciting information on item prices
	Viewpoint	from the customers
Question	Q8	Do you have mobile phone/device?
Metric	M1	Percentage of respondents who have mobile phone/device
Question	Q9	Can you access internet with your mobile phone/device?
Metric	M2	Percentage of respondents who can access
Question	Q10	Can you take photo using your mobile phone/device?
Metric	M3	Percentage of respondents who can take photo
Question	Q11	Do you bring your mobile phone/device with you during your shopping trip?
Metric	M4	Percentage of respondents who bring
Question	Q12	Do you have the time to use your mobile phone/device while shopping?
Metric	M5	Percentage of respondents who have the time to access
Question	Q13	How many hours in average that you usually spend browsing the internet or accessing mobile applications in a day?
Metric	M6	Mod of total duration accessing the internet and mobile applications
Question	Q14	What is the subscription type of your mobile phone/device?
Metric	M7	Percentage of respondents who subscribe to post-paid with mobile data plan
Question	Q15	Do you face network problem (referring to internet accessibility) when using your mobile phone/device?
Metric	M8	Percentage of respondents who face it occasionally and less.
Question	Q16	Do you share photos through your mobile phone/device with others?
Metric	M9	Percentage who share.
Question	Q17	If you find out that an item is cheaper than its usual price, do you usually share about it with your friends?
Metric	M10	Percentage who do
Question	Q18	If you find out that an item is cheaper than its usual price, do you think you should share about it with your friends?
Metric	M11	Percentage who think they should

researchers. This was believed to be the result of the pilot testing performed earlier. For the online version of the questionnaire, invitations to participate were sent through emails and social media. The online version of the questionnaire was made accessible for the duration of one month. Since price hike is the issue that concerned everyone, we did not specify explicit criteria for the target population. Everyone who did grocery shopping was eligible to participate.

IV. Results and Analysis

After the two research exhibitions and the one month period were over, a total of 138 responses was received, comprising 74 (53.6%) hardcopy responses and 64 (46.4%) online responses. Prior to the analysis, three hardcopy responses had to be excluded due to incomplete information. One question each from these responses was not answered. The results presented below are therefore based on 135 (97.8%) valid responses. Due to the branching questions, the number of analysed responses per question varies as shown in Table III. Descriptive statistics method was used in analysing the results obtained as listed in Table II.

TABLE III: Number of respondents after each branching

	Hardcopy	Online	Total
Received responses	74	64	138
Valid responses	71	64	135
Proceed to Q9	71	64	135
Proceed to Q10	70	62	132
Proceed to Q16	67	62	129
Proceed to Q11 and the rest	70	62	132

TABLE IV: Profiles for participants

Characteristics	Frequency	Percentage
Gender		
Male	51	37.78
Female	84	62.22
Marital status		
Single	42	31.11
Married	93	68.89
Income earner		
Sole	26	19.26
Joint	109	80.74
Main breadwinner	48	35.56
Record expenses	74	54.81

A. Demographics information

From the total of 135 respondents, 84 (62.22%) of them were females and 51 (37.78%) were males as shown in Table IV. As can be seen from the table too, 93 (68.89%) of the respondents were married and 42 (31.11%) were not married. Demographics information then looked at the income and expenditure of the respondents. From the total of 135 respondents, only 26 (19.26%) of them were the only income earner of the family as shown in Table IV. A significant majority of them (80.74%) were not. From these, 48 (35.56%) respondents were the main breadwinners of the household. This means that an additional 23 respondents who were not the only income earner of the household, were also the main breadwinner of the household. When asked whether or not they recorded their monthly expenditures, it was discovered that more than half of them (54.81%) recorded their monthly expenditure, as also shown in Table IV.

With regard to the income range, 13 (9.63%) of the respondents lived with the monthly income of less that RM1,200 as shown in Figure 2. RM1,200 was used specifically because this was the threshold value used by the government to identify low income earners. In other words, in Malaysia, household with income of RM1,200 or less is categorised as low income household [34] and hence qualifies for a number of benefits and initiatives offered by the government. The majority of the respondents (54.07%) had household income between RM1,200 and RM8,000. There was also a notably large number of respondents (36.30%) with household income of more than RM8,000. Household income is the combined income of the whole house, usually of husband and wife,

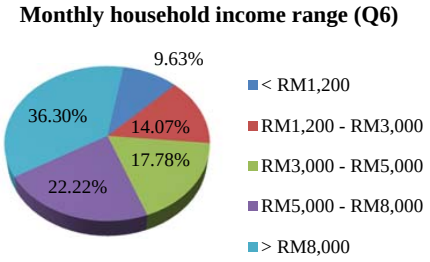


Fig. 2: Household income range of respondents

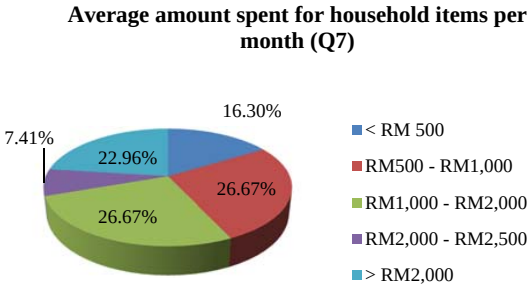


Fig. 3: Average amount spent on household items in a month

therefore, RM8,000 seems reasonable. The ceiling household income in our survey was capped at RM8,000, because this survey was part of the research that looks into ways to help the consumers cope with the increasing prices of items. Therefore, the focus was more on the low and middle income earners.

Finally, demographics information looked into the average monthly amount spent for household items. Results obtained showed that the majority (53.34%) of the respondents spent between RM500 and RM2,000 for monthly household expenditure as presented in Figure 3. Only 22 (16.30%) respondents spent less than RM500 per month on household expenditure. It is believed that this group of respondents came from those who were single and who were with lowest household income, although no effort was made to look into the results in detail with regard to this. A bubble plot of monthly household income versus average monthly household expenditure in Figure 4 shows that the household expenditure is proportional to the household income.

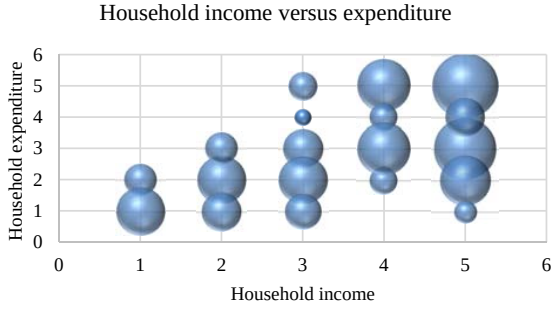


Fig. 4: Relationship between monthly income and expenses

TABLE V: Availability and capability of mobile devices

Characteristics	Frequency	Percentage
Devices availability	135	100.00
Internet accessibility	131	97.04
Photo taking feature	132	97.73

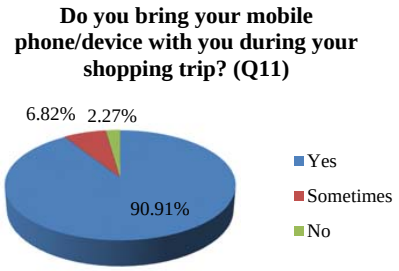


Fig. 5: Mobile phones availability during shopping trip

B. Availability and capability of mobile devices

With regard to the availability of mobile devices, it came as no surprise when the result showed that all participants (100%) possessed mobile phone or devices as shown in Table V. We kind of expected this and the result confirmed our expectation. When asked whether the mobile devices that they have can be used to access the internet or not, 131 (97.04%) of them were able to do that using their mobile devices. This represented a significant majority of the respondents. Noticeably, about the same percentage of respondents was found to be able to take photos using their mobile phone or devices as also shown in Table V.

C. Pattern of mobile devices usage

Next, the questionnaire explored about the pattern of mobile devices usage amongst the respondents, particularly during their shopping trips. It was found that 129 (97.73%) respondents have their mobile devices with them during their shopping trips as shown in Figure 5. We further asked whether or not they are able to somehow use their mobile devices during the shopping trips and the result showed that 119 (91.54%) of them were able to access their mobile devices while doing their shopping as shown in Figure 6.

The last question in this section asked about the daily average cumulative hours spent browsing the internet or ac-

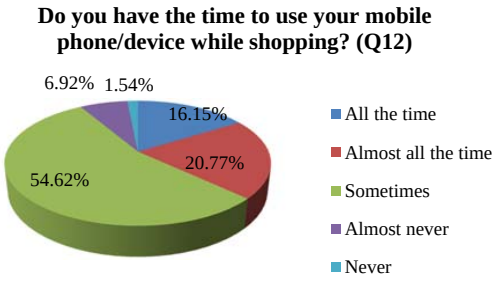


Fig. 6: Mobile phones accessibility during shopping trip

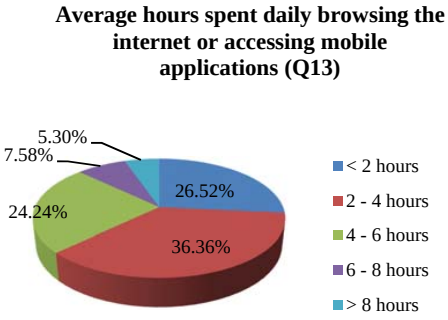


Fig. 7: Average hours spent daily using mobile phones to access the internet

TABLE VI: Types of mobile phone plan subscription

Types of subscription	Description
Prepaid	Subscribers can access the internet and mobile application as long as they have available credit value. Credit value needs to be topped up when it runs out.
Post-paid with data plan	Internet access is included in the subscribed package. Users can access the internet continuously as long as the data limit is not reached. Users will be billed at the end of each month.
Post-paid without data plan	Package only covers phone call. Internet access is not included. Users will have to subscribe to the data plan on ad hoc basis whenever needed.

cessing mobile applications. As can be seen from Figure 7, the majority of the respondents (68.18%) spent between two and eight hours accessing the internet and mobile applications daily.

D. Supporting infrastructure

With regard to the supporting or enabling infrastructure, we first asked the respondents on the types of their mobile plan subscription. Generally, there are two types of subscription available in this country; prepaid and post-paid. Post-paid can be further categorised into post-paid with data plan and post-paid without data plan. Their differences with regard to internet accessibility is shown in Table VI.

As can be seen from the table, post-paid with data plan users will experience least disruption with their internet access. For this category of subscribers, internet accessibility will become a problem only when the bill is not paid and this is less likely to happen, compared with the other categories of subscribers. If the allocated data limit is reached, they can easily top it up and the cost will appear in the next bill. For post-paid with mobile data plan subscribers, the package enables them to access the internet and mobile applications, usually capped at certain data limit depending on the chosen plan. Subscription to this type of mobile plan enables them to experience uninterrupted internet access as long as the allocated limit is not exceeded. On the other hand, prepaid subscribers will have to ensure that their credit values are sufficient before they can use their line to access the internet.

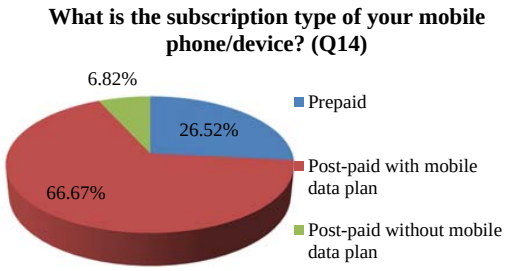


Fig. 8: Types of mobile phone plan subscription

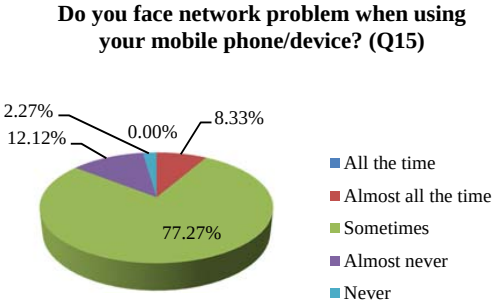


Fig. 9: Network availability

Based on the responses received, 88 (66.67%) respondents are subscribed to post-paid with mobile data plan as shown in Figure 8.

The next question probed about their experience with the (mobile) network. Result showed that only 11 (8.33%) respondents experienced network problem almost all of the time as shown in Figure 9. A vast majority of them (91.67%) had no problem with the network. Intermittent and irregular interruptions in the network are expected and acceptable. It is beyond this research to investigate their causes, let alone to figure out ways to improve the situation.

E. Readiness to share information

The questionnaire eventually explored about the consumers readiness to share information by looking at their current practices and thought. With regard to their practice in photos sharing, 125 (96.9%) respondents whose mobile devices had the capability to take photos had ever shared photos with others. Only that the frequency differs between all the time and hardly as shown in Figure 10.

When asked whether they were presently sharing information about prices of items with their friends whenever they found them cheaper than their usual prices, it was quite interesting and surprising at the same time to discover that 54 (40.91%) of them firmly did so with an additional 65 (49.24%) respondents occasionally did so as shown in Figure 11. Furthermore, regardless of whether they were presently sharing the information on prices of items with their friends or not, almost all of them (96.97%) were positive about the idea as shown in Figure 12.

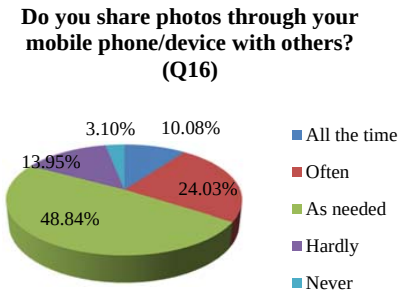


Fig. 10: Practice with regard to photo sharing

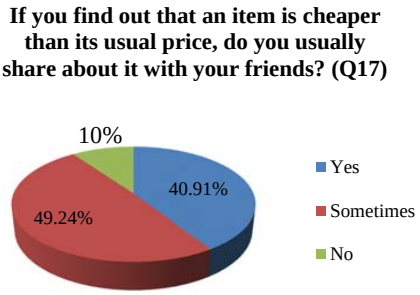


Fig. 11: Practice with regard to price information sharing

V. Discussion

The purpose of the survey was to quantitatively identify the potential of using mobile crowdsourcing to share information on prices of household items from the perspective of the consumers. Prior to analysing the collected data and coming out with the findings, we first looked at the relevance and reliability of the data. From the demographics information collected, it can be seen that the majority of the respondents are not the sole income earner of their family. Only 26 (19.26%) of them are. Further analysis was also made to analyse the correlation between marital status and sole income earner status. It was found that from 93 respondents who are married, only 14 (15.05%) of them are the sole income earner of their family. This means that for the majority of the respondents who are married (68.89%), both husbands and wives are contributing to the household income. Further analysis also showed that only 37 of the respondents who are married, are also the main breadwinner of the household.

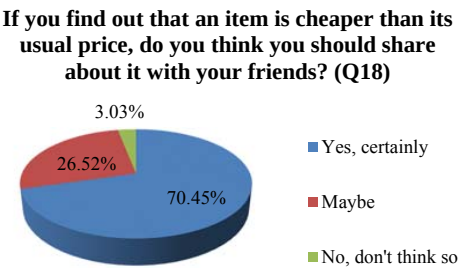


Fig. 12: Opinion on price information sharing

This can be inferred from the fact that the majority of the respondents are female (62.22%) and the majority of the married respondents, 60 out of 93 (64.52%), are also female. By convention, husbands are the main breadwinners of the households, although not necessarily. Although the majority of the respondents are female and are not the sole income earner or main breadwinner of their households, the data also show that both equally contribute to the household income and hence, manage the expenditures. Therefore, they have the required knowledge to answer the questions and subsequently provide reliable data. Furthermore, the majority of the respondents (54.8%) do record their monthly expenditures, which further strengthen the reliability of the collected data. Additionally, from the bubble plot in Figure 4, a monotonic correlation pattern can be seen between household income amount and household expenditure amount, with a few exclusions of those who gain most but spend least. The monotonic relationship is further supported by means of Spearman rank order correlation test that yields a p value of 0.9993, which indicates strong positive correlation between the two. This implies their normal and wise spending behaviour.

A. Ubiquity of mobile devices

In order to determine the potential use of mobile crowdsourcing in information sharing of prices of items from the perspective of mobile device capability, we first looked at the availability of the mobile devices. Though it may be obvious to some that mobile device is something that everybody has, it is better to have empirical data that can support the claim. As can be seen from the result shown in Table V, all of the survey respondents have mobile phones or devices and almost all of the respondents who have mobile phones or devices are able to access the internet using their mobile phones or devices. Furthermore, almost all of those who can access the internet using their mobile phones or devices are able to take photos with their mobile phones or devices. In order for the mobile crowdsourcing to be successfully used in sharing of information on prices of items, it is important for the potential users to have mobile phones or devices which features enable them to access the internet and take photos of the shared items and with respect to this perspective, it can be concluded that the potential use of mobile crowdsourcing in information sharing of items prices is high.

B. Internet as part and parcel of lifestyle

The second perspective investigates the pattern of mobile phones or devices usage. Having the required mobile phones or devices alone is not sufficient if their usage pattern is not in support of the requirements of the mobile application that will be developed for the purpose of sharing of information on prices of items. Results show that a large majority of the respondents do bring their mobile phones or devices with them during their shopping trips and almost all of the respondents are able to use the phone during their shopping trips although with varying chances. Amongst the aims of the use of mobile crowdsourcing in information sharing of prices is to enable

timely sharing of information [9]. For this, immediate update of information is expected, which means that it will be best if they can update the information during the shopping trips. Furthermore, result from the survey also showed that the majority of the respondents spend more than two hours daily accessing the internet, which is regarded as sufficient for them to use the mobile applications developed for the purpose of sharing the items prices information. The fact that the majority of the respondents spend more than two hours on average accessing the internet and mobile applications also implies their familiarity with mobile applications. Hence, it is not foreseen as a problem to use yet another mobile application that can help them in reducing their monthly expenditures. Since a large majority of them do bring the mobile phones or devices with them during their shopping trips and are able to access them during the trips, and that their usage pattern is also in support of the use of mobile applications, it can be concluded that the potential use of mobile crowdsourcing in information sharing of items prices with regard to this perspective is also high. This finding confirms the informal observation and assumption on the prevalence of mobile devices that are often not backed by statistical data.

C. Continuous internet accessibility

Next, we also looked at the potential of using crowdsourcing in information sharing of items prices from the perspective of the enabling infrastructure. Results showed that with regard to the network connectivity, the majority of the respondents are experiencing smooth network connectivity with problems occurring only occasionally. From this perspective, we also probed on the types of plan that that they are currently subscribed to. Based on the result shown in Figure 8, the majority of the respondents belong to this category of subscribers, which implies that the majority of them experience least internet access disruption. Therefore, from the perspective of enabling infrastructure, we can also conclude that the potential of mobile crowdsourcing in information solicitation and sharing of items prices is also high.

D. Positive attitude toward information sharing

Finally, we identify the potential of using mobile crowdsourcing for information sharing of items prices by looking at the current practice of the respondents and their opinion about information sharing of items prices, which constitute the fourth perspective; readiness to share the information. From the survey, it was discovered that the majority of the respondents are used to sharing photos with their friends albeit with differing frequency. Only 3.10% of them who had never shared. In sharing information on prices of items, it is expected that the photos of the items need to be shared as well if the photos have never been uploaded onto the application. If they are already used to sharing photos with others, they will not be facing any problem to do the same when sharing information on items prices with others. It was also quite surprising to discover that the majority of the respondents are also presently sharing the information on items prices with their friends despite the absence of a proper mechanism or application to do that. Only 10% of the respondents did not share the information and interestingly, 70% of respondents who are not currently sharing the information agree that they should share. Therefore, it can be concluded that from the perspective of the readiness to share information on prices of items, the potential is also high. Their mindset is more than ready to share the information on items prices.

Given that findings from all perspectives are positively supporting the potential of using mobile crowdsourcing in information solicitation and sharing of items prices, it can be concluded that based on the data collected from the survey, the potential of using mobile crowdsourcing in information solicitation and sharing on prices of items is high.

VI. Threats to Validity

While every care has been taken to ensure the reliability of the information gathered, its representativeness cannot be totally guaranteed as the data are obtained from sampled population. However, the following measures have been taken to mitigate the possible threats to data validity. On construct validity, a pilot test was performed prior to the actual questionnaire distribution to ensure that the potential respondents share common understanding of the questions in the questionnaire. As a result, a number of questionnaires were rearranged and rephrased for clarity. One question was also removed. Furthermore, there was no technical jargons or terms in the questionnaire that are difficult to understand and may cause misunderstanding. On external validity, as explained earlier, the questionnaires were distributed at research exhibitions that were open to public visitors. There was also the online version of the questionnaire that was able to reach respondents from various backgrounds. On internal validity, as can be seen in this article, only basic descriptive statistics are used in analysing the results, which are derived directly from the raw data gathered. Finally, on reliability, detailed descriptions on the survey method and questionnaire structure have been included to enable replication of the study.

VII. Conclusion

This article presents the results of a survey performed that explored the potential of using mobile crowdsourcing as means to share information on prices of items. The purpose of sharing prices information is to enable consumers to perform convenient and timely comparison of household items prices. The means being able to do so at the comfort of their house but getting up-to-date information. The comparison is necessary due to the ever increasing prices of items that has directly contributed to the overall increase in the cost of living. Findings from the analyses showed that the potential of using mobile crowdsourcing to solicit and share the information is high, with all four perspectives that determined the potential returned positive findings. Future research work will focus on the development of a model to realise the information sharing of prices through mobile crowdsourcing, which is expected to be able to help people to save on their household expenditures

despite the continuous price hike. Findings from this research contributes to the broader community in mobile crowdsourcing research for consumers as they are backed by data collected from the actual consumers. With the confirmed potential, more research opportunities should be explored that will bring more benefits and convenience to the users.

ACKNOWLEDGEMENT

Information presented in this paper constitutes a part of the research titled The Construction of a Pricewatch Information Solicitation and Sharing Model for Timely Price Comparison of Household Products using Mobile Crowdsourcing (FRGS/1/2014/SS07/UNITEN/02/1) funded by the Ministry of Higher Education Malaysia.

REFERENCES

[1] Department of statistics malaysia. http://www.statistics.gov.my/portal/index.php?option=com_content&view=article&id=767&Itemid=111&lang=en#2, Accessed on 10th July 2014.

[2] Pdrm mydistress rescue doctrine application. http://mydistress.net/main/app/mydistress.

[3] Singapore consumer price index. http://www.singstat.gov.sg/news/press_releases/cpifeb2014.pdf, Accessed on 10th July 2014.

[4] Callaghan innovation, 2014. http://callaghaninnovation.govt.nz.

[5] E. Akhir, Y. Chen, G. K. Nee, S. Sugathan, and S. Obama, "Information sharing through football website - Equatorial Guinea (EG) case study", in Information Technology (ITSim), 2010 International Symposium in, volume 3, pp. 1576–1580, June 2010.

[6] S. L. Alam and J. Campbell, "A conceptual framework of influences on a non-profit GLAM crowdsourcing initiative: A socio-technical perspective", in 24th Australasian Conference on Information Systems, ACIS 2013, dec 2013.

[7] N. Almadhoun, P. Dominic, and L. F. Woon, "Perceived security, privacy, and trust concerns within Social Networking Sites: The role of Information sharing and relationships development in the Malaysian Higher Education Institutions' marketing", in Control System, Computing and Engineering (ICCSCE), 2011 IEEE International Conference on, pp. 426–431, Nov 2011.

[8] H. Aris, "Local Pricewatch Information Solicitation and Sharing Model using Mobile Crowdsourcing", in 1st International Conference on Communication and Computer Engineering, pp. 449–457, 2014.

[9] H. Aris and M. Md Din, "On Using Mobile Crowdsourcing for Timely Information Solicitation and Sharing of Prices", in Proceedings of the 9th International Conference on Software Engineering and Applications, ICSOFT-EA 2014, pp. 518–523. SCITEPRESS, 2014.

[10] D. C. Brabham, K. M. Ribisl, T. R. Kirchner, and J. M. Bernhardt, "Crowdsourcing Applications for Public Health", Am. J. Prev. Med., 46(2):179–187, 2014.

[11] J. Brownstein. Healthmap: Outbreaks near me, 2014. https://itunes.apple.com/us/app/healthmap-outbreaks-near-me/id328358693?mt=8, Accessed on 24/6/2014.

[12] K. M. Calo, K. Cenci, P. Fillottrani, and E. Estevez, "Information sharing-benefits", J. Comput. Sci. Tech., pp. 49–55, 2012.

[13] V. Chanal and M.-L. Caron-Fasan, "How to invent a new business model based on crowdsourcing : the Crowdspirit case", in Conférence de l'Association Internationale de Management Stratégique, pp. 1–27, Sophia-Antipolis, France, May 2008.

[14] E. Chu, Y.-L. Chen, J.-Y. Lin, and J. Liu, "Crowdsourcing support system for disaster surveillance and response", in Wireless Personal Multimedia Communications (WPMC), 2012 15th International Symposium on, pp. 21–25, Sept 2012.

[15] H.-M. Chuang and C.-C. Shen, "A study on the applications of information-sharing concepts to the teaching in elementary school", in Machine Learning and Cybernetics, 2008 International Conference on, volume 1, pp. 174–179, July 2008.

[16] N. Eagle. "Internationalization, Design and Global Development: Third International Conference, IDGD 2009, Held as Part of HCI International 2009, San Diego, CA, USA, July 19-24, 2009. Proceedings", chapter txteagle: Mobile Crowdsourcing, pp. 447–456. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009.

[17] E. Estellés-Arolas and F. González-Ladrón-de Guevara, "Towards an integrated crowdsourcing definition", J. Inform. Sci., 38(2):189–200, 2012.

[18] A. Foncubierta Rodríguez and H. Müller, "Ground Truth Generation in Medical Imaging: A Crowdsourcing-based Iterative Approach", in Proceedings of the ACM Multimedia 2012 Workshop on Crowdsourcing for Multimedia, CrowdMM '12, pp. 9–14, New York, NY, USA, 2012. ACM.

[19] P. Fraternali, A. Castelletti, R. Soncini-Sessa, C. V. Ruiz, and A. Rizzoli, "Putting humans in the loop: Social computing for Water Resources Management", Eviron. Modell. Softw., 37:68–77, 2012. Environmental Modelling Software.

[20] F. Fuchs-Kittowski and D. Faust. Architecture of mobile crowdsourcing systems. in N. Baloian, F. Burstein, H. Ogata, F. Santoro, and G. Zurita, editors, *Collaboration and Technology*, volume 8658 of *Lecture Notes in Computer Science*, pp. 121–136. Springer International Publishing, 2014.

[21] H. Gao, G. Barbier, and R. Goolsby, "Harnessing the Crowdsourcing Power of Social Media for Disaster Relief", IEEE Intell. Syst., 26(3):10–14, 2011.

[22] M. F. Goodchild and J. A. Glennon, "Crowdsourcing geographic information for disaster response: a research frontier", Int. J. Digit. Earth, 3(3):231–241, 2010.

[23] Healthmap. Contagious disease surveillance and virus awareness, 2014. http://healthmap.org.

[24] J. Heinzelman, R. Brown, and P. Meier, "Mobile Technology, Crowdsourcing and Peace Mapping: New Theory and Applications for Conflict Management", pp. 39–53, 2011.

[25] J. Howe, "The Rise of Crowdsourcing", Wired Magazine, 14(6), 2006.

[26] J. M. Leimeister. Crowdsourcing as a new way of organizing work.

[27] P. Marshall, R. Cain, and S. R. Payne, "Situated crowdsourcing: a pragmatic approach to encouraging participation in healthcare design", Pervasive Computing Technologies for Healthcare (PervasiveHealth), 2011, pp. 555–558, 2011.

[28] A. McCoy, D. Sittig, and A. Wright, "Comparison of Association Rule Mining and Crowdsourcing for Automated Generation of a Problem-Medication Knowledge Base", in Healthcare Informatics, Imaging and Systems Biology (HISB), 2012 IEEE Second International Conference on, pp. 125–125, Sept 2012.

[29] D. McKinley. Non-profit crowdsourcing, 2015. http://nonprofitcrowd.org/crowdsourcing/.

[30] J. Mtsweni and L. Burge, "The potential benefits of mobile microwork services in developing nations: Research opportunities and challenges", in IST-Africa Conference Proceedings, 2014, pp. 1–10, May 2014.

[31] U. B. of Labor Statistics'. Consumer expenditures in 2012. Technical report, US Bureau of Labor Statistics, 2014.

[32] StatPac. Qualitites of a good question. http://www.statpac.com/surveys/question-qualities.htm, accessed on 5th July 2014.

[33] H. Väätäjä, T. Vainio, and E. Sirkkunen, "Location-based Crowdsourcing of Hyperlocal News: Dimensions of Participation Preferences", in Proceedings of the 17th ACM International Conference on Supporting Group Work, GROUP '12, pp. 85–94, New York, NY, USA, 2012. ACM.

[34] W. M. Vagias. Likert-type scale response anchors. https://www.clemson.edu/centers-institutes/tourism/documents/sample-scales.pdf, Clemson International Institute for Tourism & Research Development, Department of Parks, Recreation and Tourism Management. Clemson University.

[35] R. van Solingen, V. Basili, G. Caldiera, and H. D. Rombach. "Goal Question Metric (GQM) Approach". John Wiley Sons, Inc., 2002.

[36] S. N. Wan-Ahmad-Makki-Mashor. Dbkl akan wujud aplikasi harga barangan (press cutting). http://www.hmetro.com.my/articles/DBKLakanwujudaplikasihargabarangan/Article/, Accessed on 10th July 2014.

[37] S. Wang, J. Chen, and F. Xie, "Intermediating R&D and marketing value creation by open innovation", in 2011 IEEE International Conference on Industrial Engineering and Engineering Management, pp. 1170–1174. IEEE, 2011.

[38] P. Whitla, "Crowdsourcing and Its Application in Marketing Activities", Contemporary Management Research CMR, 5(1):1628, 2009.

[39] B. Yu, M. Willis, P. Sun, and J. Wang, "Crowdsourcing Participatory Evaluation of Medical Pictograms Using Amazon Mechanical Turk", J. Med. Internet Res., 15(6), Mar 2013.

Genetic-Based Task Scheduling Algorithm in Cloud Computing Environment

Safwat A. Hamad
Department of Computer Science,
Faculty of Computers & Information, Cairo University,
Cairo, Egypt

Fatma A. Omara
Department of Computer Science,
Faculty of Computers & Information, Cairo University,
Cairo, Egypt

***Abstract*—Nowadays, Cloud computing is widely used in companies and enterprises. However, there are some challenges in using Cloud computing. The main challenge is resource management, where Cloud computing provides IT resources (e.g., CPU, Memory, Network, Storage, etc.) based on virtualization concept and pay-as-you-go principle. The management of these resources has been a topic of much research. In this paper, a task scheduling algorithm based on Genetic Algorithm (GA) has been introduced for allocating and executing an application's tasks. The aim of this proposed algorithm is to minimize the completion time and cost of tasks, and maximize resource utilization. The performance of this proposed algorithm has been evaluated using CloudSim toolkit.**

Keywords—Cloud computing; Task Scheduling; Genetic Algorithm; Optimization Algorithm

I. INTRODUCTION

Due to the development of virtualization and Internet technologies, Cloud computing has emerged as a new computing platform [1]. Cloud computing can be defined as a type of distributed system consisting of a collection of interconnected and virtualized computers that are dynamically provisioned. It provides one or more consolidated computing resources based on service-level agreements (SLA) between the service providers and service consumers [2].

Cloud computing has some challenges (e.g., security, performance, resource management, reliability, etc.) [3]. One of the resource management issues is related to task scheduling. Task scheduling on Cloud computing refers to allocating the users' tasks on the available resources to improve execution of tasks, and increase resource utilization [4].

As the allocation of Cloud resource is based on SLA, the task execution cost is considered one of the main performance parameters of the task scheduling algorithm [5]. On the other hand, the task scheduling algorithm is considered a complex process because it must schedule a large number of tasks into the available resources. In the other side, there are many parameters that should be taken into consideration to develop a task scheduling algorithm. Some of these parameters are important from the Cloud user perspective (i.e., tasks compilation time, cost, and response time). Other parameters are important from the Cloud provider perspective (i.e., resource utilization, fault tolerant, and power consumption) [6].

The task scheduling problem is considered NP-Complete problem. Therefore, optimization approaches could be used to solve it by considering performance parameters (i.e., completion time, cost, resource utilization, etc.) [6]. The aim of this paper is to develop a task scheduling algorithm in the Cloud computing environment based on Genetic Algorithm for allocating and executing independent tasks to improve task completion time, decrease the execution cost, as well as, maximize resource utilization.

The rest of the paper is as follows: Section 2 discusses the related work. In Section 3, the principles of the modified GA-based task scheduling are described. The configuration of the CloudSim simulator, implementation of the proposed Genetic Algorithm, as well as, performance evaluation are discussed Section 4. Finally, conclusion and future work are given in Section 5.

II. RELATED WORK

In recent years, the problem of task scheduling on a distributed environment has caught the attention of researchers. The main issue is the execution time which should be minimized. On the other hand, scheduling of tasks is considered a critical issue in the Cloud computing environment by considering different factors like completion time, the total cost for executing all users' tasks, utilization of the resource, power consumption, and fault tolerance.

GE Junwei [6] has presented a static genetic algorithm by considering total task completion time, average task completion time, and cost constraint.

One of the scheduling issues is to allocate the correct resource to the arriving tasks. The dynamic scheduling process is considered complex if several tasks arrive at the same time, Therefore, S. Ravichandran and D. E. Naganathan [7] have introduced a system to avoid this problem by allowing the arrived tasks to wait in a queue and the scheduling will recompute and sort these tasks. Therefore, the scheduling is done by taking the first task from the queue and allocated to the resource that will be the best fit using GA. The objective of this system is to maximize utilization of resources as also reduce execution time.

R. Kaur and S. Kinger [5] have proposed task scheduling algorithm-based enhancement GA. They use a new fitness function based on mean and grand mean values. They claim that this algorithm could be implemented on both task and resource scheduling.

A comparative study of three task scheduling algorithms on the Cloud computing environment - round-robin, pre-emptive priority and shortest remaining time first algorithms - has been done in [8].

V. V. Kumar and S. Palaniswami [9] have introduced a study focusing on increasing the efficiency of the task scheduling algorithm for real-time Cloud computing services. Additionally, they have introduced an algorithm to utilize the turnaround time by assigning high priority for the task of early completion time and less priority for abortion issues of real-time task.

Z. Zheng, *et al.* [10] have proposed an algorithm based on GA to deal with scheduling problem in the Cloud computing environment called Parallel Genetic Algorithm (PGA) to achieve the optimization or sub-optimization for Cloud scheduling problems mathematically.

Furthermore, one of the main goals of task scheduling from the perspective of a Cloud provider is to maximize the profit by utilizing resource efficiently. Therefore, K. Thyagaarajan, *et al.* [11] have introduced a model for task scheduling in the Cloud computing environment for an effective gain of profits on the Cloud computing service provider.

In [12], S. Singh has provided an elaborate idea about GA by introducing several variants for task scheduling in the Cloud computing environment. He has introduced an algorithm to solve task scheduling problem by modifying GA in which initial population is generated by Max-Min approach to get more optimum results in term of "makespan".

III. The Proposal Genetic Based Task Scheduling Algorithm

The Cloud provider should guarantee optimal scheduling of user's tasks in the Cloud computing environment according to SLA. At the same time, he should guarantee the best throughput and good utilization of the Cloud resources.

Generally, by increasing the users' tasks, the complexity of scheduling these tasks in the Cloud computing environment will be increased proportionally. Therefore, the Cloud provider needs a good algorithm to schedule the users' tasks on the Cloud to satisfy QoS, minimize makespan, and guarantee good utilization of the Cloud resources [5]. Therefore, task scheduling is classified as an optimization problem.

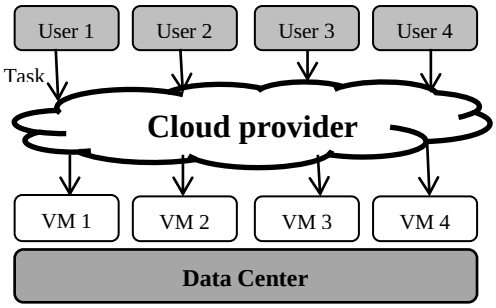


Fig. 1. Task scheduling principles

Fig. 1 illustrates the task scheduling process where each user introduces his application's tasks, and the Cloud provider uses the appropriate approaches to schedule these tasks by considering some optimization parameters, such as minimum makespan, resources utilization, and minimum cost.

Therefore, the optimization problem can be solved using heuristic algorithm such as Genetic Algorithm (GA), Particle Swarm Optimization (PSO), and Ant Colony Optimization (ACO).

In this work, the proposed task scheduling algorithm in the Cloud environment is based on the default GA with some modifications. According to these modifications, the parents will be considered in each population beside the produced child after the crossover process. Also, the Tournament Selection is used to select the best chromosomes to overcome the limitation of the population size. Therefore, the proposed algorithm is called **T**ournament **S**election **G**enetic **A**lgorithm (TS-GA).

A. Genetic Algorithm

Genetic Algorithm (GA) is based on the biological concept of generating the population. GA is considered a rapidly growing area of Artificial Iintelligence [1] [2]. By Darwin's theory of evolution was inspired the Genetic Algorithms (GAs). According to Darwin's theory, term "Survival of the fittest" is used as the method of scheduling in which the tasks are assigned to resources according to the value of fitness function for each parameter of the task scheduling process [13]. The main principles of the GA are described as follows [1] [2]:

1) Initial Population

The initial population is the set of all individuals that are used in the GA to find out the optimal solution. Every solution in the population is called as an individual. Every individual is represented as a chromosome for making it suitable for the genetic operations. From the initial population, the individuals are selected, and some operations are applied on them to form the next generation. The mating chromosomes are selected based on some specific criteria.

2) Fitness Function

The productivity of any individual depends on the fitness value. It is the measure of the superiority of an individual in the population. The fitness value shows the performance of an individual in the population. Therefore, the individuals survive or die out according to the fitness or function value. Hence, the fitness function is the motivating factor in the GA.

3) Selection

The selection mechanism is used to select an intermediate solution for the next generation based on the Darwin's law of survival. This operation is the guiding channel for the GA based on the performance. There are various selection strategies to select the best chromosomes such as roulette wheel, Boltzmann strategy, tournament selection, and selection based on rank.

4) Crossover

Crossover operation can be achieved by selecting two parent individuals and then creating a new individual tree by alternating and reforming the parts of those parents. Hybridization operation is a guiding process in the GA and it boosts the searching mechanism.

5) Mutation

After crossover, mutation takes place. It is the operator that introduces genetic diversity in the population. The mutation takes place whenever the population tends to become homogeneous due to repeated use of reproduction and crossover operators. It occurs during evolution according to a user-defined mutation probability, usually set to fairly low. Mutation alters one or more gene values in the chromosome from its initial state. This can produce the entirely new gene values being added to the gene pool. With this new gene values, the genetic algorithm may be able to produce a better solution than was previously (see Fig. 2) [1].

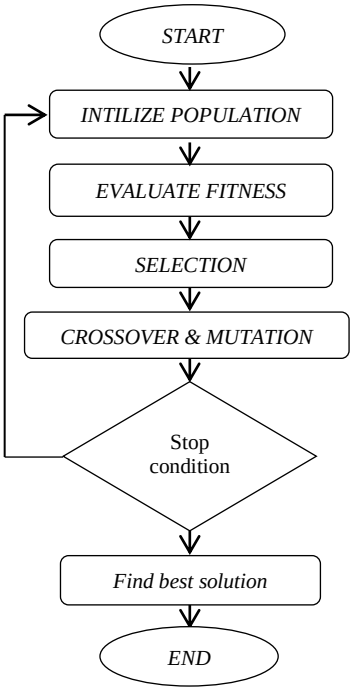


Fig. 2. Genetic Algorithm [14]

B. The Proposed Tournament Selection Genetic Algorithm (TS-GA)

In this work, a modified GA is proposed to solve task scheduling problem in Cloud computing environment to enhance the completion time for executing all tasks on the VMs, in the same time, minimize the total cost of usage the resource and maximize utilization of the resource. The main idea of this proposed algorithm (i.e., TS-GA) is that after each selection in the population, there is a solution that might satisfy good fitness function, but it is not selected to crossover process. By the proposed algorithm, this solution is not removed from the population, but it is chosen and added to the population when next iteration is started. This step is considered as a good step as some of the iterations can generate the best solution.

1) Initialize Population

According to the proposed TS-GA algorithm, the population is randomly generated using encoded binary (0, 1).

Therefore, the representation of solutions in task scheduling for each gene or (chromosome) consists of VM ID and ID for each task to be executed on these VM (see Fig. 3).

Each VM and the executed tasks on it are encoded into the binary bit (e.g., VM3: - TS4-TS8-TS9 → [0110 – 0100-1000-1001]).

2) The Fitness Function Representation

The main objective of task scheduling in the Cloud computing is to reduce completion time for execution all tasks on the available resources. Therefore, the completion time of task Ti on VMj as CTij is defined using equation "(1)" [15]:

$$\text{Completion Time} = \text{CTmax}[i,j]$$
$$i \in T, \quad i = 1, 2, 3, \dots n$$
$$j \in VM \quad , \quad j = 1, 2, 3, \dots m \qquad (1)$$

Where CTmax denotes maximum time for complete Task i on VMj. *n* and *m* denote the number of tasks and virtual machines respectively.

Therefore, to reduce the completion time which can be denoted as CTmax, the execution time of each task for each virtual machine must be calculated for the scheduling purpose. If the processing speed of virtual machine VMj is PSj, then the processing time for task Pi can be calculated by equation "(2)" [15]:

$$\text{Pij} = \frac{\text{Ci}}{\text{Psj}} \qquad (2)$$

Where, Pij the processing time for task Pi on VMj and Ci computational complexity of task Pi

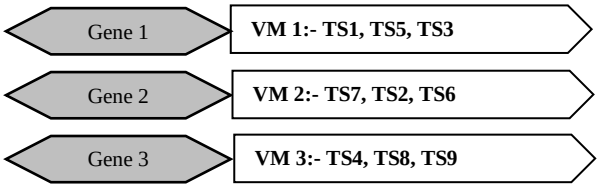


Fig. 3. Representation tasks and VMs

The processing time of each task in the virtual machine can be calculated by equation "(3)" [15]:

$$\text{Pj} = \sum_{i=1}^{n} \text{Pij} \qquad (3)$$

3) Selection Process

Tournament selection is computationally more efficient and more amenable to parallel implementation [16]. Therefore, the developed TS-GA algorithm, Tournament Selection is used to overcome the limitation of the population size. Two individuals are chosen at random from the population. A random number r is then chosen between 0 and 1. If r < k (where k is a

parameter, for example, 0.75), the fitter of the two individuals is selected to be a parent; otherwise the less fit individual is selected. The non-chosen individuals are then returned to the original population and could be selected again.

4) Crossover

In the proposed TS-GA algorithm, the new crossover has been used differently from the used crossover in the original GA. Therefore, two chromosomes which are selected to crossover process to generate two offspring will be considered as offspring also. So, the proposed crossover produces four children (see Figure 4). After that, the two best children are chosen from these.

5) Initialize Subpopulation

After each iteration, subpopulations (i.e., new populations after crossover) are added into old populations (i.e., parents). This step can enhance the diversity of population. This idea is introduced by the modified TS-GA algorithm.

6) Keep Best Solution

There is a solution that might satisfy good fitness function, but it is not selected during the crossover process. By the proposed TS-GA algorithm, this solution is not removed from the population, but it is chosen and added to the population when next iteration is started. This step is considered as good step as some of the iterations can generate the best solution.

Generally, according to the modified TS-GA algorithm, a set of modifications have been introduced. These modifications are as follows.

- The tournament is used instead of the roulette wheel in the selection process to select the best solution.
- The solutions not chosen in the selection process are considered and added to the new population. This might help in generating the best solution in the next generations.
- The new crossover is introduced by considering parents individuals as new child (see Fig. 4)

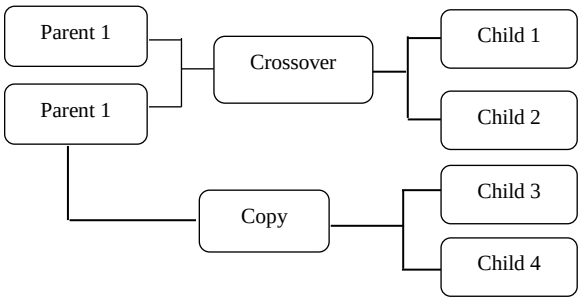


Fig. 4. Crossover process

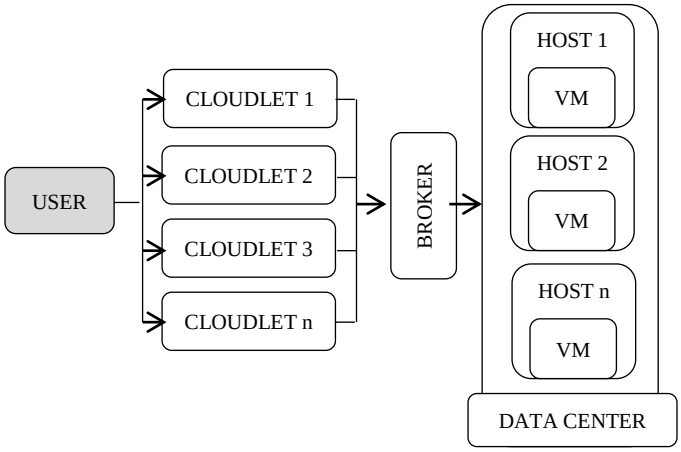


Fig. 5. CloudSim Behavior [17]

- After each iteration, subpopulations (i.e., new populations after crossover) are added into old populations (i.e., parents).

IV. P PERFORMANCE EVALUATION

In this section, the experimental evaluation of the proposed TS-GA algorithm on the original GA, and Round-Robin algorithms is presented, starting by describing the experimental environment.

A. The Experimental Environment

The CloudSim toolkit helps the researchers to simulate cloud computing environment, where it is released by the Cloud Computing and Distributed Systems Laboratory, University of Melbourne [17]. In other words, it provides the features of modeling and simulation of Cloud computing environment. According to CloudSim, the user tries to submit his requests in the form of cloudlets. Each cloudlet has the properties of file size, the number of instructions to be executed, etc. These cloudlets will be submitted to the broker to schedule onto VMs according to scheduling. CloudSim has an advantage of the building of broker driven policies. The defined class in CloudSim, VM, represents the virtual machine which can be created on the hosts. Creation of hosts depends on the broker where it allocates each VM to the different host. Datacenter has the capability to hold a maximum number of hosts and the broker can dynamically change the setup of hosts and VMs (see Fig. 5) [17].

B. Experimental Results

By using CloudSim toolkit, the proposed TS-GA is implemented, and a comparative study has been made among three algorithms; Round-Robin (RR), the default GA, and the improved TS-GA algorithms. Five parameters are considered to evaluate the performance. These parameters are the completion time, cost, resource utilization, speedup, and efficiency.

1) Completion Time:

Table 1, and Fig. (6) represent the completion time of RR, default GA and the proposed TS-GA algorithms using 8 VMs.

TABLE I. THE COMPLETION TIME OF RR, GA, AND TS-GA ALGORITHMS USING EIGHT VMS

No. of Task	RR	GA	TS-GA	Improve TS-GA vs. GA	No. VM
25	174.78	292.51	122.19	58.22 %	8
50	578.77	420.21	286.12	31.91 %	
75	582.16	680.84	405.97	40.37 %	
100	954.37	812.89	513.7	36.8 %	

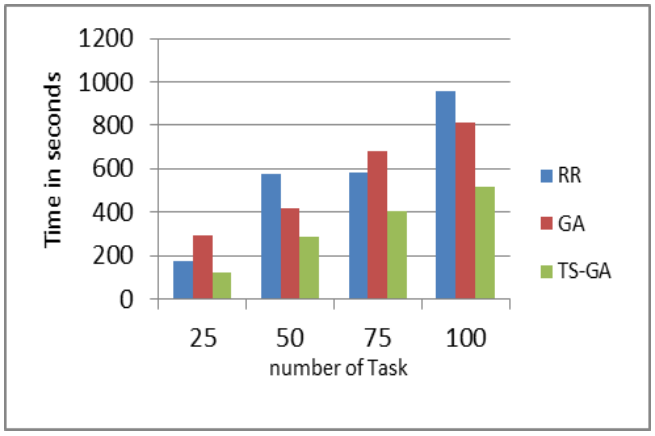


Fig. 6. the comparison completion time of three algorithms RR, GA and TS-GA

According to the results in Fig. (6), it is found that the completion time of the proposed TS-GA algorithm is reduced by (41.83%) and (39.26%) about the default GA, and RR algorithms respectively.

2) Execution Cost:

In addition, the total cost of execution of all tasks on the available VMs is calculated as "(4)" [18]:

$$Total\ Cost = \frac{Task\ length\ *\ Cost\ per\ seconds}{VM\ mips} + Processing\ Cos \quad (4)$$

Table 2. and Fig. (7) represent the execution cost of RR, default GA and the proposed TS-GA algorithms using 8 VMs.

TABLE II. THE EXECUTION COST OF RR, GA AND TS-GA ALGORITHMS

No. of Task	RR	GA	TS-GA	IMPROVE TS-GA vs. GA	No. VM
25	3017.96	3012.98	2915.21	3.24 %	8
50	7543.22	6813.46	6603.88	1.3 %	
75	10517.14	10549.22	9910.48	6.05 %	
100	13515.05	13118.87	12618.22	3.81 %	

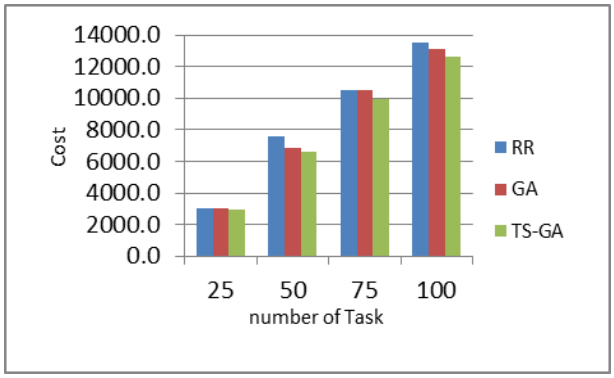


Fig. 7. The comparison cost of three algorithms RR, GA and TS-GA

According to the results in Fig. (7), the cost of the proposed TS-GA algorithm is reduced by (3.6%) and (6.07%) relative to the default GA and RR algorithms respectively.

3) Resource Utilization:

On the other side, the utilization of resources represents the ratio between the total busy time of Virtual Machine and the total finish execution time of the parallel application. It is defined as "(5)" [19]:

$$\text{utilization} = \frac{\text{final VMs available time}}{\#\text{VMs} \ * \ \text{schedule time}} * 100 \quad (5)$$

Table 3, and Fig. (8) represent the resource utilization of RR, default GA and the proposed TS-GA algorithms using 8 VMs.

TABLE III. THE RESOURCE UTILIZATION OF RR, GA AND TS-GA ALGORITHMS

No. of Task	RR	GA	TS-GA	IMPROVE TS-GA vs. GA	No. VM
25	55.13	31.51	75.89	58.47	8
50	44.36	47.88	84.43	43.29	
75	69.4	39.96	80.75	50.51	
100	52.04	48.7	75.78	35.73	

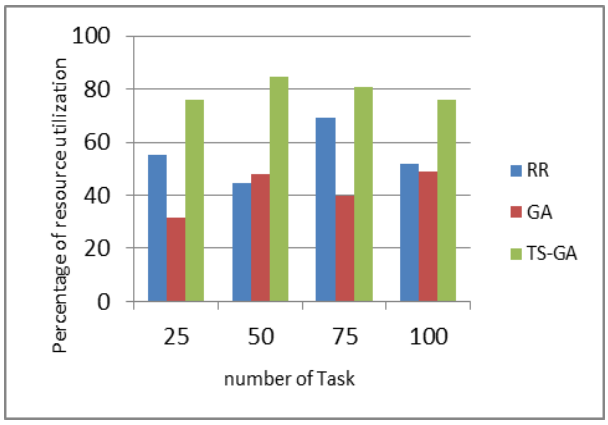


Fig. 8. Resource Utilization of RR, GA and TS-GA algorithms

According to the results in Fig. (8), it is found that the resource utilization of the proposed TS-GA algorithm is improved by (47%) and (30.04%) relative to the default GA and RR algorithms respectively.

4) Speedup and Efficiency:

The speedup and efficiency for each VM are calculated as "(6)", "(7)" follows [19]:

$$\text{Speedup} = \frac{final\ sum\ of\ excute\ time\ of\ tasks\ to\ one\ VM/\#VMs}{schedule\ time} \quad (6)$$

$$\text{Efficiency} = \frac{speedup}{\#\ VMs} \quad (7)$$

Table 4, Fig. (9) and Fig. (10) represent the speedup and efficiency of RR, default GA and the proposed TS-GA algorithms using 8 VMs.

TABLE IV. SPEEDUP AND EFFICIENCY FOR RR, GA, AND TS-GA

No. TASK	Algorithm	Speedup	Efficiency
25	RR	0.193	0.241
	GA	0.143	0.179
	TS-GA	0.248	0.311
50	RR	0.314	0.393
	GA	0.419	0.524
	TS-GA	0.558	0.697
75	RR	0.598	0.748
	GA	0.543	0.679
	TS-GA	0.829	1.037
100	RR	0.622	0.778
	GA	0.691	0.864
	TS-GA	1.105	1.319

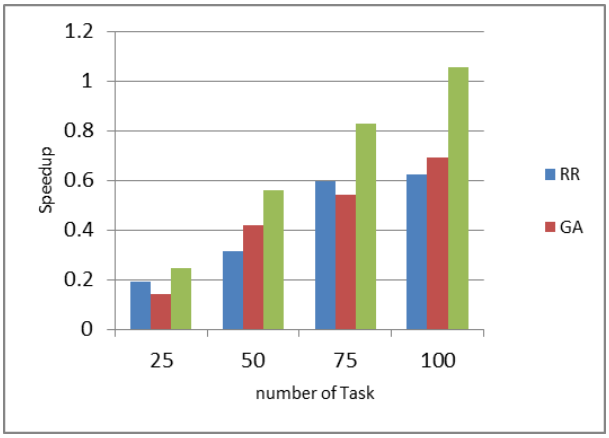


Fig. 9. The comparison speedup of three algorithms RR, GA and TS-GA

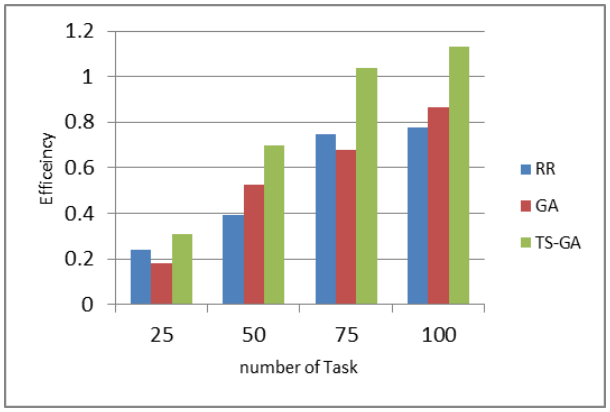


Fig. 10. The comparison efficiency of three algorithms RR, GA and TS-GA

According to the results in Fig. (9), it is found that the speedup of the proposed TS-GA algorithm is improved by (34.03%) and (33.65%) about the default GA and RR algorithms respectively. Also, the efficiency of the proposed TS-GA algorithm is improved by (34.06%) and (33.66%) about the default GA and RR algorithms respectively (see Fig. 10).

The average improved of speedup and efficiency of the proposed TS-GA algorithm about the default GA and RR algorithms are presented in Table 5 and Table 6 respectively.

TABLE V. AVERAGE IMPROVEMENT OF SPEEDUP AND EFFICIENCY FOR TS-GA ALGORITHM RELATIVE TO DEFAULT GA ALGORITHM

No. TASK	25	50	75	100	Average improvement
Speedup	42.35	24.82	34.51	34.45	34.03 %
Efficiency	42.44	24.82	34.52	34.49	34.06 %

TABLE VI. AVERAGE IMPROVEMENT OF SPEEDUP AND EFFICIENCY FOR TS-GA ALGORITHM RELATIVE TO ROUND-ROBIN ALGORITHM

No. TASK	25	50	75	100	Average improvement
Speedup	22.22	43.57	27.86	40.98	33.65 %
Efficiency	22.23	43.58	27.88	40.9	33.69 %

V. CONCLUSION AND FUTURE WORK

This paper proposes an improved Genetic Algorithm for task scheduling problem in the Cloud computing environment. The proposed algorithm targets to minimize completion time and cost, and maximize resource utilization. The completion time for the proposed TS-GA algorithm is reduced by

(41.83%) and (39.26%) about the default GA, and RR algorithms, respectively. The cost of the proposed TS-GA algorithm is reduced by (3.6%) and (6.07%) about the default GA and RR algorithms respectively. The resource utilization of the proposed TS-GA algorithm is improved by (47%) and (30.04%) about the default GA and RR algorithms, respectively. The speedup of the proposed TS-GA algorithm is improved by (34.03%) and (33.65%) about the default GA and RR algorithms, respectively. The efficiency of the proposed TS-GA algorithm is improved by (34.06%) and (33.66%) about the default GA and RR algorithms respectively.

For future work, the proposed algorithm can be extended to add possibility dynamic characteristic of VMs through run GA. Moreover, more parameters can be added based on the users' requirements.

REFERENCES

[1] S. H. Jang, T. Y. Kim, J. K. Kim, and J. S. Lee, "The study of genetic algorithm-based task scheduling for cloud computing," International Journal of Control and Automation, vol. 5, pp. 157-162, 2012.

[2] T. Goyal and A. Agrawal, "Host Scheduling Algorithm Using Genetic Algorithm In Cloud Computing Environment," International Journal of Research in Engineering & Technology (IJRET) Vol, vol. 1, 2013.

[3] B. Furht, "Armando Escalante Handbook of Cloud Computing," ISBN 978-1-4419-6523-3, Springer2010.

[4] F. Etro, "Introducing Cloud Computing," in London Conference on Cloud Computing For the Public Sector, 2010, pp. 01-20.

[5] R. Kaur and S. Kinger, "Enhanced Genetic Algorithm based Task Scheduling in Cloud Computing," International Journal of Computer Applications, vol. 101, 2014.

[6] J. W. Ge and Y. S. Yuan, "Research of cloud computing task scheduling algorithm based on improved genetic algorithm," in Applied Mechanics and Materials, 2013, pp. 2426-2429.

[7] S. Ravichandran and D. E. Naganathan, "Dynamic Scheduling of Data Using Genetic Algorithm in Cloud Computing," International Journal of Computing Algorithm, vol. 2, pp. 127-133, 2013.

[8] V. Vignesh, K. Sendhil Kumar, and N. Jaisankar, "Resource management and scheduling in cloud environment," International Journal of Scientific and Research Publications, vol. 3, p. 1, 2013.

[9] V. V. Kumar and S. Palaniswami, "A Dynamic Resource Allocation Method for Parallel DataProcessing in Cloud Computing," Journal of computer science, vol. 8, p. 780, 2012.

[10] Z. Zheng, R. Wang, H. Zhong, and X. Zhang, "An approach for cloud resource scheduling based on Parallel Genetic Algorithm," in Computer Research and Development (ICCRD), 2011 3rd International Conference on, 2011, pp. 444-447.

[11] K. Thyagarajan, S. Vasu, and S. S. Harsha, "A Model for an Optimal Approach for Job Scheduling in Cloud Computing," in International Journal of Engineering Research and Technology, 2013.

[12] S. Singh and M. Kalra, "Scheduling of Independent Tasks in Cloud Computing Using Modified Genetic Algorithm," in Computational Intelligence and Communication Networks (CICN), 2014 International Conference on, 2014, pp. 565-569.

[13] R. Buyya, R. Ranjan, and R. N. Calheiros, "Modeling and simulation of scalable Cloud computing environments and the CloudSim toolkit: Challenges and opportunities," in High Performance Computing & Simulation, 2009. HPCS'09. International Conference on, 2009, pp. 1-11.

[14] J. S. Raj and R. M. Thomas, "Genetic based scheduling in grid systems: A survey," in Computer Communication and Informatics (ICCCI), 2013 International Conference on, 2013, pp. 1-4.

[15] B. Kruekaew and W. Kimpan, "Virtual Machine Scheduling Management on Cloud Computing Using Artificial Bee Colony," in Proceedings of the International MultiConference of Engineers and Computer Scientists, 2014.

[16] M. Mitchell, An introduction to genetic algorithms: MIT press, 1998.

[17] R. N. Calheiros, R. Ranjan, C. A. De Rose, and R. Buyya, "Cloudsim: A novel framework for modeling and simulation of cloud computing infrastructures and services," arXiv preprint arXiv:0903.2525, 2009.

[18] R. Sahal and F. A. Omara, "Effective virtual machine configuration for cloud environment," in Informatics and Systems (INFOS), 2014 9th International Conference on, 2014, pp. PDC-15-PDC-20.

[19] D. M.Abdelkader, F.Omara," Dynamic task scheduling algorithm with load balancing for heterogeneous computing," system Egyptian Informatics Journal, Vol.13, PP.135–145, 2012.

The Methodology for Ontology Development in Lesson Plan Domain

Aslina Saad
Computing Department
Faculty of Art, Computing & Creative Industry
UPSI
Tanjung Malim, Malaysia

Shahnita Shaharin
Computing Department
Faculty of Art, Computing & Creative Industry
UPSI
Tanjung Malim, Malaysia

***Abstract*—Ontology has been recognized as a knowledge representation mechanism that supports a semantic web application. The semantic web application that supports lesson plan construction is crucial for teachers to deal with the massive information sources from various domains on the web. Thus, knowledge in lesson plan domain needs to be represented accordingly so that the search on the web will retrieve relevant materials only. Essentially, such retrieval needs an appropriate representation of the domain problem. The emergence of semantic web technology provides a promising solution to improve the representation, sharing, and re-use of information to support decision making. Thus, the knowledge of lesson plan domain needs to be represented ontologically to support efficient retrieval of semantic web application in the domain of lesson plan. This paper presents a new methodology for ontology development representation of lesson plan domain to support semantic web application. The methodology is focused on the important model, tools, and techniques in each phase of the development. The methodology consists of four phases, namely requirements analysis, development, implementation, evaluation and maintenance.**

Keywords—knowledge representation; methodology; ontology development; lesson plan

I. INTRODUCTION

Ontology is widely used for knowledge representation in artificial intelligence, information retrieval and semantic web [15]. Ontology provides a common understanding of specific domains and is also expressed as a formal representation of knowledge by a set of concepts within a domain and the relationship between these concepts [18]. [5] has built an ontology of lesson plan in the form of hierarchical taxonomy that shows the semantic relationships between terms in lesson plan domain. The taxonomy was then used to produce relevant search results using semantics approach in a case based reasoning (CBR) system. This is parallel to what was stated by [8], ontology can enhance CBR systems in many dimensions. Related terms are looked up from the lesson plans ontology, a structured data source. An evaluation study was designed to examine the effectiveness of the system using this representation and have shown positive results. However, the search result is limited to cases defined in the database.

Past research shows that lesson planning imposes significant burden on teachers and causes excessive workload among teachers as they need to spend a lot of time to prepare their lessons. Several efforts have been carried out to overcome this predicament, including the development of a web-based system to assist teachers in such a task.

Various platforms have been established to enable the sharing of information among teachers, such as web pages, blogs, online systems, Slideshare[1], and even social networking sites such as Facebook[2]. However, searching relevant materials or contents using such multiple platforms will result in too much information being fed to teachers. Consequently, teachers need to filter that information to select materials that really meet their teaching needs. Invariably, such effort is laborious and taxing, which further burdens their teaching workload.

Furthermore, almost all of the web-based applications use attribute-value representation and they are stored in databases, which need constant updating and verification by the system administrator. Arguably, the uses of such databases have both benefits and limitations. According to [13], the usefulness of the database management system based on the three models, namely hierarchical, network and relational models, is severely restricted by the failure to take into account the semantic of databases.

The above limitation can be overcome by applying semantic web technology. In particular, such semantic system can be used for inter organization data sharing and reuse. The purpose of this system is to let information on the Internet to have richer semantics in order to facilitate computers to determine information that is important and relevant to various users' needs, thus improving the interoperation among the entities on the Internet.

In essence, the semantic web technology is capable to connect a particular website to other websites through the use of knowledge representation. Each site in the network of internet is connected to each other by an existing relationship that has been defined in terms of knowledge representation. Thus, the search for online information can be implemented more intelligently by focusing on the relevant domain. Given this capability, knowledge representation can be applied to a dedicated semantic web application to help diverse users, especially teachers in lesson planning.

Knowledge representation is important to produce intelligent systems based on knowledge as a key element to enable the process of reasoning and decision making. According to [2], knowledge representation is one of the core

Authors would like to thank the Ministry of Education for funding us with a two year research grant. (2013-0157-109-72)

[1] http://www.slideshare.net/
[2] https://www.facebook.com/

elements in the field of artificial intelligence, which is an important aspect of problem solving. Such focus on problem solving based on knowledge representation is also stressed by [16], who state that a computer system that is capable of performing tasks that require human intelligence entails such representation.

Essentially, the purpose of understanding what knowledge is and what are the types of knowledge that exist allows us to use it in artificial systems [6]. Thus, in the context of lesson plan construction, the issue of deciding what is to be stored and how memory should be organized in order to retrieve and reuse previously prepared teaching plans effectively and efficiently needs to be addressed urgently.

In general, many different architectures have been used for knowledge representation, including ontology. According to [3], ontology is widely considered as a promising approach for capturing and representing knowledge. On one hand, [20] asserts that ontology as an explicit formal specification of a shared conceptualisation. On the other hand, according to [16], ontology is a method that defines terms which are commonly accepted for a particular domain to enable the effective sharing of information among researchers. In essence, this definition encompasses the concepts and their relationships in that domain.

Accordingly, ontology will help ensure that the terms and symbols used are defined with clear intention. Moreover, computing is the key component that enables logic and ontology-based representation to be implemented in a computer program. In this regard, [22] describe that the semantic web ontology language should include five key components, namely the concept, taxonomy, relations and functions, axioms, and instances. Particularly, the concept involves the explanation on common issues, attributes, and facets.

II. Existing Methodologies

The literature is replete with studies in which several scholars and researchers have proposed several ontology development methodologies. According to [19], here are two ways of conceiving ontology construction, the bottom up and top down approach. Such scholars include [3], [14], [12], and [10], where the third and the forth researchers introduced ontology development models called *Methontology* and *Model-based and Incremental Knowledge Engineering* (MIKE), respectively. In essence, each methodology comprises several phases, which are classified and contrasted as summarized in Table I.

Clearly, there are some methodologies that are more comprehensive compared to others, such as [12] model, that also focus on phases of evaluation, documentation and maintenance. In contrast, [3] model does not emphasize on the evaluation phase prior to the maintenance phase. In addition, the development phases of [14] model and [10] terminate at the implementation phase, without the evaluation, documentation, and maintenance phases.

TABLE I. The Comparison Among Existing Ontology Development Models

Ohgren and Sandkuhl (2005)	**Uschold and King (1995)**	**Methontology (Fernandez, 1997)**	**MIKE (Angele, 1998)**
Requirement analysis	Specification	Specification	
		Knowledge acquisition	Elicitation
Development	Conceptualization	Conceptualization	Interpretation
	Formalization	Formalization	Formalization
		Integration	Design
Implementation	Implementation	Implementation	Implementation
Evaluation Maintenance		Evaluation	
		Documentation	
		Maintenance	

Interestingly, the model proposed by [12] is more comprehensive as evidenced by its nine development phases as opposed to other development models that comprises of four or five development phases only. Essentially, [3] model is similar with the development methodology of any application domains. In their model, the phases involve requirement analysis, development, implementation, evaluation, and maintenance, which are more distinct in the computing realm. Such development phases can help developers to familiarise with the terms or nomenclatures that they are not used to in the ontology field, such as conceptualization, formalization, integration, and interpretation. The development model introduced by [3] consists of requirement analysis, development, implementation, evaluation, and maintenance phases. Fig. 1 summarizes the four development phases, together with their end results.

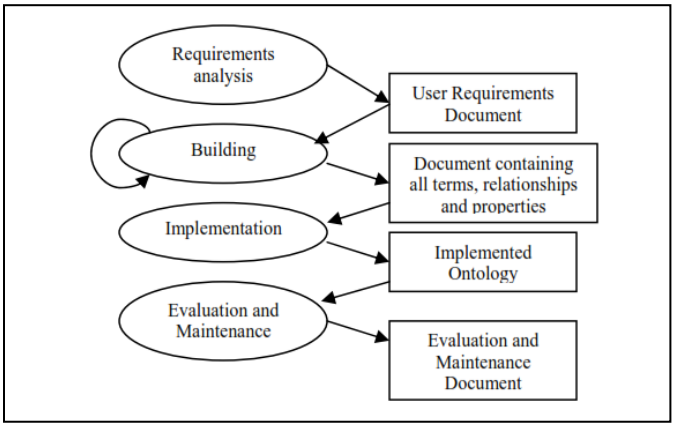


Fig. 1. Ontology development phases from [3]

Apparently, all the above methodologies emphasize requirement analysis as the first phase although different terms have been used. This phase is then followed by ontology development based on the acquired knowledge. Later, the developed ontology is then implemented and finally evaluated.

III. Applied Methodology for Lesson Plan Ontology Development

Based on the comparison of the above methodologies, a new methodology was formulated based on [3] and [12] models to serve as a guideline for knowledge representation methodology as summarized in Table II.

A. Phase I: Requirement analysis

Requirement analysis is the first phase of the methodology that comprises two sub phases, namely specification and knowledge acquisition.

TABLE II. The Comparison Among Existing Ontology Development Models

Phase		Activities
Requirement analysis	Specification	Identifying ontology specification includes: • The purpose of the developed ontology • Target users of the ontology • Ontology usage scenarios • Scope of the ontology • User requirements • Requirements of equipment and software
	Knowledge acquisition	Acquiring informal information related to knowledge and problem-solving process of subject matter experts using surveys, structured interview, observation, document analysis, and structuring techniques.
Development	Conceptualization	Developing knowledge representation in a semi-formal format using graphical representation.
	Formalization	Changing the semi-formal knowledge representation to formal knowledge representation.
	Integration	Identifying any appropriate existing ontology that can be integrated into the ontology being developed.
Implementation	Implementation	Transforming human-readable representation into machine-readable representation.
Evaluation Maintenance	Evaluation and Maintenance	Evaluating and assessing the developed ontology in meeting the requirement specifications. Identifying individuals to update and maintain the developed ontology.

1) Specification: This phase involves identifying all specifications of ontology requirements, which includes the objectives, target users, usage scenarios, scope, needs, and requirements of equipment and software in the development process ontology. For example, such needs for equipment and software include yEd[3] Graph Editor (to construct the conceptual modelling), StarUML[4] (which is to generate UML) and Protégé 5.0[5] (as an ontology language) that will be used to develop ontologies in the implementation phase. Identifying the specifications of ontology include:

- *The purpose of the ontology*: The problem of the domain, involving the construction of lesson plans for semantic application.
- *The target users of the ontology*: Teachers of any levels, including trainee teachers, inexperienced teachers, or experienced teachers.
- *The ontology usage scenarios*: Information retrieval related to lesson plan construction. The scope of the ontology for daily lesson plan.
- *The user requirements*: To support information retrieval based on keywords that are inserted by users to the semantic application.
- *The requirements of equipment and software*: The software to support ontology development conceptually and physically. For example, yEd graph editor were used for modeling, (semantic net and class diagram) and Protégé 5.0 for formalizing the developed ontology.

2) Knowledge acquisition: The first phase in the acquisition of knowledge was the elicitation process. In this phase, the procurement process related to information of the selected domain was implemented. Various techniques can be done for this procurement including the use of concept maps as a means of expression for the expert [17]. Among the activities carried out in this phase was a review of related literature, an analysis of related documents, a survey, and a structured interview with domain experts. Information obtained from these activities was recorded in the form of a natural language representation that is human-readable. The following are the techniques of knowledge acquisition implemented in this study.

a) Survey: A preliminary study was carried out using a quantitative approach with the aim to get feedback from Malaysian teachers about their daily lesson planning. An online survey questionnaire was developed using Google Form[6] to facilitate faster distribution and administration involving a wider participation compared to the conventional survey questionnaire. Essentially, the procedure of the survey involved three main steps. First, the construction of questions was performed by focussing on the components of a lesson plan, factors that influence lesson plan preparation, and materials or resources for lesson planning. Second, a survey questionnaire was distributed online. Third, the collected survey data were analyzed using appropriate descriptive statistical method.

- *Respondents*: The sample size of the online survey was made up of 150 teachers consisting of 117 female teachers and 33 male teachers. These respondents came from a diverse background, who had teaching experiences ranging from one to 30 years, and they worked in several Malaysian schools across the nation.
- *Research Instrument*: The research instrument used in the survey consists of three parts, which are demographics, lesson plan preparation, and materials parts.

b) Interview: A structured interview method involving 10 respondents was carried out with the main aim of eliciting further information by asking several pertinent questions to

[3] https://www.yworks.com/products/yed
[4] http://staruml.io/
[5] http://protege.stanford.edu/
[6] https://docs.google.com/forms/

verify some of the findings from the survey. In addition, a teach back technique was embedded in the interview sessions, in which important components or elements of a lesson plan based on [4] finding was presented to the interviewees. In particular, they were required to rank each element according to its perceived importance to the retrieval of information. The questions comprise several aspects, namely the process of producing a lesson plan, references in preparing the lesson plan, factors that are considered vital in producing the lesson plan, problems faced by teachers in preparing the lesson plan, and the necessity for tools to help them perform the task with ease.

c) Document analysis: Lesson plan documents were collected from the teachers during the interview sessions other than online resources. The documents were compared and analysed, such as to identify standard elements in a a lesson plan and the tools used for lesson planning. The analysis revealed that some of the available tools to support lesson planning were SmartLP[7], INTIME[8], KITE[9], The Lesson Planner Lesson Planning System (LPS)[10], RPH Online[11], PlanBookEdu[12], Planboard[13], Common Curriculum[14], Core Lesson Planner[15], and PlanBook[16].

B. Phase II: Development

The construction phase was a repetitive process, occurring in a cycle. In essence, this phase comprised three sub phases, namely the conceptualization, formalization, and integration sub phases. In each cycle, an evaluation was performed, and any changes were implemented to improve the constructed representation. Further explanation of each sub phase is as follow:

1) Conceptualisation phase: In this sub phase, the conceptual model for a specific domain (i.e., the lesson plan) was developed using semantic net as shown in Fig. 2. The knowledge for the domain was represented using a graph consisting of vertices to denote objects, concepts, domain entity, and edges. And each edge is basically a line connecting two vertices. As the lesson plan ontology construction was carried out within Malaysian context, the terms shown in Fig. 2 to Fig. 6 were constructed in Malay language.

2) Formalisation phase: In this phase, the conceptual model was transformed into a semi-formal representation using Unified Modeling Language (UML), specifically a class diagram. This is shown in Fig. 3. Then, a formal representation was formed to construct the ontology of the lesson plan domain. The ontology editing package Protégé 5.0 was used to develop the ontology for Daily Lesson Plan domain as shown in Fig. 4. Fig. 5 illustrates examples of the instance of the constructed class.

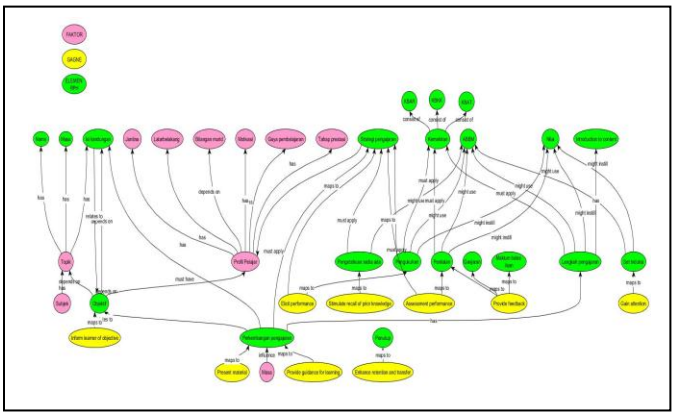


Fig. 2. Semantic net for lesson plan domain

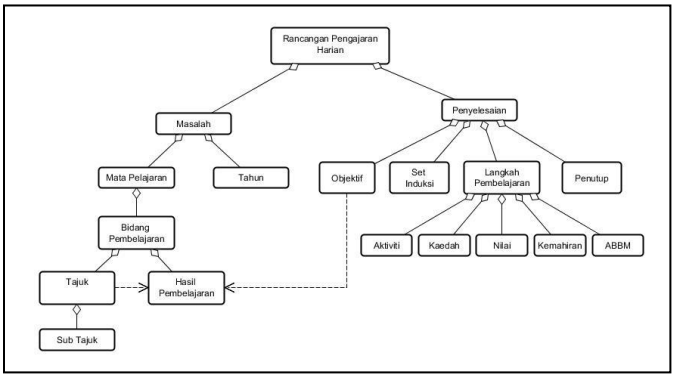


Fig. 3. Class diagram for lesson plan domain

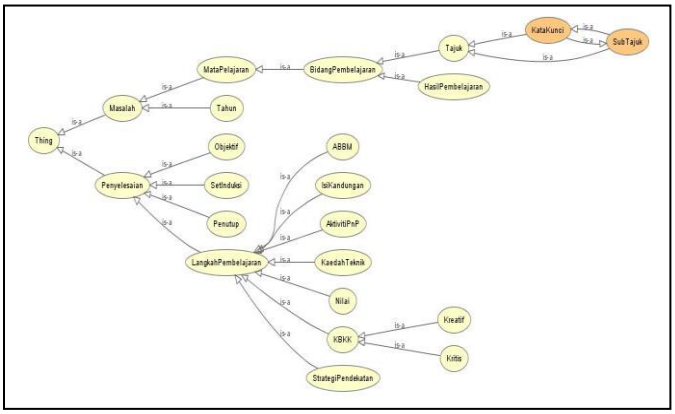


Fig. 4. Daily lesson plan ontology developed using Protégé 5.0

3) Integration phase: Any existing ontology for the lesson plan domain was identified in this phase by processing which parts of the ontology were appropriate or otherwise. If such ontology was suitable, it would be integrated into the developed ontology. A lesson plan ontology by [4] in a hierarchical form was compared to the constructed ontology. Then, these two ontologies of daily lesson plans domain were integrated during the development process.

[7] http://smartlp.upsi.edu.my/
[8] http://www.intime.uni.edu/casestudies
[9] http://kite.missouri.edu/
[10] https://www.oncoursesystems.com/products/detail/lessonplanner
[11] http://rphonline.teknologihijau.net/
[12] http://planbookedu.com/
[13] https://planboard.chalk.com/
[14] http://www.commoncurriculum.com/
[15] http://www.coreplanner.com/
[16] https://www.planbook.com/

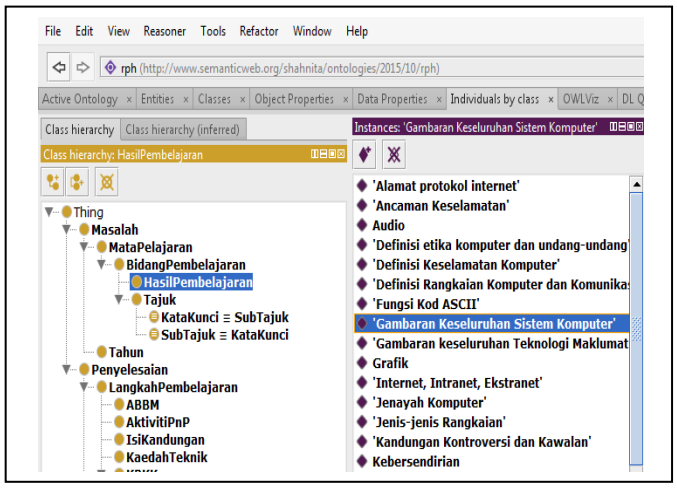


Fig. 5. Protégé 5.0 interface

C. Phase III: Implementation

The main aim of this phase is to change the human-readable representation to machine-readable representation. According to [7], RDF is a standard model for data interchange on the Web. Both RDFS and OWL are modeling languages for describing RDF data. For example, RDFS allows users to express the relationships among data by standardizing them using a flexible, triple-based format and then providing relevant vocabulary or keywords, such as "rdf:type" or "rdfs:subClassOf", which can be used to express such data. On the other hand, OWL is more powerful as it describes data models more efficiently using appropriate database queries and automatic "reasoners". Furthermore, OWL provides useful annotations to help transform the data models into the real world.

Such machine-readable representations include Ontology Web Language (OWL) format, which can be understood by the computer as illustrated in Fig. 6. In this study, Protégé 5.0 was used to convert the conceptual model to such representation.

Fig. 6. The representation in OWL format

D. Phase IV: Evaluation and Maintenance

The third phase involved evaluating and assessing the developed ontology to determine whether it meets the requirement specifications which is to support retrieval. This is in line with was discussed by [22] which state that the organization of elements in knowledge representation must facilitate the retrieval of useful information. As a mean to prove the developed ontology in lesson plan domain, the constructed ontology was implemented in i-Rph[17] system, a semantic application for lesson planning.

This implementation involved the development of a system prototype using the developed knowledge representation. More importantly, the development of the system was based on the prototype development model comprising six (6) phases, namely preliminary study, requirement definition, system design, development and evaluation, implementation and maintenance. Fig. 7 shows all the six phases of the i-Rph system implementation.

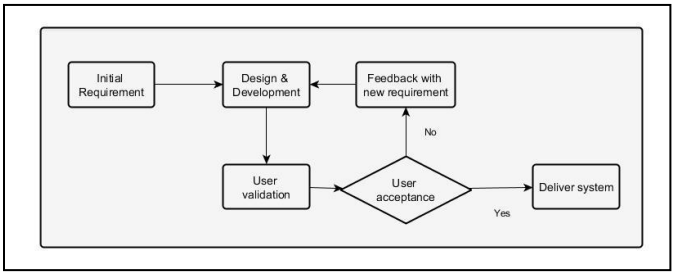


Fig. 7. Evolutionary prototyping model (Adapted from [11])

This phase is challenging as, according to [9], information systems are not easy to be assessed, and there are many aspects to be considered in the assessment process. Moreover, with the emergence of new internet technologies, it is now more difficult to measure IS effectiveness. This is especially true given that internet provides a borderless, non-stop, and flexible communication medium.

Such evaluation will be performed using a quantitative approach by means of an experiment. This experimental approach is selected because it is one of the effective means in evaluating the implementation of a software system [1]. The evaluation phase will be carried out to evaluate the effectiveness of the constructed ontology based on the retrieval mechanism supported by the representation.

The evaluation of the system will be based on the Form Four's ICT curriculum as the scope of the study involves the Daily Lesson Plan (DLP) of the same subject matter. The respondents will be required to construct the lesson plan based on specific details, such as a particular topic or learning objectives. In constructing such lesson plan, they have to use three different applications: a) i-Rph which is the system developed using ontology representation, b) SmartLP system which uses attribute value representation, and c) any free, non-proprietary search engine. They are then required to answer a survey concerning aspects of information quality based on [21] IS Effectiveness Model.

The sample of this study will involve 20 trainee teachers from the Computing Department of the Faculty of Art, Computing, and Creative Industry, UPSI who will be undergoing teaching practicum at several secondary schools. These trainees will teach the Information and Communication Technology (ICT) subject at such schools. The selection of the sample will involve students majoring in computing because of their extensive exposure to the use of technology in education.

The analysis of the empirical data will be performed using Analysis of Variance (ANOVA). Essentially, this statistical

[17] http://irph-dev.upsi.edu.my/

procedure helps compare the mean scores of relevant variables among the three groups of the same population. The following is the information related to the experimental study to be carried out.

- *Sample:* 20 trainee teachers who will be undergoing teaching practicum at several selected schools.
- *Variables :* The independent variables are the criteria of the DLP that will be created. The dependent variables are the matched returned result based on the specified criteria.
- *Null Hypothesis*: The differences in the mean scores of information quality among the three groups are not significant.

$$(H_0: \mu_1 = \mu_2 = \mu_3) \quad (1)$$

IV. Conclusion

Based on the comparison among existing ontology development methodologies, a new ontology development methodology was proposed for the lesson plan domain. In light of the discussed problems, this proposed methodology will serve as a comprehensive, systematic guideline to help system developers produce an ontology for other domains based on a knowledge representation that supports web semantics. The activities involved within each phase, and techniques applied for each activity were clearly explained. Ultimately, this guideline can help in the development of high quality ontology to support all users to perform their task with greater efficacy.

This methodology to support based on a web semantic application can help users gain access to information that is not only ample but also relevant to the preparation of lesson plan. This application can also help overcome the unmanageable amount of information typically produced with the use of normal search engines.

In addition, the same application can help overcome the limitations of databases based on attribute value knowledge representation, which invariably need constant updating by the system administrator. However, the terms in this lesson plan domain were mainly defined using Malay language which might limits the search result in other language. This can be overcome by using alternative terms in an international language such as English and Arabic which is a future plan for this research to support all users to perform their task with greater efficacy.

Acknowledgment

Authors would like to thank the Ministry of Education for funding us with a two year research grant (2013-0157-109-72).

References

[1] A. Dix, “Human-computer interaction,” Springer, US, 2009, pp. 1327-1331.

[2] A. Newell, “The knowledge level,” Artificial intelligence, 18(1), 1982, pp. 87-127.

[3] A. Öhgren, and K. Sandkuhl, “Towards a methodology for ontology development in small and medium-sized enterprises,” In IADIS AC, 2005, pp. 369-376.

[4] A. Saad, “A case-based system for lesson plan construction,” Doctoral Dissertation, Faculty Of Science of Loughborough University, 2011.

[5] A. Saad, P. W. Chung, and C. W. Dawson, “Effectiveness of a case-based system in lesson planning,” Journal of Computer Assisted Learning, 30(5), 2014, pp. 408-424.

[6] C. Ramirez, and B. Valdes, “A General Knowledge Representation Model of Concepts,” Advances in Knowledge Representation, InTech, 2012.

[7] Cambridge Semantic, 2016. Retrieved from http://www.cambridgesemantics.com/semantic-university/rdfs-vs-owl.

[8] D. H. Le, and V. T. Dang, “Ontology-based disease similarity network for disease gene prediction,” Vietnam Journal of Computer Science, 2016, pp.1-9.

[9] G. Tokdemdr, “An Assessment Model For Web-Based Information System Effectiveness,” Doctoral Dissertation, School of Informatics Of The Middle East Technical University, 2009.

[10] J. Angele, D. Fensel, D. Landes and S. Studer, “Developing knowledge-based systems with MIKE,” Automated Software Engineering, 5(4), 1998, pp. 389-418.

[11] J. Mishra, and A. Mohanty, Software Engineering. New Delhi, India: Dorling Kindersley, 2012.

[12] M. Fernández-López, A. Gómez-Pérez, and N. Juristo, “Methontology: from ontological art towards ontological engineering,” 1997.

[13] M. L. Brodie, and J. Mylopoulos, “Knowledge bases and databases: semantic vs. computational theories of information,” New Directions for Database Systems, Ablex, 1986, pp. 186-218.

[14] M. Uschold, and M. King, “Towards a methodology for building ontologies,” Edinburgh: Artificial Intelligence Applications Institute, University of Edinburgh, 1995, pp. 15-30.

[15] N. Xue, S. Jia, J. Hao, and Q. Wang, “Scientific ontology construction based on interval valued fuzzy theory under Web 2.0,” Journal of Software, 8(8), 2013, pp.1835-1842.

[16] P. Hitzler, M. Krötzsch, and S. Rudolph, “Foundations of semantic web technologies,” Chapman and Hall/CRC, Taylor and Francis Group, 6000, 2010, pp. 33487-2742.

[17] R. R. Starr, and J. M. P. De Oliveira, “Concept maps as the first step in an ontology construction method,” Information systems, 38(5), 2013, pp.771-783.

[18] R. Subhashini, and J. Akilandeswari, “A survey on ontology construction methodologies,” International Journal of Enterprise Computing and Business Systems, 1(1), 2011, pp.60-72.

[19] T. Aaberge, and R. Akerkar, “Ontology and Ontology Construction: Background and Practices,” IJCSA, 9(2), 2012, pp.32-41.

[20] T. R. Gruber, “A translation approach to portable ontology specifications,” Knowledge acquisition, 5(2), 1993, pp. 199-220.

[21] W. H. DeLone, and E. R. McLean, “The DeLone and McLean Model of Information Systems Success: A TenYear Update,” Journal of Management Information Systems, 19 (4), 2003, pp. 9-30.

[22] Y. Sun, and Z. Li, “Ontology-based domain knowledge representation,” In 2009 4th International Conference on Computer Science and Education, 2009, pp. 174-177.

A Novel Broadcast Scheme DSR-based Mobile Adhoc Networks

Muneer Bani Yassein*
Department of Computer Science
Jordan University of Science and Technology
Irbid, Jordan, *Corresponding Author

Ahmed Y. Al-Dubai
School of Computing
Edinburgh Napier University
Edinburgh EH10 5DT, UK

***Abstract*— Traffic classification seeks to assign packet flows to an appropriate quality of service (QoS). Despite many studies that have placed a lot of emphasis on broadcast communication, broadcasting in MANETs is still a problematic issue. Due to the absence of the fixed infrastructure in MANETs, broadcast is an essential operation for all network nodes. Although the blind flooding is the simplest broadcasting technique, it is inefficient and lacks resource utilization efficiency. One of the proposed schemes to mitigate the blind flooding deficiency is the counter based broadcast scheme that depends on the number of received duplicate packets between the node and its neighbors, where the node compares the duplicate packet itself and each neighbor node that previously re-broadcasted a packet. Due to the fact that existing counter-based schemes are mainly based on the fixed counter based approach, these schemes are not efficient in different operating conditions. Thus, unlike existing studies, this paper proposes a dynamic counter based threshold value and examines its effectiveness under the Dynamic Source Routing Protocol (DSR) which is one of the well-known on-demand routing protocols. Specifically, we develop in this paper a new counter based broadcast algorithm under the umbrella of the DSR, namely, Inspired Counter Based Broadcasting (DSR-ICB). Using various simulation experiments, DSR-ICB has shown good performance especially in terms of delay and the number of redundant packets.**

Keywords*— *a dynamic counter based; Broadcasting; DSR

I. INTRODUCTION

Mobile Ad-hoc Network (MANET) is an infrastructure-less and self-configuring network that consists of mobile devices which are connected by wireless links. In MANETs, each node is able to move freely and randomly in any direction. Thus, it frequently changes its links to the other nodes and can facilitate applications that require nodes to communicate without depending on a specific infrastructure, such as those used for crisis-management application, emergency, military application, group and collaboration communication and personal network [1, 2].

MANETs are distinguished by some characteristics that make them a good option in different scenarios and networked systems. However, these features pose several challenges and problems to the vast deployment of MANETs. The most important characteristics and challenges of MANETs are summarized as follows: infrastructure-less, mobility, energy conservation, multi-hop routing, self-organization and bandwidth-constrained. In addition, many provocations that need to be overcome such as transmission range and bandwidth limitations and routing overhead [3, 4].

Broadcasting is a fundamental operation in ad hoc networks whereby a source node sends the same packet to all the nodes in the network. In the one-to-all communication pattern, the transmission by each node can reach all nodes that are within its transmission radius. Broadcasting has many significant applications and several ad hoc network protocols assume the availability of an underlying broadcast service [5]. The route discovery in reactive protocols is one if the applications that benefit from the broadcast communication. For instance, some MANET routing protocols such as Dynamic Source Routing (DSR) [6] use broadcasting or simple flooding as one of its derivatives to establish routes that cause the broadcast storm problem. The broadcast storm problem can be avoided by deploying efficient broadcast algorithms that aim to reduce the number of nodes that retransmit the broadcast packet and ensure, at the same time, that all nodes receive the packet. Proper use of a counter based broadcasting method can reduce the number of rebroadcasts, and as a result reduces the chance of contention and collision among neighboring nodes. Motivated by these observations, this study will focus on the performance of DSR routing protocol used in MANET using proper counter based schema algorithm called Inspired Counter Based Broadcasting (DSR-ICB).

DSR is a source-routed reactive routing protocol in which each data packet has complete routing information to reach the destination, and each node uses caching technology to maintain routing information. When a new route is created the node updates its route caches. It works in two phases, namely, route discovery and route maintenance. When the source node wants to send a packet to a destination, it searches in the node's route cache to find out if it is already has got a route to the destination. If it exists then, it uses that route to send the packet. However, if it is not the case, it starts its route discovery phase by broadcasting a route request packet using blind flooding. That contains the address of the source, destination, and a unique id number. Each intermediate node checks whether it has a route to the destination. If not, it adds its address to the route record of the packet and forwards the packet to its neighbors. If the node has not seen the packet before and its address does not exist in the route record of the packet, the route request is performed using a new counter-based broadcast algorithm present a main cut of the number of redundant rebroadcast packets along with good results in terms of reachability, saved rebroadcasts, latency, packet delivery ratio and routing overhead and an extensive comparison against counter based schemes have been performed using Qualnet simulator [7]. The rest of this paper is organized as follows. In Section 2, we review the previous work of broadcasting,

Section 3 presents the new algorithm proposed a counter based broadcast scheme, Section 4 the obtained simulation results and analysis of the proposed scheme are described. Finally, Section 5 concludes this study.

II. Related Work

In MANETs, any node can broadcast the packet at any time, and there is no acknowledgment for receiving that packet. In MANET, due to node mobility, broadcasting is expected to be executed more frequently to perform essential operations such as finding a route to a particular node [2, 3]. Because radio signals are likely to overlap with others in a given geographical area, a straight forward broadcast by flooding is usually expensive and results in the broadcast storm problem. In [5, 8], they have proposed several schemes to reduce redundant rebroadcasts and differentiate timing of rebroadcasts to alleviate this trouble. Williams and Camp [9] have classified the broadcast protocols into simple flooding, probability based schemes, counter-based schemes, distance based, location-based schemes and neighbor knowledge schemes. Bani Yassein et al. [10] have proposed a probabilistic scheme where the probability p is computed from the local density n to achieve reachability of the broadcast. However, this scheme has the disadvantage of being "locally uniform" in that each node of a given area receives a broadcast and determines the probability according to a static efficiency parameter to achieve the reachability as well as from its local density [4, 11].

Several research studies have suggested broadcast algorithms [12, 13] that are based on the clustered architecture as a way to improve broadcasting reliability. In these algorithms, each cluster has one cluster head that dominates all other members in the cluster and computes its forward node set locally. Therefore, there is a need to distribute the responsibility of being a cluster head to all nodes (load-balancing). While the load-balancing algorithms have anticipated for the routing problem in the mobile ad-hoc network, no attempt has so far been made to introduce such algorithms within the context of broadcasting. Bani Khalaf et al. [9] have used velocity aware probabilistic route discovery models to exclude unstable nodes while constructing routes between the source and its destination. Zhang et al. [14] proposed an adoption broadcast technique. This scheme is based on the RTS/CTS frames at the MAC layer. However, using RTS/CTS frames can affect the exposed station problem. Bani Yassein et al. [15] proposed a probabilistic, fuzzy logic-based, distance broadcasting scheme for MANETs using a broadcasting scheme based on the fuzzy logic concept at every node for generating a dynamic probability value based on the node location. The node location was classified into four locations (border, internal border, exterior, and interior) which were assigned four probability values, namely, high, medium, low, very low respectively. The results have shown that the fuzzy logic control scheme was much better than the smart probabilistic scheme. The results also indicated that the proposed algorithm generated much higher saved re-broadcast and throughput. The drawback of this study is that it assigned the always higher probability to the border than other locations which are considered to be incorrect for all cases. This is because in some cases the other location might be better than the border. In [12] authors have proposed a local broadcast algorithm to reduce the number of transmissions. Unfortunately, however, the local broadcast algorithms are based on static approach, and it is worth indicating that finding the minimum connected dominating set is an NP-complete algorithm. Bani Yassein, et al. [16-18], have proposed counter based solutions that can achieve a higher degree of reachability. This scheme enables the nodes to rebroadcast the message only if the number of received duplicate packets is less than a threshold value by taking into account the node density either in dense or sparse areas.

In an attempt to fill in this gap, the proposed scheme maintains a better performance than the existing counter-based schemes in terms of different metrics. However, the proposed scheme would provide more efficient and dynamic result if it contains more counter values instead of four values. The proposed algorithm was implemented under AODV reactive routing protocol and the highly adjusted counter-based scheme will be used within DSR.

III. The Proposed Scheme

In the light of the above observations, our goal has been set to design an efficient and scalable broadcast algorithm based on some received duplicate packets. Since the node counter increases by one with every time of receiving the same broadcast message. Predefined threshold value is already given to each node to initiate rebroadcasting according to that threshold value. Hence, a new counter-based broadcast algorithm will be developed for MANETs. The algorithm has to present a major decrease in the number of redundant rebroadcast packets along with good results regarding reachability.

The Adaptive Counter-based Broadcasting Scheme is based on the idea of initiating a counter c that keeps track of the received packets and therefore counts their number. Another counter threshold defined according to the node neighboring status. Since the nodes located in a sparse area have different threshold value than the nodes located in Medium, dense or even extra dense area. The threshold values threshold1, threshold2, threshold3 and threshold4, threshold5 assigned to the sparse, medium, dense and extra dense area respectively [17]. After determining these threshold values according to the neighboring information, the comparison process will take place between the counter c which indicates the number of received duplicate packets and one of those threshold values. Whenever the counter c is less than the threshold value, the broadcasting will continue. Otherwise, it will stop. The proposed algorithm performs as follows: whenever a node X hears a broadcasted packet m, the node rebroadcasts that packet if it's received for the first time in addition to taking the node density into consideration as the description bellow where threshold1, threshold2, threshold3, threshold4 and threshold5 are predefined values and threshold1<threshold2 < threshold3<threshold4<threshold5.

Firstly, node X rebroadcasts the packet according to threshold1 if the node located in a sparse area, which means its neighbor numbers is less than or equivalent to the minimum average number of neighbors avg_neighbors1. Secondly node X rebroadcasts the packet according to threshold2 if the node is located in a medium area, i.e., its neighbor numbers is greater

than the minimum number of neighbors avg_neighbors1, and less than or equal the maximum numbers of neighbors avg_neighbors2. Thirdly, node X rebroadcasts the packet according to threshold3, if the node is located in a dense area, i.e., its neighbor numbers is greater than the maximum numbers of neighbors avg_neighbors2, and less than or equal the maximum extra numbers of neighbors avg_neighbors3. Fourthly, node X rebroadcasts the packet according to threshold4 if the node is located in extra dense area, which means its neighbor numbers is greater or equal the extra maximum numbers of neighbors avg_neighbors3.

In this section, we evaluate the performance of DSR-ICB or DSR-5C as a broadcasting algorithm which can be applied to any broadcast operation. Such an operation is the dissemination of route requests (RREQ) in the route discovery process of on-demand routing protocols. For the purpose of discovering a route to a destination, it suffices that the RREQ reaches those nodes with a route to the desired destination.

To study the impact of ICB on the route discovery process, we implemented it as the basis for deciding which nodes should broadcast RREQ messages in the route discovery process of DSR. We named the resulting protocol (DSR-ICB) or DSR-5C and implemented it in Qualnet [7]. To compare DSR-5C against DSR with flooding (DSR-F), we use traffic and mobility models similar to those previously reported for the performance of DSR [6]. To address reliability, we have used three versions of DSR. First, DSR with DSR- ICB we implement (DSR- ICB) or (DSR- 5C) as described in Fig. 1. Second, DSR with an adjusted counter based (DSR-4C) [17, 8] and the last one DSR with flooding (DSR- F) [6].

Algorithm 1: Inspired Counter Based Broadcasting: (DSR-ICB) or (DSR-5C).

Input : BROADCAST MESSAGE (MSG)
Output: DECIDE WHETHER TO REBROADCAST MSG OR Get the broadcast ID from the packet; n1 minimum numbers of neighbour; n2 Medium number of neighbour and n3 maximum number of neighbour, n4 extra maximum n5 Ultra extra maximum number of neighbour

```
ON HEARING A BROADCAST PACKET AT NODE X
Get degree n of node X
if n < n1 then
    Node X has a low degree: the low threshold value
    (threshold=c1);
end
else if n > n1 and n ≤ n2 then
    Node X has a medium degree: the medium
    threshold value (threshold=c2);
end
else if n > n2 and n ≤ n3 then
    Node X has a high degree: the high threshold value (threshold = c3) ;
end
else if n > n3 and n ≤n4 then
    Node X has a very high degree: the high threshold value
    (threshold=c4);
end
else if n > n4 and n ≤ n5 then
    Node X has Ultra degree: the high threshold value
    (threshold=c5)
end
while not hearing a message do
    Wait for a random number of slots.;
    Submit the packet for transmission and wait until the transmission
    actually start;
    Increment c
end
if c < threshold then
    Submit the packet for rebroadcast
end
else
    Drop the packet from rebroadcast
end
```

Fig. 1. Inspired Counter Based Broadcasting

IV. Simulation Classification Results and Analysis

The Qualnet simulator has been used in this study which is a well- known discrete-event simulator, originally designed for wired networks and has been subsequently extended to support simulations in mobile wireless (and MANET) settings [7]. Experiments are repeated for 20 trials with different random-number seeds, traffic endpoints. Hence, all protocols are compared having identical node mobility and traffic demands. Each data point represents the average of the 20 trials. Due to the fact that the proposed algorithm is based on a counter based approach DSR with five counters (DSR-5C), it does not fit in every case. It is worth indicating that here is a small chance that the route requests cannot reach the destination in our counter based algorithm. In such cases, we have to generate the route request again if the previous route request failed to reach the destination. The DSR protocol, on the other hand, uses flooding in route discovery phase. Therefore, all route requests will reach their destinations if the network is not partitioned. Based on this analysis, our algorithm should perform better than DSR with flooding (DSR-F) and DSR with four counters (DSR-4C) in dense networks with heavy traffic [17, 8].

In our simulations, two nodes are selected as data sources. A CBR traffic generator has been attached to the sources. The mobility model chosen is the random waypoint model due to its suitability for the ad hoc environments. 50 nodes are placed randomly on a 1000m x 1000m area and having a bandwidth of 2Mbps.

1) Effects traffic loads

To study the effects traffic loads, i.e. in this work varied traffic load has been used from light traffic through moderate to heavy traffic. To do this, the following rates of broadcast packets generated at the source node will be considered:

- Light traffic load: 1 packets/sec;
- Medium traffic load: 5 packets/sec;
- Heavy traffic load: 11 packets/sec;

We measure the broadcast latency for three approaches. We record the start time of broadcast as well as the time when the broadcast packet reaches the last node. The difference between these two values is used as the broadcast latency. Since packet

rebroadcasts collide and content for the channel with each other, and the counter based approach incurs the lowest number of rebroadcasts, it should have the lowest latency, which is affected mainly by the number of total packets transmitted in the channel. If the number of packets is high, then the number of collisions is also high, and in turn more retransmissions are needed. As a result, fewer packets lead to lower delays. Fig. 2 shows the end-to-end delay for different levels of traffic load. As expected, our DSR-5C exhibit lower latency than the DSR-4C and DSR-F.

Fig. 3 illustrates that DSR-5C algorithm can significantly reduce the number of rebroadcasts for a network with 50 nodes and maximum speed 1 m/s. It also shows the number of route request rebroadcast increases when the traffic load grows. Fig. 4 depicts that the reachability increases when the network traffic load increases, regardless of what kind of the algorithms is used. The DSR-5C algorithm has the best performance in terms of reachability, reaching nearly 1. The performance of DSR-4C shows that the reachability is above 95% in network traffic load equal 10.

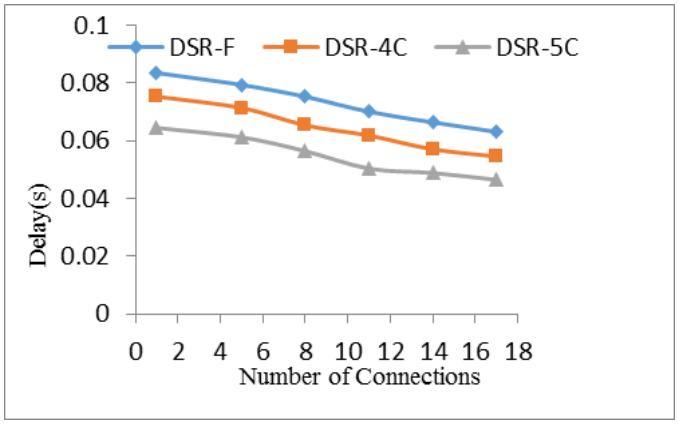


Fig. 2. End To End Delay Vs. Number Of Connections

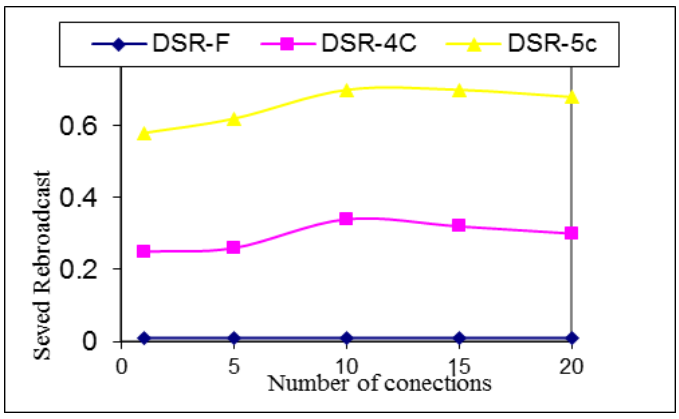


Fig. 3. Saved Rebroadcast vs. number of connections

In Fig. 5, we compare the packet delivery ratio for different network traffic load. It shows packet delivery ratio decreases when the number of connections increases. The drop of packet delivery ratio is caused by the fact that the movement of destination or intermediate nodes may incur route expiration and route request retransmission. Fig.5 also shows that DSR-5C outperforms DSR-F and DSR-4C. As mentioned before, the packet delivery ratio improvement of DSR-5C is due to its reduction of rebroadcasting. Fig. 6 shows routing overhead increases when traffic load increases. The more connections created, the more route requests that lead to more rebroadcasts, higher bandwidth consumption and higher routing overhead. It also shows that DSR-5C outperforms DSR-4C, DSR-F by about 20%.

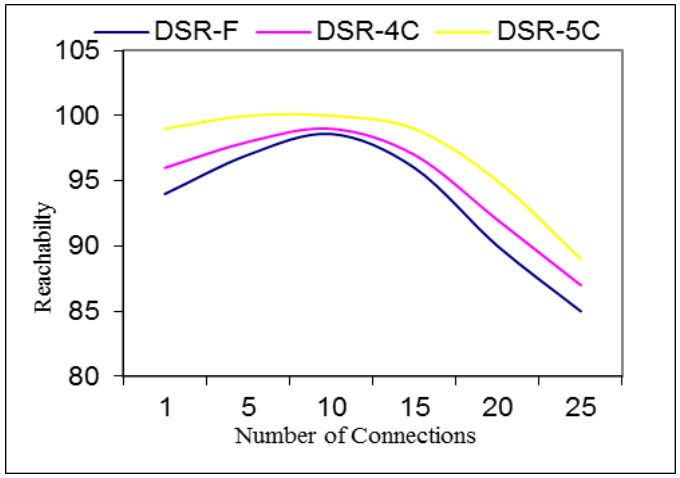


Fig. 4. Reachability vs. number of connections

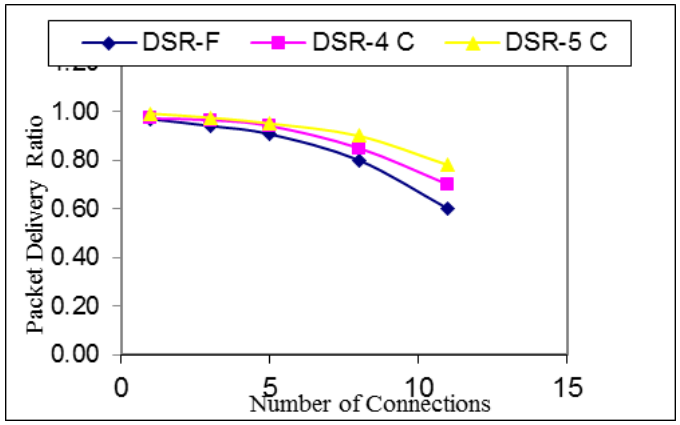


Fig. 5. Packet delivery ratio vs. number of connection

2) Effects of node Speed

In these experiments, we are investigating the effect of the node speed on the performance of the examined protocols using 50 nodes. In all scenarios of this experiment, the range of the maximum node speed is from 1m/s up to 20 m/s. The network is considered to be sparse when the number of nodes is 25 nodes and examined to be dense when the number of nodes is 100 nodes.

Fig. 7 presents the average end-to-end delay for the three schemes evaluated at different node speeds. It reveals that when the node speed increases, the delay is also growing. This is because that whenever the node speed increases, there are more generated data packets competing for the limited buffer space. As a result, more queuing and buffer overhead experienced, and thus, higher delay in addition to the higher

node speed, which affect the stability of the network. That is why the network could have more breakage links and therefore failure in delivering the packet to the destination. This will result in generating extra RREQ and exhibiting growing end-to-end delay. The new proposed scheme outperforms the other schemes in terms of the average delay by 66% in compared with blind flooding and 35% compared with three counter-based scheme for all node speeds.

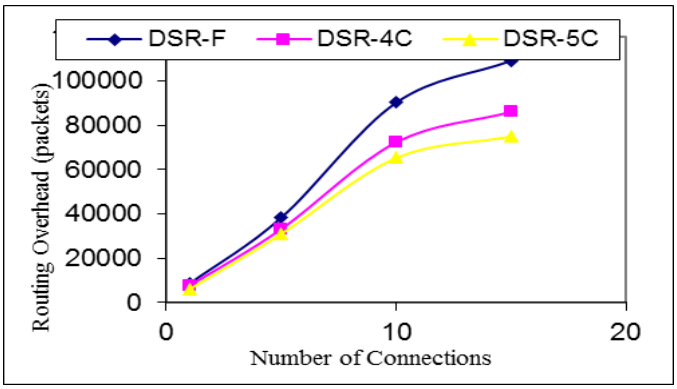


Fig. 6. routing overhead vs. number of connections

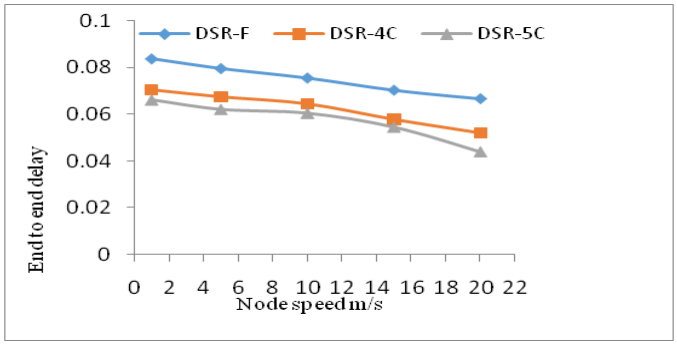


Fig. 7. Average end-to-end delay vs. node speed

Fig. 8 presents delivery ratio for the three examined schemes against different node speeds. Fig.8 shows that whenever the node speed increases the PDR decreases, because the more speed generated the less stable network links.

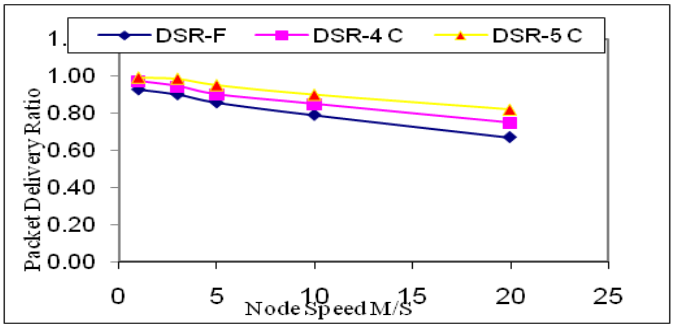


Fig. 8. Packet delivery ratio vs. node speed

In Fig.8 it can be noticed that our proposed scheme maintains almost the same performance level that obtained from the previous examined schemes with varying node speeds. Fig. 9 shows that DSR-5C algorithm can significantly reduce the number of rebroadcasts for a network with 50 nodes and different node speeds = 1, 5, 10 and 20 m/s.

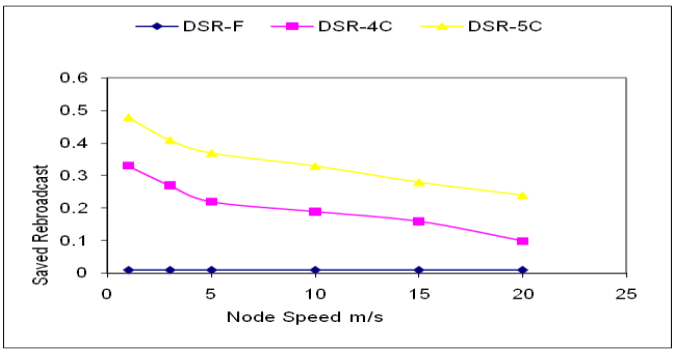


Fig. 9. Saved Rebroadcast vs. Node speed

Furthermore, our algorithm shows its superiority over DSR-F and DSR-4C for all speed values in terms of packet delivery ratio. We can conclude from the results that our protocol maintains the high level of enhancement over the other two protocols DSR-F and DSR-4C despite the number of nodes involved. The results of the end to end delay show how our protocol exceeds the two other protocols DSR-F and DSR-4C for all speed values.

V. CONCLUSION

In this study, we have presented a new dynamic counter based route discovery algorithm for mobile ad hoc networks. Various simulation experiments have been conducted to examine the proposed algorithm under different operating conditions. The results concluded that the new dynamic counter based broadcasting scheme overcomes the limitations of the other evaluated schemes in terms of alleviating the broadcast storm problem, collision, contention and redundant packets transmission. In addition, the proposed scheme achieves high packets delivery ratio with a reduced level of delay while keeping the routing overhead to a minimum. To improve the route discovery process of on-demand routing protocols, a new counter based broadcast is implemented in DSR (the new variant is named (DSR-5C) as the mechanism for disseminating RREQ messages. Our results confirm that DSR-5C improves, in all aspects, the performance of DSR in traffic load scenarios.

We also show the performance of DSR with flooding based on both fixed and adjusted counter based under different working conditions. A performance analysis has revealed that the DSR-5C algorithm has superior latency characteristics over those of the well-known DSR with flooding algorithm. One of the key results is that the performance of the proposed algorithms scales up well with the growing saved rebroadcasts. The dynamic counter-based broadcasting finds a solution to the challenging problem of route discovery in MANETs. However, there are still several interesting issues and unsolved problems that require further investigation. As a next step, we are willing to investigate the effect of the proposed approach with a broad range of routing protocols and different mobility models such as Manhattan Model, Trace-based Models and Pathway model. Besides, we aim to explore the correlation between the rebroadcast probability and the counter between mobile nodes.

Moreover, we will also develop a hybrid approach that combines the features of both counters based and distance based schemes. It is anticipated that this direction will enable us to explore fully the routing discovery challenging issues and present efficient solutions accordingly

ACKNOWLEDGMENT

A preliminary version of this paper appeared in 14th IEEE International Conference on Ubiquitous Computing and Communications (IUCC-2015), At Liverpool, England, UK, 26-28 October 2015,. This version includes a concrete analysis on Counter Based Broadcasting in Dynamic Source Routing Protocol. This research was supported by Jordan University of Science and Technology. We express our thanks and my sincere appreciation to Jordan University of Science and Technology, for their financial and logistical support for providing me with the necessary guidance concerning this work during my sabbatical leave at Edinburgh Napier University. I am also grateful to the school of computing at Edinburgh Napier University for technical support

REFERENCES

[1] M. Varghese, V. Parthasarathy , "An Improved Highly Dynamic Choice Routing Scheme (I-HDCR) for Mobile Ad Hoc Networks," International Journal of Applied Engineering Research, Vol. 9 Issue 22,, pp. 17709-17718, 2014.

[2] Mary, Arul, S. A. Sahaaya, and Gnanadurai Jasmine Beulah. "Intra-Cluster Optimization in Zone-Based Wireless Sensor Networks Using DBSCAN."*International Journal of Applied Engineering Research* 10.7: p18811-18822, 2015.

[3] DG. Reina, SL Toral, P Johnson, F Barrero, "A survey on probabilistic broadcast schemes for wireless ad hoc networks." *Ad Hoc Networks* 25 (2015): 263-292.

[4] Bani Yassein M, Bani Khalaf M, and Al-Dubai A. "A performance comparison of smart probabilistic broadcasting of ad hoc distance vector (AODV)." *Journal of Supercomputing*, vol. 53, no.1, pp. 196-211 Springer Journal, 2010.

[5] Williams, Brad, and Tracy Camp. "Comparison of broadcasting techniques for mobile ad hoc networks." *Proceedings of the 3rd ACM international symposium on Mobile ad hoc networking & computing*. (MOBIHOC 2002), 194–205, 2002.

[6] D. A. Maltz, D. B. Johnson and Y. Hu. "The dynamic source routing protocol (DSR) for mobile ad hoc networks for IPv4. RFC 4728," The Internet Engineering Task Force, Network Working Group, Feb 2007. http://www.ietf.org/rfc/rfc4728.txt.

[7] QualNet and SCALABLE Network Technologies." http://web.scalable-networks.com/content/qualnet. [Accessed: 28-12-2013].

[8] Bani Yassein M, Al-Hameed A, and Constandinos Mavromoustakis. "Adaptive counter-based broadcasting scheme in mobile ad hoc networks." *Proceeding of the 15-th ACM MSWiM 2012, HP-MOSys, 1-st ACM Workshop on High Performance Mobile Opportunistic Systems (HP-MOSys 2012) pages 47-52, Paphos, Cyprus, 2012.*

[9] Bani Khalaf M, Al-Dubai A Y, and Geyong Min. "New efficient velocity-aware probabilistic route discovery schemes for high mobility Ad hoc networks." *Journal of Computer and System Sciences* 81.1, pp. 97-109, 2015.

[10] Kabir, Tanjida, Novia Nurain, and Md Humayun Kabir. "Pro-AODV (Proactive AODV): Simple modifications to AODV for proactively minimizing congestion in VANETs." *Networking Systems and Security (NSysS), 2015 International Conference on*. IEEE, 2015.

[11] Bani Yassein M, Alslaity A N, Ababneh I. M, "Performance Evolutions of Velocity Aware Routing Protocol for Mobile Ad hoc Networks ", Proceeding of the Sixth International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies(UBICOMM 2012, - Barcelona, Spain, September 23-28, 2012.

[12] Wu, Jie, and Hailan. Li. "On calculating connected dominating set for efficient routing in ad hoc wireless networks." *, vol. 18, no. 1–3, 2001*.

[13] Lin, Chunhung Richard, and Mario Gerla. "Adaptive clustering for mobile wireless networks." *IEEE JSAC*, 15.7 (1997): 1265-1275.

[14] Q. Zhang and D. P. Agrawal. Dynamic probabilistic broadcasting in MANETs, Journal of Parallel Distributed Computing, volume 65(2), pages 220-233, May 2005.

[15] Bani Yassein M, Al-Hameed A, Mardini W, and Khamayseh Y, Performance Analysis of Adjusted Counter Based Broadcasting in Mobile Ad Hoc Networks. *Communications and Network*, Vol. 5 No. 4, 353-359, November *2013*.

[16] Bani Yassein M, Nimer S F, and Al-Dubai A. "A new dynamic counter-based broadcasting scheme for Mobile Ad hoc Networks."Simulation Modelling Practice and Theory 19.1 (2011),pp. 553-563.

[17] Muneer Bani Yassein and Ahmed Yassin Al-Dubai Inspired Counter Based Broadcasting in Dynamic Source Routing Protocol,, Conference: 14th IEEE International Conference on Ubiquitous Computing and Communications (IUCC-2015), At Liverpool, England, UK, 26-28 October 2015.

[18] Tseng, Yu-Chee. "The broadcast storm problem in a mobile ad hoc network." *Wireless networks* vol. 8, no.2, pp.153-167, 2002.