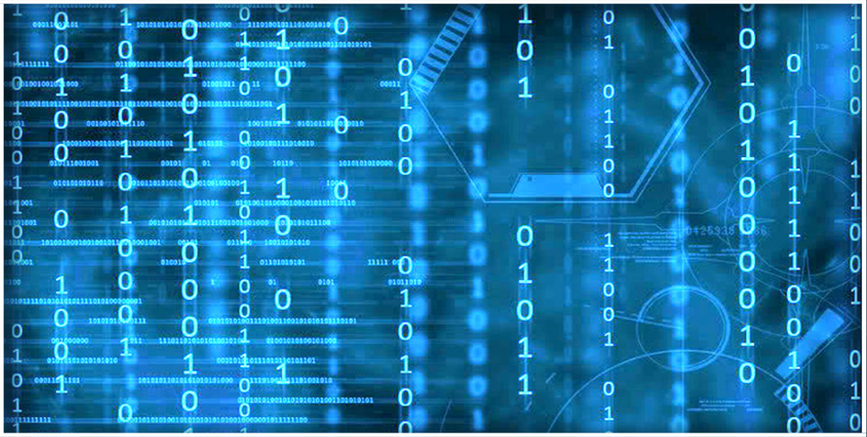
IJACSA

WHERE WISDOM SHARES

International Journal of Advanced Computer Science and Applications

Volume 9 Issue 6

June 2018



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

SAI

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 9 Issue 6 June 2018
ISSN 2156-5570 (Online)
ISSN 2158-107X (Print)

Editorial Board

Reviewer Board Members

Sudarson Jena

Sudipta Roy

Suhail Sami Owais Sami Owais Owais

Suhas J Manangi

SUKUMAR SENTHILKUMAR

Süleyman Eken

Sumazly Sulaiman

Sumit Goyal

Sunil Phulre

Suparerk Janjarasjitt

Suresh Sankaranarayanan

Surya Narayan Panda

Susarla Venkata Ananta Rama Sastry

Suseendran G

Suxing Liu

Syed Asif Ali

T C.Manjunath

T V Narayana rao Rao

T. V. Prasad

Taghi Javdani Gandomani

Taiwo Ayodele

Talal Bonny

Tamara Zhukabayeva

Taner Tuncer

Tanvi Banerjee

Tanweer Alam

Tanzila Saba

TAOUFIK SALEM SAIDANI

Tarek Fouad Gharib

tarig ahmed

Taskeed Jabid

Tasneem Bano Rehman

thabet Mohamed slimani

Totok R. Biyanto

Touati Youcef

Tran Xuan Sang

TSUNG-CHUAN MA

Tsvetanka Georgieva-Trifonova

Uchechukwu Awada

Udai Pratap Rao

Urmila N Shrawankar

V Baby Deepa

Vaidas Giedrimas

Vaka MOHAN

Venkata Raghavendran Chaluvadi

VENKATESH JAGANATHAN

Vijay Bhaskar Semwal

Vijayarani Mohan S

Vijendra Singh

Vinayak K Bairagi

VINCE PAUL A

Visara Urovi

Vishnu Narayan Mishra

Vitus S.W. Lam

VNR SAIKRISHNA K

Voon Ching Khoo

VUDA SREENIVASARAO

Wali Khan Mashwani

Wei Wei

Wei Zhong

Wenbin Chen

Wenzhao Zhang

Wichian Sittiprapaporn

Xi Zhang

Xiao Zhang

Xiaojing Xiang

Xiaolong Wang

Xunchao Hu

Y Srinivas

Yanping Huang

Yao-Chin Wang

Yasser M. Alginahi

Yaxin Bi

Yi Fei Wang

YI GU

Yihong Yuan

Yilun Shang

Yu Qi

Zacchaeus Oni Omogbadegun

Zaffar Ahmed Shaikh

Zairi Ismael Rizman

Zarul Fitri Zaaba

Zeki Yetgin

Zenzo Polite Ncube

ZHENGYU YANG

Zhigang Yin

Zhihan Lv

Zhixin Chen

Zia Ur Rahman Zia

Ziyue Xu

Zlatko Stapic

Zne-Jung Lee

Zuraini Ismail

CONTENTS

Generating Classification Rules from Training Samples

Arun D. Kulkarni
Department of Computer Science
University of Texas at Tyler
Tyler, TX, USA

***Abstract*—In this paper, we describe an algorithm to extract classification rules from training samples using fuzzy membership functions. The algorithm includes steps for generating classification rules, eliminating duplicate and conflicting rules, and ranking extracted rules. We have developed software to implement the algorithm using MATLAB scripts. As an illustration, we have used the algorithm to classify pixels in two multispectral images representing areas in New Orleans and Alaska. For each scene, we randomly selected 10 per cent of the samples from our training set data for generating an optimized rule set and used the remaining 90 per cent of samples to validate the extracted rules. To validate extracted rules, we built a fuzzy inference system (FIS) using the extracted rules as a rule base and classified samples from the training set data. The results in terms of confusion matrices are presented in the paper.**

Keywords—Fuzzy membership functions; classification; rule extraction; multispectral images

I. Introduction

Many methods have been used to classify pixels in multispectral images using training samples. These include parametric methods such as the maximum likelihood, support vector machines, decision trees, neural networks, fuzzy-neural systems, and fuzzy inference systems. In supervised classification methods during the learning phase, a model is built to map an input feature vector to output classes, and during the classification phase the model is used to classify an unknown sample. The maximum likelihood classification algorithm assumes normal distribution and uses the mean vector and covariance matrix of each class to find the posterior probability. It then assigns a pixel to the class with the higher posterior probability. The Support Vector Machine (SVM) partitions the feature space by using hyper-planes that maximize the distance between the two classes in the feature space [1]. It has been shown that the SVM algorithm yields higher classification accuracy for small datasets compared to conventional classifiers [2]. Neural networks provide a nonparametric method for classification. Neural network models learn from training samples. During the learning process weights are updated using a gradient descent method such that the mean squared error between the desired and actual outputs is minimized [3]. During the decision-making phase the model is used to classify pixels based on their spectral signatures.

Fuzzy-neural systems have been used to classify pixels in Landsat images [4]. Fuzzy logic provides a tool to process information using linguistic rules. Fuzzy logic in the form of approximate reasoning provides decision support and expert systems with powerful reasoning capabilities. In fuzzy logic class memberships based on a degree of compatibility with the concepts presented are used [5]. A fuzzy inference system (FIS) provides a method to classify pixels in Landsat images. However, the potential of fuzzy inference systems has not been fully explored by the remote sensing community as of yet. The main task in implementing a FIS is to develop a rule base. Classification rules can be generated from training samples or can be obtained from expert's knowledge. These classification rules then can be used to build the FIS. Several methods to generate classification rules from training samples have been reported in the literature. They include extracting classification rules using fuzzy membership functions, decision trees, neural networks, and black-box models. Wang and Mendel [6] suggested a method to extract fuzzy rules from data samples using fuzzy membership functions. They have used the method for a time-series prediction problem, where the output function is a continuous function. Chiu [7] developed a method called subtractive clustering to efficiently extract rules from a high dimensional feature space. The method was able to produce a much simpler fuzzy classifier and could be used to extract rules for function approximation as well as pattern classification. Kulkarni and McCaslin [8] have generated classification rules from neural network models and have built a FIS to classify pixels in Landsat images. Fung et al. [9] developed a cost-efficient method to quickly extract rules from SVMs trained with thousands of samples. Their algorithm forms rule sets that can be easily understood by humans, and only needs simple multivariable optimization problems to be solved. Sicat et al. [10] developed the FIS using farmer's knowledge for agricultural land sustainability classification using fuzzy models. Reshmidevi et al. [11] have developed a fuzzy rule base system for land suitability in agricultural watersheds. They have considered two types of attributes: continuously measured attributes and thematic attributes, and the crop suitability index as the output of the fuzzy rule-based system. They have used heuristic information and farmer's knowledge aggregated through field surveys as the basis for the fuzzy rule-base. Cay and Iscan [12] have developed a fuzzy expert system for land reallocation in land consolidation. They developed a rule base system using farmer's knowledge obtained from survey questions. Meng and Pei [13] have suggested a method to extract linguistic rules from data sets using fuzzy logic and genetic algorithms. They have formalized linguistics based on complex data summaries and used a genetic algorithm to optimize the number of parameters of membership function of linguistic values. Kulkarni and Khan [14] generated rules to classify Likert-scale survey data by using a multi-layered feed forward neural network. Kulkarni and Shrestha [15] have generated rules using induction trees and built a FIS using the extracted rules.

In this paper we have used the method similar to that suggested by Wang and Mendel [6] for classification of pixels in a Landsat images. In rule extraction the main concerns are the number of extracted rules and the quality of those rules. Technically, each training sample generates a rule, and we get a large number of rules. It is important to note that the generated rules often contain redundant and conflicting rules. Also, a rule set with a large number of rules results in a model that often over-fits the data samples. Generally, rule generation is a two-step process. During the first step all possible rules are generated. In the second step, the rule set is optimized. The suggested algorithm for rule generation is as follows: First, the training data is fuzzified. From the fuzzified data, rules are generated. The generated rules may contain redundant and conflicting rules which are then eliminated. The remaining rules are ranked.

As an illustration, we have considered Landsat scenes from areas in New Orleans and Alaska. We selected training set areas interactively by displaying the scenes. We extracted classification rules from training samples. We built a FIS for each scene using the extracted rule as the rule base and classified all pixels. The outline of the paper is as follows. Section II describes a method for generating classification rules from training samples and optimizing the rule set. Section III provides implementation and results of Landsat data analysis. Section IV provides discussions and results.

II. Rule Generation and Optimization

The proposed method for extracting classification rules from data samples and finding the optimized rule set by eliminating conflicting and redundant rules is shown in Fig. 1. The process includes five steps. The first two steps are concerned with rule generation and the last three steps deal with optimization. To illustrate the method, we have chosen a classification problem with two features and three classes, and the training set contains fifty samples from each class. The method can be extended to multiple features and multiple classes. The steps are explained below.

Step-1 Fuzzify Data: We assume a set of desired input-output data pairs as shown in (1).

$$\left(x_1^1, x_2^1, y^1\right), \left(x_1^2, x_2^2, y^2\right), \ldots, \left(x_1^n, x_2^n, y^n\right) \tag{1}$$

where $\left(x_1, x_2\right)$ represents features, and y represents the corresponding class. For each feature the domain interval is 0 through 10. We divided the domain interval with three fuzzy sets {low, medium, high}. We used trapezoidal membership functions as shown in Fig. 2.

Step-2 Rule Generation: We fuzzified the input values and generated classification rules. Let the input vector $\left(2.3, 3.5\right)$ represent class C_1. From the membership functions shown in Fig. 2, membership values are given by (2), and the corresponding rule can be stated as If x_1 is low and x_2 is medium then the class is C_1

$$\begin{aligned} &\mu_{low}\left(x_1\right) = 0.7,\ \mu_{med}\left(x_1\right) = 0.2, \mu_{high}\left(x_1\right) = 0.0 \\ &\mu_{low}\left(x_2\right) = 0.0,\ \mu_{med}\left(x_2\right) = 1.0, \mu_{high}\left(x_2\right) = 0.0 \end{aligned} \tag{2}$$

We generate a rule using the highest membership values. The firing strength of a rule is given by (3).

$$\begin{aligned} \alpha &= \min\left(\mu_{low}\left(x_1\right), \mu_{med}\left(x_2\right)\right) \\ &= \min(0.7,1.0) = 0.7 \end{aligned} \tag{3}$$

Each sample pair generates a rule, and the total number of generated rules is equal to the number of samples. The extracted rules contain duplicate and conflicting rules.

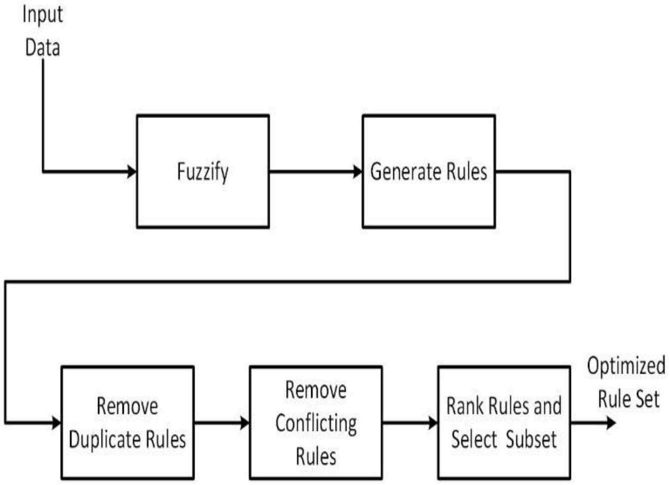


Fig. 1. Rule generation and optimization process.

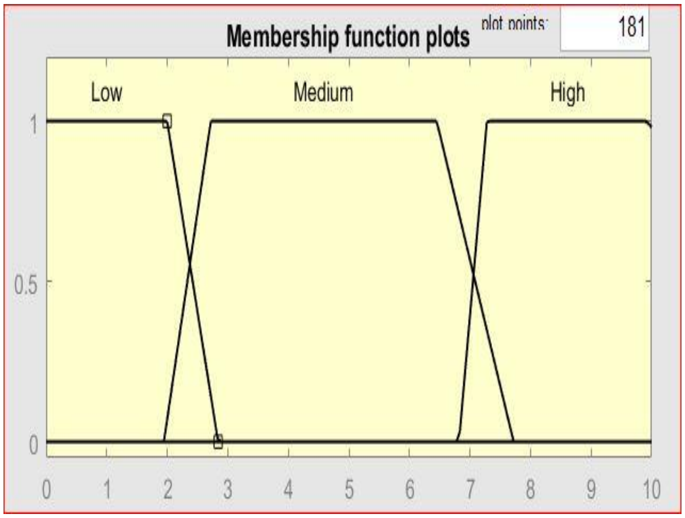


Fig. 2. Fuzzy membership functions.

Step-3 Eliminate duplicate rules: To eliminate repeated rules, extracted rules are mapped onto the Fuzzy Associative Memory (FAM) banks as shown in Fig. 3. In this example there are three classes and there are 50 samples in each class. There are 150 rules generated as each sample generates a rule. We used three FAM banks, one for each class. Each cell in a FAM bank represents a rule, and the value in the cell represents the count of that rule. It can be seen from Fig. 3 that a rule is as follows: If x_1 is low, and x_2 is low, then class is C_1. The count for the rule is 32. That means 32 samples satisfied this rule. Looking at the FAM bank in Fig. 3, we can see that by eliminating repeated rules, we get a rule set of only 10 rules. The extracted rules are shown in Table I.

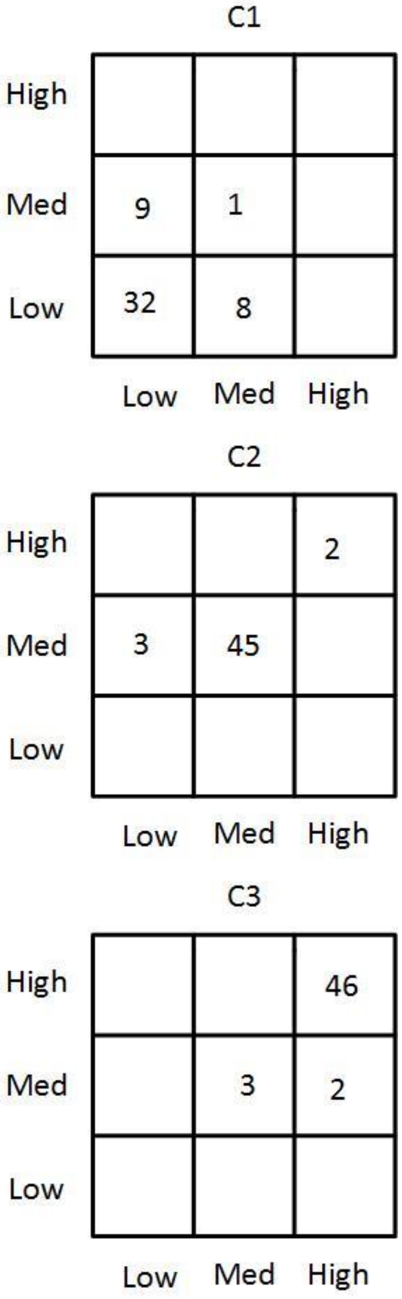


Fig. 3. Fuzzy associative memory (FAM) bank.

TABLE I. RULE SET AFTER ELIMINATING REPEATED RULES

Rule	X_1	X_2	Class	Count
R_1	low	low	C_1	32
R_2	low	medium	C_1	9
R_3	medium	low	C1	8
R_4	medium	medium	C1	1
R_5	medium	medium	C2	45
R_6	low	medium	C2	3
R_7	high	high	C2	2
R_8	medium	medium	C3	3
R_9	high	high	C3	46
R_{10}	medium	high	C3	2

TABLE II. RULE SET AFTER ELIMINATING CONFLICTING RULES

Rule	X_1	X_2	Class	Count
R_1	high	high	C_3	46
R_2	medium	medium	C_2	45
R_3	low	low	C_1	32
R_4	low	medium	C_1	9
R_5	medium	low	C_1	8

Step-4 Remove Conflicting Rules: To optimize the generated rules, it is necessary to remove conflicting rules if there are any. Two rules are considered to be conflicting when their antecedent parts are identical while the consequent parts are not the same. The conflicting rule with the highest count is retained, and the other rule is discarded. It can be seen from Table I that Rules 4 and 5 are conflicting rules. For Rule 4 the count is 1, while for Rule 5, the count is 45.Therefore Rule 4 is eliminated. This process is repeated until there are no more conflicting rules.

Step-5 Rank Rules and Select a Subset: After eliminating repeated rules, the remaining rules are organized in descending order from the highest to lowest based on their count. A subset from the ranked rules is then selected using the count as the criterion. Rules with a low count can be excluded. In our example, we removed the rules that represent less than three percent of samples. The final rule set is shown in Table II.

III. IMPLEMENTATION AND RESULTS

In this research work we developed software to generate classification rules from training samples using MATLAB scripts. We also evaluated the extracted rules by classifying pixels in two Landsat scenes. We built FISs with extracted rules as the rule base and classified training set data. The results are provided in this section.

A. Example-1 Landsat Scene from New Orleans

As an example, we considered a Landsat-8 scene from operational Land Imager (OLI) obtained on February 26, 2016; path # 22 and row # 39. We selected an area of the size 512x512 pixels from the full scene. The raw image is shown in Fig. 4. To extract classification rules, we selected six training set areas representing three classes: water, vegetation, and land.

The training set data contained a total of 7000 samples consists of 3400, 1800, and 1800 samples from three classes: water, vegetation, and land, respectively. We used band-2, band-3, band-5, and band-6 as features for classification. We selected these bands because they showed the maximum variance. We used randomly selected ten percent of training samples for generating classification rules. Spectral signatures for the classes are shown in Fig. 5.

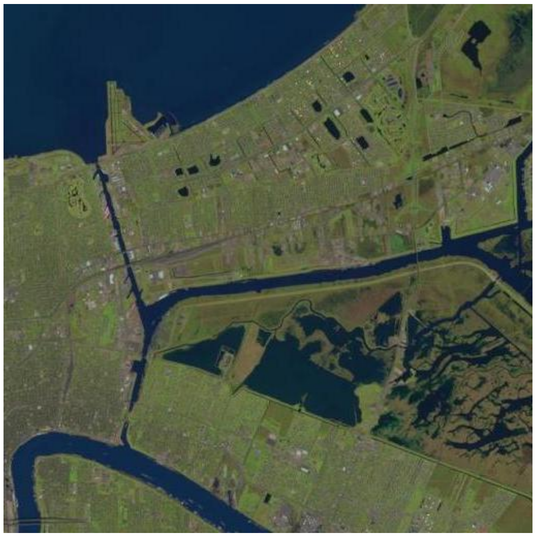


Fig. 4. Raw image-new orleans scene.

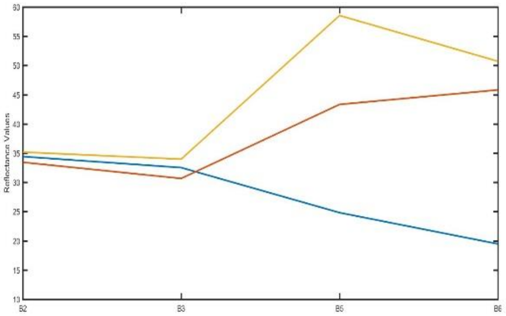


Fig. 5. Spectral signatures - new orleans scene.

TABLE III. OPTIMIZED RULE SET FOR NEW ORLEANS SCENE

	B2	B3	B5	B6	Class	Count
R_1	low	low	low	low	water	357
R_2	low	low	medium	medium	vegetation	119
R_3	low	low	high	medium	land	117
R_4	low	low	medium	high	vegetation	23
R_5	low	low	high	high	land	15
R_6	medium	medium	high	high	land	13
R_7	medium	low	high	medium	land	9
R_8	medium	medium	high	medium	land	6
R_9	low	low	low	high	vegetation	4
R_{10}	low	low	high	low	vegetation	3

TABLE IV. CONFUSION MATRIX FOR NEW ORLEANS SCENE

	water	vegetation	land
water	3055	0	0
vegetation	51	1564	0
land	0	155	1475

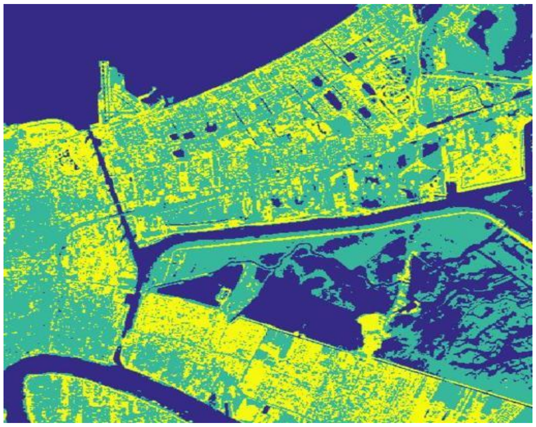


Fig. 6. Classified output new orleans scene.

We used five term sets for each feature: very-low, low, medium, high, and very-high. We used trapezoidal membership functions and generated the optimized rule set using the method outlined in Section II. The extracted optimized rule set contained sixteen rules. The first ten rules of the optimized rule set are shown in Table III. We implemented a FIS with the optimized rule set as a rule base. The process of implementing the FIS is described by Kulkarni & Shrestha [15]. The validation samples were classified using the FIS. The confusion matrix is shown in Table IV. We obtained classification accuracy of 96.73 percent with the FIS system that was built using extracted rules. The classified output is shown in Fig. 6.

B. Example-2 Landsat Scene from Alaska

In this example, we considered Landsat-8 OLI scene from Alaska obtained on June 6, 2016, path # 58 and row # 19. We considered a sub-scene of the size 512 x 512 pixels. The unclassified data for the Alaska scene is shown in Fig. 7. Spectral signatures for four classes are shown in Fig. 8.

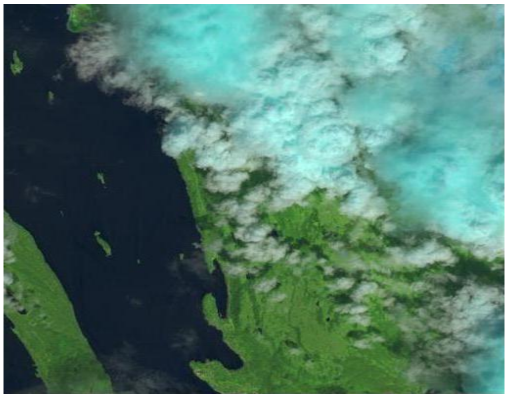


Fig. 7. Raw data-Alaska scene.

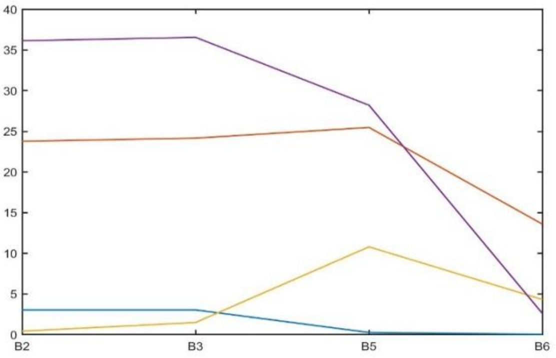


Fig. 8. Spectral signatures - Alaska scene.

To extract classification rules, we selected five training set areas representing four classes: water, vegetation, ice-land, and glaciers. Each selected training area was of the size 100x100 pixels. Our training set data consisted of 50,000 training samples. We used band-2, band-3, band-5, and band-6 as features for classification as these bands showed the maximum variance. We used randomly selected ten percent training samples for generating classification rules.

To define fuzzy membership functions, we used five term sets for each feature: very-low, low, medium, high, and very-high. We extracted fuzzy classification rules using the method described in Section II. The optimized rule set contained twenty rules. The first ten rules are shown in Table V. We implemented a FIS with the optimized rule set as a rule base, and validation samples were classified using the FIS. The confusion matrix is shown in Table VI. The obtained classification accuracy was 91.58 percent. The classified output is shown in Fig. 9.

TABLE V. OPTIMIZED RULE SET FOR ALASKA SCENE

	B2	B3	B5	B6	Class	Count
R_1	v_low	v_low	v_low	v_low	water	2068
R_2	v_low	v_low	low	medium	ice_land	851
R_3	high	high	medium	medium	glaciers	786
R_4	medium	medium	medium	low	vegetation	282
R_5	high	high	high	medium	vegetation	281
R_6	high	high	high	low	vegetation	111
R_7	medium	medium	high	low	vegetation	83
R_8	high	high	high	low	glaciers	61
R_9	medium	medium	high	medium	vegetation	56
R_{10}	medium	high	high	medium	vegetation	42

TABLE VI. CONFUSION MATRIX FOR ALASKA SCENE

	Water	Vegetation	Ice_land	Glacier
Water	16757	1223	17	0
Vegetation	10	8865	130	0
Ice_land	164	1761	7066	0
Glacier	0	136	347	8524

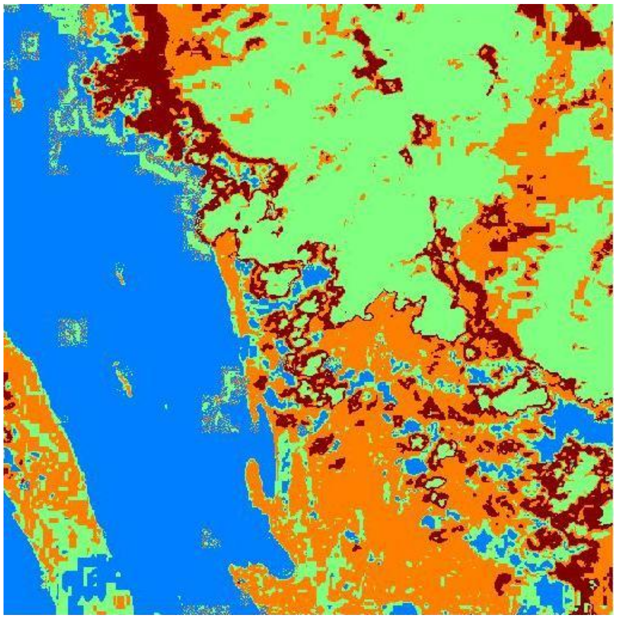


Fig. 9. Classified output - Alaska scene.

IV. DISCUSSIONS AND CONCLUSIONS

In this paper we have suggested an algorithm for generating and optimizing classification rules from training samples using fuzzy membership functions. Furthermore, we developed software using MATLAB scripts to implement the algorithm. As an illustration, we classified pixels from Landsat scenes for two areas in New Orleans and Alaska. We extracted classification rules from training samples for these two scenes. To validate extracted rules, we developed a FIS for each scene using extracted rules a rule base and classified samples from the training sets. The classification accuracy for New Orleans scene was 96.73 percent, and for Alaska, the accuracy was 91.58 percent. This clearly shows that extracting rules using fuzzy membership functions is a valid approach to generate a rule set that can be used develop a FIS for classifying pixels in Landsat images. In our examples we have used five term sets to define fuzzy membership functions. It is possible to use more terms sets to increase granularity, which may lead to an increase in the number of rules in the optimized rule set. It may be noted that as the number of rules in the optimized rule set increases the classification accuracy increases; however, there is a danger of overfitting training data.

The future work includes generating rules using fuzzy membership functions with seven or nine term sets for each membership function. This may increase the number of rules in the optimized rule set and may yield better classification accuracy. Furthermore there is no well-known criterion for evaluating quality of generated rules. That needs to be developed. We also plan a bench mark study to compare accuracy of the suggested algorithm with other existing rule extraction algorithms.

The author is thankful to anonymous reviewers for their valuable comments.

References

[1] N. Vapnik and S. Kotz, Estimation of Dependences Based on Empirical Data, NewYork: Springer-Verlag, 1982.

[2] P. Mantero, G. Moser and S. b. Serpico, "Partially supervised classification of remote sensing images through SVM-based probability density estimation," IEEE Transactions on Geoscience and Remote Sensing, vol. 43, no. 3, pp. 559-570, 2005.

[3] W. Y. Huang and R. P. Lippmann, "Neural Net and Traditional Classifiers," Neural Information Processing Systems, pp. 387-396, 1988.

[4] A. D. Kulkarni, "Neural-Fuzzy Models for Multispectral Image Analysis," Applied Intelligence, vol. 8, no. 2, pp. 179-187, 1998

[5] G. Alba, "Remote Sensing Classification. Algorithms Analysis Applied to Land Cover Change," in Master in Emergency Early Warning and Response Space Application. Mario Gulich Institute, 2014.

[6] L.X. Wang and J. M. Mendel, "Generating Fuzzy Rules by Learning from Examples," IEEE Transactions on Systems, Man, and Cybernetics, vol. 22, no. 6, pp. 1414-1427, 1992.

[7] S. L. Chiu, "An Efficient Method for Extracting Fuzzy Classification Rules from High Dimensional Data," Advanced Computational Intelligence, vol. 1, no. 1, pp. 1-7, 1997.

[8] A. Kulkarni and S. McCaslin, "Knowledge Discovery from Multispectral Satellite Images," IEEE Geoscience and Remote Sensing Letters, vol. 1, no. 4, pp. 246-250, 2004.

[9] G. Fung, S. Sandilya and R. B. Rao, "Rule Extraction from Linear Support Vector Machines," in Procs. ACM/SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, Chicago, 2005.

[10] R. S. Sicat, E. J. M. Carranza and U. B. Nidumolu, "Fuzzy Modeling of Farmers' knowledge for land suitability classification," Agricultural Systems, vol. 83, pp. 49-75, 2005.

[11] T. V. Reshmidevi, T. I. Eldho and R. Jana, "A GIS-Integrated Fuzzy Rule-Based Inference System for Land Suitability Evaluation in Agricultural Watersheds," Agricultural Systems, vol. 101, no. 1-2, pp. 101-109, 2009.

[12] T. Cay and F. Iscan, "Fuzzy Expert System for Land Reallocation in Land Consolidation," Expert Systems with Applications, vol. 38, no. 9, pp. 11055-11071, 2011.

[13] D. Meng and Z. Pei, "Extracting Linguistic Rules from Data Sets Using Fuzzy Logic and Genetic Algorithms," Neurocomputing, vol. 78, pp. 48-54, 2012.

[14] I. Khan and A. Kulkarni, "Knowledge Extraction from Survey Data Using Neural Networks," Procedia Computer Science, vol. 20, pp. 433-438, 2013.

[15] A. Kulkarni and A. Shrestha, "Multispectral Image Analysis using Decision Tress," International Journal of Advanced Computer Science and Applications, vol. 8, no. 6, pp. 11-18, 2017.

Robust Control of a 3D Space Robot with an Initial Angular Momentum based on the Nonlinear Model Predictive Control Method

Tatsuya Kai
Department of Applied Electronics
Faculty of Industrial Science and Technology
Tokyo University of Science
6-3-1 Niijuku, Katsushika-ku, Tokyo 125-8585 JAPAN

***Abstract*—This paper considers robust control problems for a 3D space robot of two rigid bodies connected by a universal joint with an initial angular momentum. It is particularly difficult to measure an initial angular momentum in parameters of space robots since the value of an initial angular momentum depends on the situations. Hence, the main purpose of this paper is to develop a robust controller with respect to initial angular momenta for the 3D space robot. First, a mathematical model, some characteristics, and two types of control problems for the 3D space robot are presented. Next, for the robust attitude stabilization control problem of the 3D space robot, a numerical simulation is performed by using the nonlinear model predictive control method. Then, for the robust trajectory tracking control problem of the 3D space robot, another numerical simulation is carried out. As a result, it turns out that this approach can realize robust control on initial angular momenta for the two control problems. In addition, computation amount is reduced by this approach and real-time control of the 3D space robot can be achieved.**

***Keywords*—3D space robot; universal joint; initial angular momentum; nonlinear model predictive control; robustness; attitude stabilization control; trajectory tracking control**

I. INTRODUCTION

Since early times, "the falling cat phenomenon" is well known and this means that a cat can land on her feet despite it drops upside down from a tree. This phenomenon has been focused and investigated in the field of classical mechanics [1]. Similar examples of the falling cat phenomenon can be easily found in the research field of engineering; robots and craft in outer space, freely falling multi-body systems [2], and gymnastic robots [3]. In such systems, the total angular momentum is conserved, and then the conservation law of the total angular momentum can be regarded as a nonholonomic constraint [4]. The word "nonholonomic" means that constraints in the differential equation form are nonintegrable. Therefore, it is possible to change the attitude of the system by changing its shape due to the existence of the nonholonomic constraint.

In the research field of control theory, various researches on space robots have been done so far. In [1], some interesting results on the falling cat phenomenon are shown. In [5], a space robot model of two rigid bodies is dealt with and a near-optimal control law using finite dimensional Fourier basis is developed. In [6], a chained-form based control strategy is proposed to control the attitude of a planer space robot. In [7], a control method by the genetic algorithm for a space robot is derived. In most researches about control of space robots, it is assumed that space robots do not have initial angular momenta. However, space robots often have initial angular momenta in realistic situations. For example, a mother ship gives a space robot out, the space robot obtains an initial angular momentum. Thus, the authors have focused on 3D space robots with initial angular momenta and developed a control strategy based on the near-optimal control method [8]–[10]. In these studies, it is confirmed that the proposed method can make the state of the 3D robot transfer to a desired one at a desired time (the state transition control problem). Moreover, in order to deal with other control purposes except the state transition control problem, The authors apply the nonlinear model predictive control method, which is one of the feedback controllers, to 3D space robots [11], [12]. In [11], [12], the two kinds of control problems, called the attitude stabilization control problem and the trajectory tracking control problem, are considered. In these work, it is assumed that initial angular momenta can be measured accurately. However, as compared to physical quantities such as mass, length, and inertia moment, it is quite difficult to measure initial angular momenta because the values of initial angular momenta vary according to the situations. So, when we measure an initial angular momentum of a 3D space robot, there exists error of measurement. To control 3D space robots with initial angular momenta, we have to overcome this problem, that is to say, we have to design robust controllers for initial angular momenta.

The main aim of this research is to construct robust controllers in terms of physical parameter errors for two types of control problems of a 3D space robot with an initial angular momentum, and verify robustness of the new control methods via numerical simulations. The contents of this paper are as follows. First, in Section II, a mathematical model of the 3D universal joint space robot with an initial angular momentum and its some characteristics are presented. Next, in Section III, a numerical simulation on robust attitude stabilization control of the 3D space robot is carried out. Then, Section IV illustrates a numerical simulation on robust trajectory tracking control of the 3D space robot. Finally, Section V gives the concluding remarks of this research and future work.

II. 3D Universal Joint Space Robot with an Initial Angular Momentum

A. Mathematical Model of 3D Space Robot with I.A.M.

First, this subsection explains a mathematical model of a 3D universal joint space robot with an initial angular momentum. See [8]–[11] for more details. We consider a space robot that consists of two rigid bodies in the 3-dimensional space as illustrated in Fig. 1. In the space robot, two rigid bodies (Rigid Body 1 and 2) are connected by a universal joint via two links (Link 1 and 2). We denote coordinates of the inertial space, Rigid Body 1 and 2 by C_0, C_1 and C_2, and assume that the origins of C_1 and C_2 correspond to the centroids of Rigid Body 1 and 2, respectively.

Let $A_i \in SO(3)$ be the attitude of Rigid Body i $(i = 1, 2)$ with respective to the inertial space C_0, and $w_i \in \mathbf{R}^3$ be the angular velocity of Rigid Body i. Note that $\hat{w}_i = A_i^\top \dot{A}_i$ holds. We use the following notations: m_i: the mass of Rigid Body i ($\epsilon = m_1 m_2/(m_1 + m_2)$), l_i: the length of Link i, $s_i = [\, 0\ \ 0\ \ -l_i \,]^\top \in \mathbf{R}^3$: the vector showing the position of the joint w.r.t. C_0, $I_i \in \mathbf{R}^3$: The inertia tensor of Rigid Body i ($J_i = I_i + \epsilon \hat{s}_i^\top \hat{s}_i$, $J_{12} = \epsilon \hat{s}_1^\top A_1^\top A_2 \hat{s}_2$), where ˆ is the operator that transforms a 3-dimensional vector $v = [\, v_1\ v_2\ v_3 \,]^\top \in \mathbf{R}^3$ into a 3×3 skew-symmetric matrix:

$$\hat{v} = \begin{bmatrix} 0 & -v_3 & v_2 \\ v_3 & 0 & -v_1 \\ -v_2 & v_1 & 0 \end{bmatrix}. \tag{1}$$

It is also noted that $A := A_1^\top A_2$ represents the shape of the space robot and

$$w_2 = A^\top w_1 + w \tag{2}$$

holds for the angular velocity of the joint $w \in \mathbf{R}^3$, $\hat{w} = A^\top \dot{A}$. In this paper, we adopt the universal joint depicted in Fig. 2 as a joint connecting the two rigid bodies. It must be noted the universal joint can twist and the degree of freedom is 2. Let $\theta_1 \in \mathbf{R}$ and $\theta_2 \in \mathbf{R}$ be angles of Link 1 and 2, respectively, and we use the notation: $\theta = [\, \theta_1\ \theta_2 \,]^\top \in \mathbf{R}^2$. By considering coordinates of the space robot, we can show the following:

$$A = \begin{bmatrix} \sin\theta_1 \sin\theta_2 & \cos\theta_1 & -\sin\theta_1 \cos\theta_2 \\ \cos\theta_2 & 0 & \sin\theta_2 \\ \cos\theta_1 \sin\theta_2 & -\sin\theta_1 & -\cos\theta_1 \cos\theta_2 \end{bmatrix}. \tag{3}$$

In this paper, we consider the case where the universal model has an initial angular momentum, hence we denote the initial angular momentum by $P_0 \in \mathbf{R}^3$. Then, the conservation law of total angular momentum of the space robot is represented by

$$(A_1 J_1 + A_2 J_{12}^\top) w_1 + (A_2 J_2 + A_1 J_{12}) w_2 = P_0, \tag{4}$$

and we can easily confirm that (4) is represented as $A(q) + B(q)\dot{q} = 0$ using the generalized coordinate q, and thus this is an affine constraint [14]. From the result in [14], it can be checked that (4) is completely nonholonomic. Assume that angular velocities of the universal joint can be controlled, that is, $u_1 := \dot{\theta}_1$, $u_2 := \dot{\theta}_2$. Then, we have

$$w = \underbrace{\begin{bmatrix} \cos\theta_2 \\ 0 \\ \sin\theta_2 \end{bmatrix}}_{b_1} u_1 + \underbrace{\begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}}_{b_2} u_2. \tag{5}$$

Substituting (2) and (5) into (4), we obtain

$$A_1 I_u w_1 + A_1 (A J_2 + J_{12})(b_1 u_1 + b_2 u_2) = P_0, \tag{6}$$

where we define the new notation:

$$I_u := J_1 + A J_2 A^\top + A J_{12}^\top + J_{12} A^\top. \tag{7}$$

To represent the attitudes of Rigid Body 1, we use the Cayley-Rodrigues parameter, and hence the attitude of Rigid Body 1 A_1 is expressed as (8). using the parameter $\alpha = [\, \alpha_1\ \ \alpha_2\ \ \alpha_3 \,]^\top \in \mathbf{R}^3$. The relationship between w_1 and α is expressed by

$$w_1 = U_1(\alpha)\dot{\alpha},\ U_1(\alpha) = \frac{2(I - \hat{\alpha})}{1 + \alpha^\top \alpha}. \tag{9}$$

Substituting (9) into (6) and solving for $\dot{\alpha}$, we have

$$\begin{aligned} \dot{\alpha} = {} & U_1^{-1} I_u^{-1} A_1^\top P_0 \\ & - U_1^{-1} I_u^{-1} (A J_2 + J_{12})(b_1 u_1 + b_2 u_2). \end{aligned} \tag{10}$$

Consequently, we derive *the universal joint model with an initial angular momentum* (11) with the variables $q := [\theta^\top\ \alpha^\top]^\top \in \mathbf{R}^5$, $u := [\, u_1\ u_2 \,]^\top \in \mathbf{R}^2$. Thus, the system (11) is represented as a 5-state and 2-input nonlinear affine control system.

B. Control Problems for 3D Space Robot with I.A.M.

Next, some characteristics of the universal joint model with an initial angular momentum (11) are investigated from the viewpoint of nonlinear control theory. If an initial angular momentum exists, that is, $P_0 \neq 0$, the drift term of the system (11) satisfies $f(q) \neq 0,\ \ \forall q$. This means that the system (11) does not have any equilibrium point, that is, the space robot cannot stop and keeps moving. For nonlinear control systems, the concepts “local accessibility” and “local controllability” are quite important in order to investigate the range of movement of the system [13], [14]. The next proposition on local accessibility and local controllability of the system (11) can be obtained [8]–[11].

Proposition 1: The universal joint model with an an initial angular momentum (11) is locally strongly accessible at any point $q = [\, \theta^\top\ \alpha^\top \,]^\top \in \mathbf{R}^5$. Moreover, if the control input u is sufficiently large, (11) is small-time locally controllable at any point q. □

Since there is no equilibrium point in the model of the universal joint model with an initial angular momentum (11), we cannot deal with a normal stabilization problem for the system. However, Proposition 1 guarantees that we have some possibilities of other control purposes except normal stabilization. For the universal joint model with an initial angular momentum (11), the next three types of control purposes

$$A_1(\alpha) = \frac{1}{1+||\alpha||^2}\left[\begin{array}{ccc} 1+\alpha_1^2-\alpha_2^2-\alpha_3^2 & 2(\alpha_1\alpha_2-\alpha_3) & 2(\alpha_1\alpha_3+\alpha_2) \\ 2(\alpha_1\alpha_2+\alpha_3) & 1-\alpha_1^2+\alpha_2^2-\alpha_3^2 & 2(\alpha_2\alpha_3-\alpha_1) \\ 2(\alpha_1\alpha_3-\alpha_2) & 2(\alpha_2\alpha_3+\alpha_1) & 1-\alpha_1^2-\alpha_2^2+\alpha_3^2 \end{array}\right] \tag{8}$$

$$\left[\begin{array}{c} \dot{\theta}_1 \\ \dot{\theta}_2 \\ \dot{\alpha} \end{array}\right] = \underbrace{\left[\begin{array}{c} 0 \\ 0 \\ U_1^{-1}I_u^{-1}A_1^{\mathsf{T}}P_0 \end{array}\right]}_{f(q)} + \underbrace{\left[\begin{array}{cc} 1 & 0 \\ 0 & 1 \\ -U_1^{-1}I_u^{-1}(AJ_2+J_{12})b_1 & -U_1^{-1}I_u^{-1}(AJ_2+J_{12})b_2 \end{array}\right]}_{g(q)} \left[\begin{array}{c} u_1 \\ u_2 \end{array}\right] \tag{11}$$

can be considered: (i) the state transition control problem: a controller to transfer of the space robot to a desired state at a desired time; (ii) the attitude stabilization control problem: a controller to stabilize only the attitude of the space robot with ignoring the shape of it; (iii) the trajectory tracking control problem: a controller to track the space robot to a given reference trajectory. In this paper, we will tackle the problems (iii): the trajectory tracking control problem.

III. Robust Attitude Stabilization Control

A. Problem Setting

In this section, we shall deal with "the robust attitude stabilization control problem" for the universal joint model with an initial angular momentum (11). The purpose of this control problem is that we design a controller to stabilize only the attitude of the space robot α with ignoring the shape θ in the presence of a parameter error for the initial angular momentum. For example, this kind of the control problem includes the situation where we move the space robot to the direction of a given point of the earth in order to send and receive information. The robust attitude stabilization control problem is formulated as the next.

Problem 1 [Robust Attitude Stabilization Control]: For the universal joint model with an initial angular momentum (11), find control inputs such that the attitude of Rigid Body 1 α is stabilized to a desired value α_d under the assumption on the existence of a parameter error for the initial angular momentum ($\hat{P}_0$: an estimated value, P_0 : a real value). □

In this paper, we take the nonlinear model predictive control approach in order to solve Problem 1. Especially, we use *the C/GMRES method* [15], which is a real-time optimal control algorithm for nonlinear systems. In the simulation, we use the parameters of the 3D space robot: $l_1 = l_2 = 1, m_1 = m_2 = 1$, $I_1 = I_2 = \mathrm{diag}\{\,1/2,\ 1/2,\ 1\,\}$, the estimated value of the initial angular momentum: $\hat{P}_0 = [0.10.1{-}0.1]^{\mathsf{T}}$, the real value) of the initial angular momentum: $P_0 = [\,0.07\ 0.07\ \ -0.01\,]^{\mathsf{T}}$, the initial state: $q_0 = [\,\pi/2\ \pi/2\ 1\ 1\ 1\,]^{\mathsf{T}}$, the desired attitude: $\alpha_d = [\,0\ \ 0\ \ 0\,]^{\mathsf{T}}$. For the nonlinear model predictive control method, we use the next cost function:

$$\begin{aligned} J =& \frac{1}{2}\int_t^{t+T(t)} (\alpha(\tau)-\alpha_d)^{\mathsf{T}} Q(\alpha(\tau)-\alpha_d)d\tau \\ &+ \frac{1}{2}\int_t^{t+T(t)} u(\tau)^{\mathsf{T}} R u(\tau) d\tau \\ &+ \frac{1}{2}(\alpha(t+T)-\alpha_d)^{\mathsf{T}} S(\alpha(t+T)-\alpha_d) \end{aligned} \tag{12}$$

with the weight matrices:

$$\begin{aligned} Q &= \mathrm{diag}\{2.0,\ 1.0,\ 3.0\}, \\ R &= \mathrm{diag}\{0.01,\ 0.01\}, \\ S &= \mathrm{diag}\{0.8,\ 0.2,\ 0.4\}. \end{aligned} \tag{13}$$

Note that the cost function (12) evaluates only α and ignores θ since we consider attitude stabilization. In (12), the evaluation interval is set as $T(t) = T(1-e^{-at})$, $T = 6.5$, $a = 0.05$. Moreover, we also use the parameters of controller: the division number of the evaluation interval: $N = 50$, the stabilization parameter of the continuation method: $\zeta = 20$, the number of iterations of the GMRES method: $k_{max} = 3$, the sampling time: $\Delta t = 0.05\,[\mathrm{s}]$, the simulation time: $20\,[\mathrm{s}]$.

B. Simulation Results

A numerical simulation is performed based on the problem setting in subsection III-A. Fig. 2 and 3 illustrate the simulation results. Fig. 2 shows the time series of θ and α of the 3D space robot, and Fig. 3 depicts snapshots of the behavior of the 3D space robot. From these results, we can confirm that the attitude of Rigid Body 1 α is stabilized to the desired value $\alpha_d = [\,0\ 0\ 0\,]^{\mathsf{T}}$. The computation time of this simulation is $0.95\,\mathrm{s}$, and hence we can see that real-time robust control can be achieved by the proposed method. It is also confirmed that the control purposes can be achieved for other problem settings (physical parameters of the 3D space robot, initial and desired states, and an initial angular momentum) with tuning the weight matrices in (12).

IV. Robust Trajectory Tracking Control

A. Problem Setting

Next, in this section, we shall deal with another control purpose, called "the robust trajectory tracking control problem" for the universal joint model with an initial angular momentum (11). The purpose of this control problem is that we design a controller to make the state of space robot q track to a desired trajectory data $q_d(t)$ in the presence of a parameter error for the initial angular momentum. Examples of the control problem are the situations where we track a solar panel installed into the space robot to the direction of the sun, and we shoot moving astronomical bodies with a camera installed into the space robot. The robust trajectory tracking control problem is defined as follows.

Problem 2 [Robust Trajectory Tracking Control] : For the universal joint model with an initial angular momentum (11), find control inputs such that the state q tracks to a desired trajectory $q_d(t)$ under the assumption on the existence of a parameter error for the initial angular momentum ($\hat{P}_0$: an estimated value, P_0: a real value). □

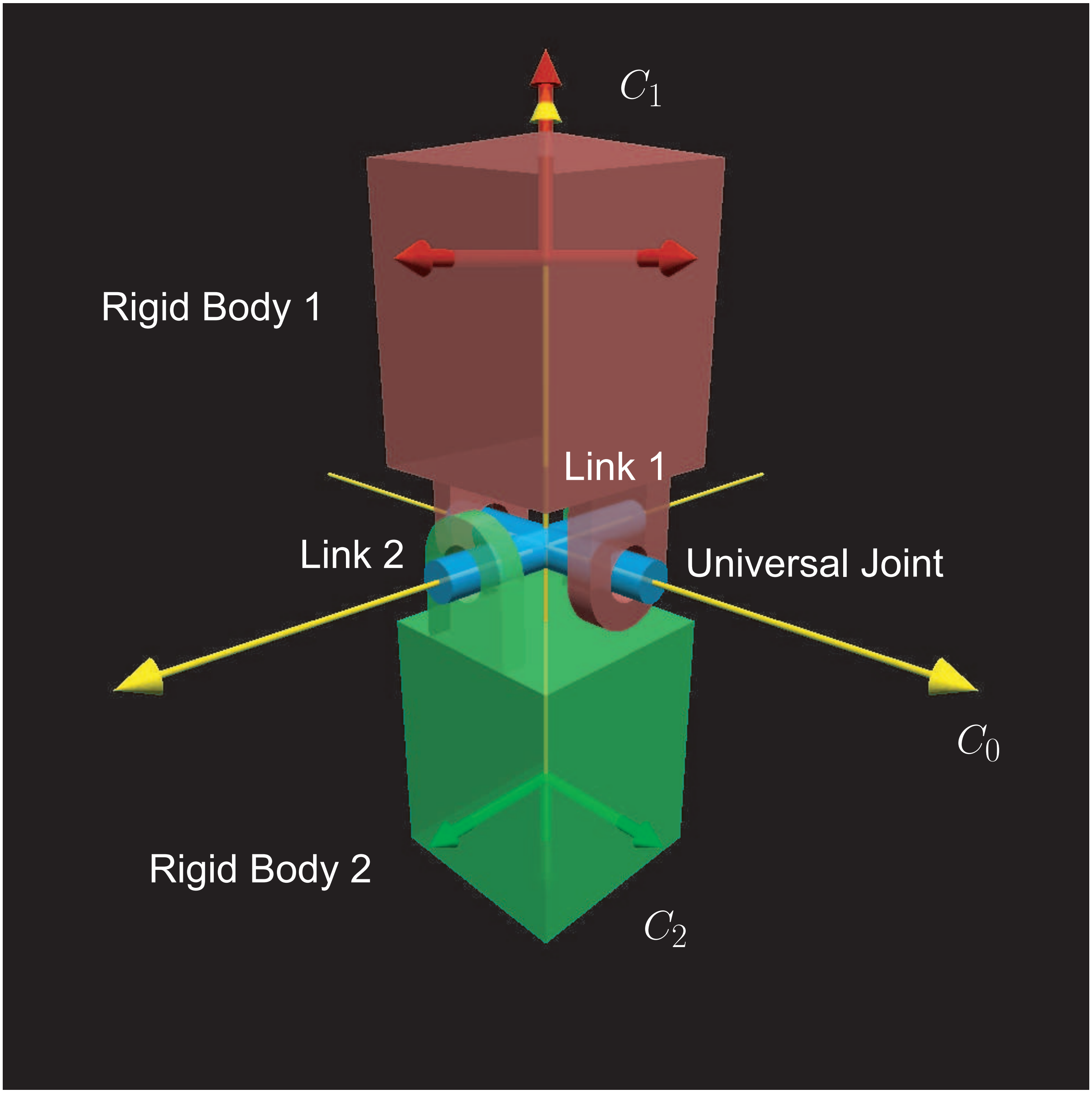


Fig. 1. The 3D universal joint space robot model.

To solve Problem 2, we also utilize the C/GMRES method [15]. In a simulation, we use the parameters of the 3D space robot: $l_1 = l_2 = 1$, $m_1 = m_2 = 1$, $I_1 = I_2 = \mathrm{diag}\{\,1/2,\ 1/2,\ 1\,\}$, the estimated value of the initial angular momentum: $\hat{P}_0 = [\,0.1\ \ -0.1\ \ 0.2\,]^\mathsf{T}$, the real value) of the initial angular momentum: $P_0 = [\,0.095\ \ -0.095\ \ 0.195\,]^\mathsf{T}$, the initial state: $q_0 = [\,-\pi/2\ \ -\pi/2\ \ -1\ \ -1\ \ -1\,]^\mathsf{T}$. The desired trajectory $q(t)$ is generated from (11) in advance (see the simulation result in the next subsection). For the C/GMRES method, we use the next cost function:

$$\begin{aligned} J =& \frac{1}{2}\int_t^{t+T(t)} (q(\tau) - q_d(\tau))^\mathsf{T} Q (q(\tau) - q_d(\tau)) d\tau \\ &+ \frac{1}{2}\int_t^{t+T(t)} u(\tau)^\mathsf{T} R u(\tau) d\tau \\ &+ \frac{1}{2}(q(t+T) - q_d(t))^\mathsf{T} S (q(t+T) - q_d(t)), \end{aligned} \tag{14}$$

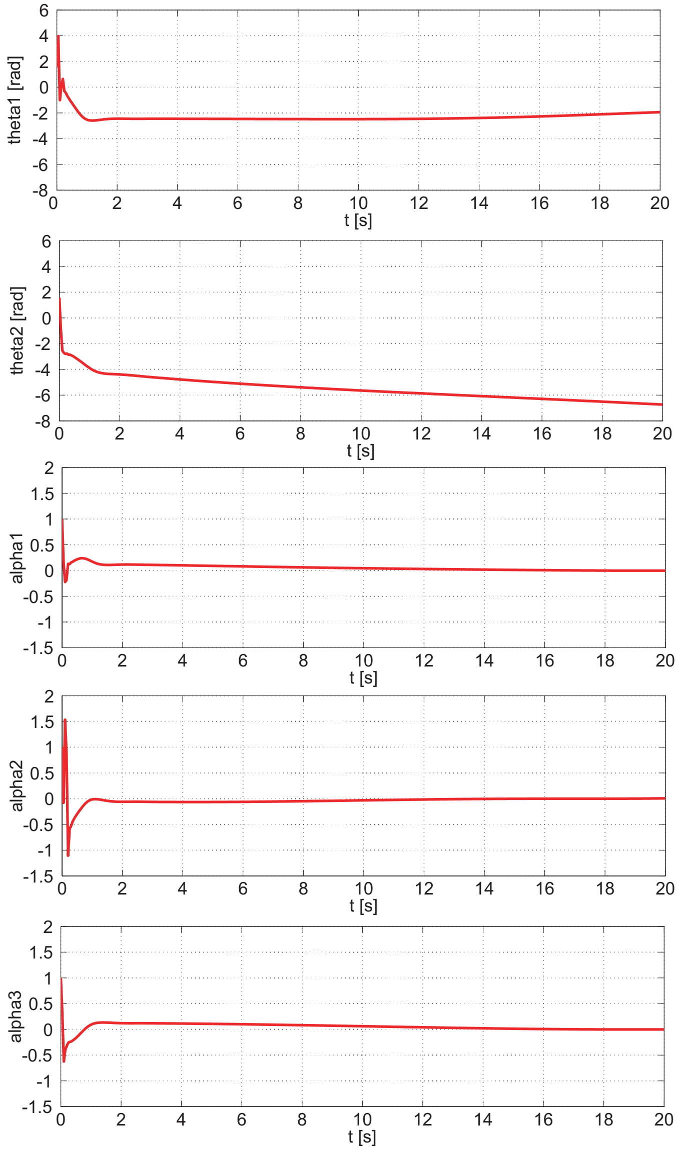


Fig. 2. The simulation result in the robust attitude stabilization control problem: the time histories of the real trajectory.

where

$$\begin{aligned} Q &= \mathrm{diag}\{0.02,\ 0.015,\ 0.04,\ 0.04,\ 0.04\}, \\ R &= \mathrm{diag}\{0.01,\ 0.01\}, \\ S &= \mathrm{diag}\{0.1,\ 0.1,\ 0.2,\ 0.2,\ 0.2\}, \end{aligned} \tag{15}$$

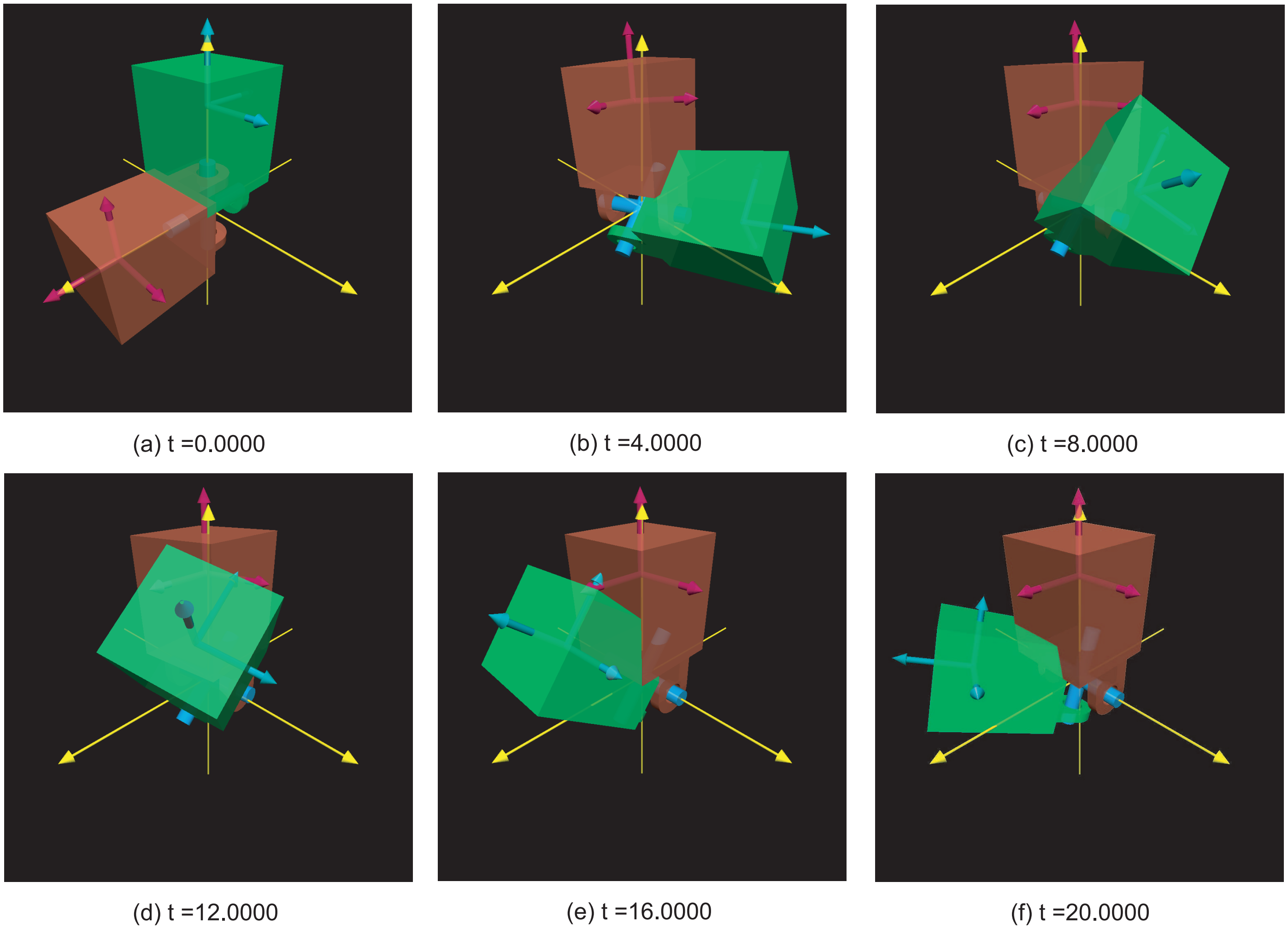
and the evaluation interval $T(t) = T(1 - e^{-at})$, $T = 6.5$, $a = 0.05$. In addition, we also use the parameters of controller: the division number of the evaluation interval: $N = 50$, the stabilization parameter of the continuation method: $\zeta = 200$, the number of iterations of the GMRES method: $k_{max} = 3$, the sampling time: $\Delta t = 0.005\,[\mathrm{s}]$, the simulation time: $40\,[\mathrm{s}]$.

B. Simulation Results

The simulation results are depicted in Fig. 4, 5, and 6. Fig. 4 shows both the time series of θ, α of the space robot and the desired trajectory. In Fig. 5, the time series of the error defined by $e(t) := q(t) - q_d(t)$ is illustrated. In addition, snapshots of the behavior of the 3D space robot for both desired trajectory and the real one are shown in Fig. 6. From these results, we can see that the state of the space robot tracks to the desired trajectory $q_d(t)$ and then the error $e(t)$ converges to 0. The computation time of this simulation is $2.71\,\mathrm{s}$, and hence it turns out that real-time robust control can be achieved by the proposed method. We can also confirm that the control purposes can be also achieved for other problem settings (physical parameters of the 3D space robot, initial and desired states, an initial angular momentum, and a desired trajectory) with tuning the weight matrices in (12).

V. Conclusions

This paper has considered two kinds of robust control problems “the robust attitude stabilization control problem” and “the robust trajectory tracking control problem” for the 3D universal joint space robot with an initial angular momentum



(a) t =0.0000

(b) t =4.0000

(c) t =8.0000

(d) t =12.0000

(e) t =16.0000

(f) t =20.0000

Fig. 3. The simulation result in the robust attitude stabilization control problem: the snapshots of the 3D space robot.

from the standpoint of the nonlinear model predictive control approach. The simulation results have shown that the control purposes are achieved in both robust control problems, and hence the proposed method has robustness for initial angular momenta. Moreover, we can see that the computation times in numerical simulations are quite short, and thus real-time control has been realized.

Our future work include the next topics: modeling and control of space robots with initial angular momenta by the quaternion representation, controller design for space robot models in the descriptor representation, and control of other types of 3D space robots.

References

[1] M. J. Enos, ed., "*Dynamics and control of mechanical systems: the falling cat and related problems*", Fields Institute Communications, 1, American Mathematical Society, 1993

[2] T. Mita, S. Hyon, and T. Nam, "Analytical Time Optimal Control Solution for Free Flying Objects with Drift Terms", *in Proc. of IEEE CDC 2000*, Sydney, Australia, pp. 91–94, 2000

[3] T. Mita, S. Hyon and T. Nam, "Analytical Time Optimal Control Solution for a Two Link Planar Acrobot with Initial Angular Momentum", *IEEE Trans. Robotics and Automation*, Vol. 17, No. 3, pp. 361–366 2001

[4] A. M. Bloch, *Nonholonomic Mechanics and Control*, Springer, 2003

[5] C. Fernandes, L. Gurvits, and Z. Li, "Near-Optimal Nonholonomic Motion Planning for a System of Coupled Rigid Bodies", *IEEE Trans. on Automatic Control*, Vol. 39, No. 3, pp. 450–463, 1994

[6] F. Matsuno and J. Tsurusaki, "Chained Form Transformation Algorithm for a Class of 3-States and 2-Inputs Nonholonomic Systems and Attitude Control of a Space Robot", *in Proc. of IEEE CDC 1999*, Arizona, USA, pp. 2126–2131, 1999

[7] X. Ge and L. Chen, "Attitude Control of a Rigid Spacecraft with Two Momentum Wheel Actuators Uging Genetic Algorithm", *Acta Astronautica*, Vol. 55, No. 1, pp. 3–8, 2004

[8] T. Kai and K. Tamaki, "3D Attitude Near-Optimal Control of a Universal Joint Space Robot Model with Initial Angular Momentum," *in Proc. of ICCA 2009*, Christchurch, New Zealand, pp.2335–2340, 2009

[9] T. Kai and K. Tamaki, "Control of a 3D Ball-in-Socket Joint Space Robot Model with Initial Angular Momentum based on the Near-Optimal Control Algorithm," *in Proc. of NOLCOS 2010*, Bologna, Italy, pp. 957–962, 2011

[10] T. Kai and K. Tamaki, "A Near-Optimal Control Approach to 3D Ball-in-Socket Joint Space Robot Models with Initial Angular Momenta," *Acta Astronautica*, Vol. 68, pp. 1702–1711, 2011

[11] T. Kai, "A Model Predictive Control Approach to Attitude Stabilization and Trajectory Tracking Control of a 3D Universal Joint Space Robot with an Initial Angular Momentum," *in Proc. of IEEE CDC 2011*, Orlando, USA, pp. 3547–3552, 2011

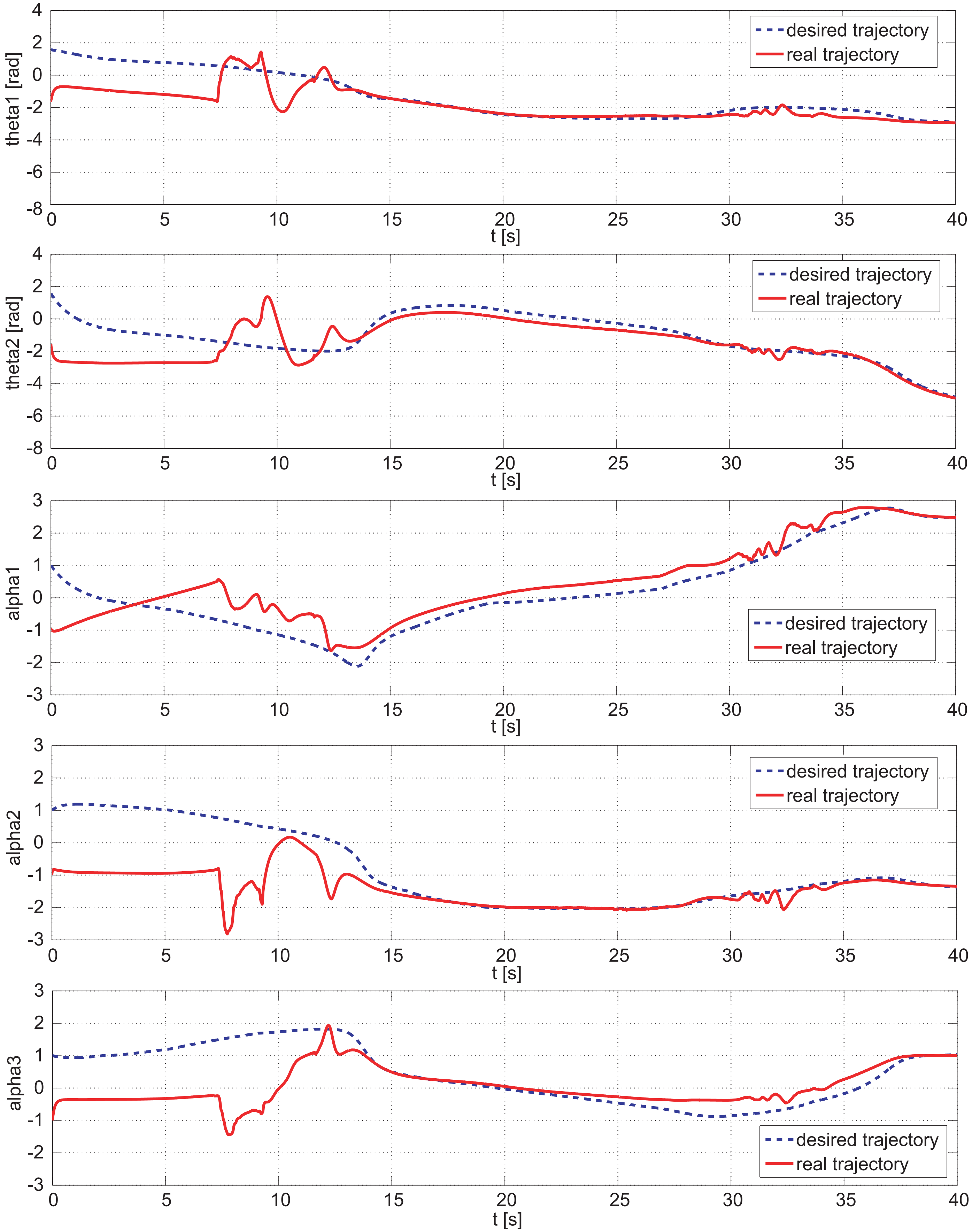


Fig. 4. The simulation result in the robust trajectory tracking control problem: the time histories of the desired and real trajectories.

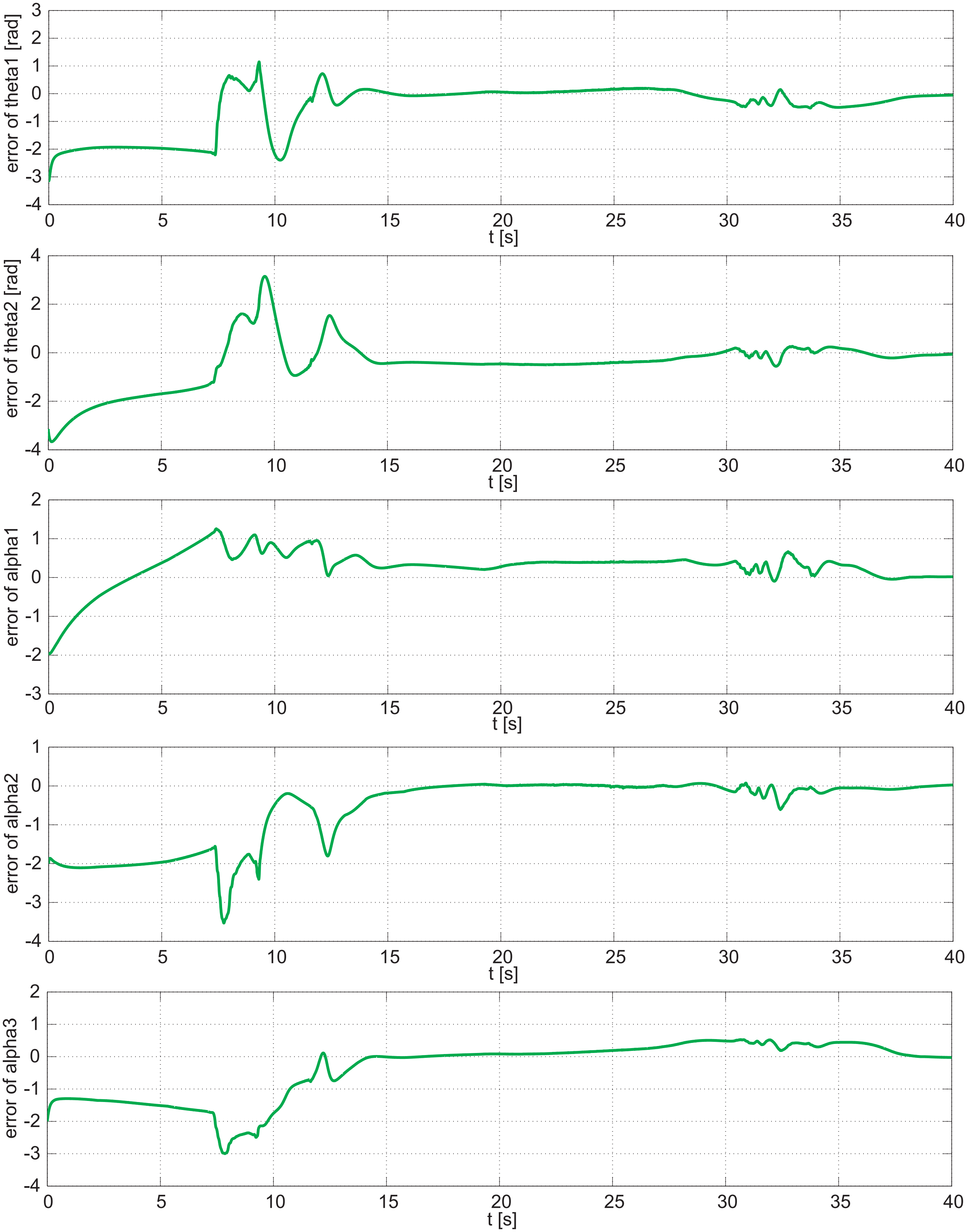
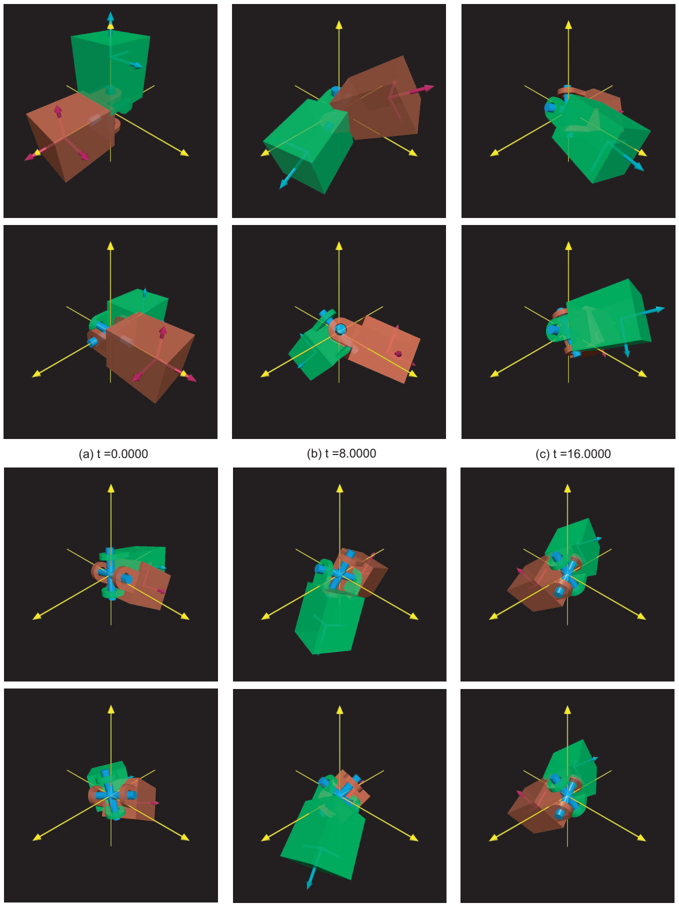


Fig. 5. The simulation result in the robust trajectory tracking control problem: the time histories of the error.



(a) t =0.0000

(b) t =8.0000

(c) t =16.0000

(d) t =24.0000

(e) t =32.0000

(f) t =40.0000

Fig. 6. The simulation result in the robust trajectory tracking control problem: the snapshot of the 3D space robot (upper: the desired trajectory, lower: the real trajectory).

[12] T. Kai, "Real-time Control of a 3D Space Robot with an Initial Angular Momentum", *International Journal of Engineering Research and Industrial Applications*, Vol. 7, No. 3, pp. 137–151, 2014

[13] J. M. Godhavn, A. Balluchi, L. S. Crawford, and S. S. Sastry, "Steering of a Class of Nonholonomic Systems with Drift Terms", *Automatica*, Vol. 35, No. 5, pp. 837–847, 1999

[14] T. Kai, H. Kimura, and S. Hara, "Nonlinear Control Analysis on Kinematically Asymmetrically Affine Control Systems with Nonholonomic Affine Constraints", *Proc. of 16th IFAC World Congress*, Seville, Spain, Paper No. Mo-M08-TO/5, 2005

[15] T. Ohtsuka, "A Continuation/GMRES Method for Fast Computation of Nonlinear Receding Horizon Control", *Automatica*, Vol. 40, No. 4, pp. 563–574, 2004

Link Prediction Schemes Contra Weisfeiler-Leman Models

Katie Brodhead
Department of Mathematics
Florida A&M University
Tallahassee, United States

***Abstract*—Link prediction is of particular interest to the data mining and machine learning communities. Until recently all approaches to the problem used embedding-based methods which leverage either node similarities or latent group memberships towards link prediction. Chen and Zhang recently developed a class of non-embedding approaches called Weisfeiler-Leman (WL) Models. WL-Models extract subgraphs around links and then encode subgraph patterns via adjacency matrices using the so-called Palette-WL algorithm. A training stage then learns nonlinear graph topological features for link prediction. Chen and Zhang compared two WL-Models – a linear regression model ("WLLR") and a neural networks model ("WLNM") – against 12 different common link prediction schemes. In this paper, all author claims are validated for WLLR. Additionally, WLLR is tested against 22 additional embedding-based link prediction techniques arising from common neighbor-, path- and random walk-based schemes. WLLR is shown not to be superior when calculable. In fact, in 80% of the datasets where comparisons were possible, one of our added implementations proved superior.**

Keywords—Weisfeiler-Leman; link prediction; machine learning; linear regression; common walk; path-based; random walk; stochastic block; matrix factorization

I. Introduction

Improvement in effective link prediction has been of particular interest to the data mining and machine learning communities. Much interest arises from diverse real-world applications. Particular applications include friend recommendation in social networks [2], product recommendation in e-commerce [3], knowledge graph completion [4], finding interactions between proteins [5], and recovering missing reactions in metabolic networks [6]. Additional interests arise from the search to overcoming a central challenge for researchers: determining which method is best for a particular situation, especially when each scheme is grounded in a particular heuristic.

Heuristics range in complexity from the more complex (e.g. stochastic block models [5], probabilistic matrix factorization [7]) to the more simplistic (e.g. common neighbors (CN) [1], Katz index [9]). Heuristics with mid-level complexities include methods which calculate node proximity scores via network topologies or random walks. Amongst the diverse methods which exist, the following two challenges have always persisted.

1) Heuristic complexity does not often translate into corresponding performance. The more simplistic often work well, are more interpretable, and scalable. The Katz and CN indices are exemplary examples. The latter asserts higher link probability as the number of common neighbors increases and is reasonably accurate with respect to links on social networks.

2) All known heuristics lack universal applicability to different kinds of networks. CN is again a prime example: its performance electrical grids and biological networks is quite poor [10] notwithstanding its excellent aforementioned successes. Resistance distance (RN) is a converse example: it performs poor where CN thrives [11]. A study of over 20 different heuristics found flaws in each, making none universally effective performance models [10].

Hitherto, the only resolutions to (1) and (2) have been expert selection or trial-and-error.

A recent KDD paper [40] modifies the Weisfeiler-Leman (WL) algorithm from graph theory towards making link predictions. The modified algorithm is called Palette-WL. Additional algorithmic additional machinery is then built on top which allow for machine learning implements to operate. The authors claim an establishment of new universal model which learns a suitable heuristic directly from a given network, thereby demolishing challenge 2. In addition, reported results demonstrate a superior performance over a wide variety of known link prediction methods, thereby ensuring the demolishment of challenge 1.

In this paper, we implement Palette-WL in MATLAB and train a linear regression model (i.e., the authors' WLLR model) towards validating author claims which the authors test on 12 common link prediction schemes. We also expand testing scope and implement 22 additional tests towards developing a more complete picture of author claims. All 34 aforementioned link prediction schemes are "lean" – that is, they do not require a neural network or advanced support for parallelism or distributed computing. The goal of this work, then, is to test author claims on lean prediction schemes contra WLLR. To that end, we test five of the authors' lean data sets (USAir, NS, PB, Yeast, C.ele). These are described in Section II with results presented in Table V. Three of the authors data sets (Power, Router, and E.coli) are not tested in this paper as these are not "lean". Our future work will perform the same type of analysis on the full WLNM model, and additional non-lean data sets. See Section VI on Future Work.

This paper is organized as follows. In Section II, a high-level overview of link prediction is presented. In Section III, we present Palette-WL and the implementation of WL-Models. Section IV details the many specific link prediction models implemented in this paper along with the key results that were obtained. Section V gives a conclusion, while Section VI presents directions for future work. The tail end of this work, following the References Section, includes an Appendix where the full set of computation results from this paper is presented in various tables.

II. Overview of Link Prediction

Historically, link prediction models have been feature-based (a.k.a. embedding-based) arising either from (1) topological features or (2) latent features.

1) Topological feature models. These models leverage node similarities, either locally or globally. Topological models do not perform well when similarity scores do not capture the network formation mechanisms. Common neighbor-based methods (e.g. CN [1], Adamic-Adar [2]), Path-based methods (e.g., Katz [9]), and random-walk based methods (e.g. PageRank [1]) all fall within this category. A breakdown of each of these categories is given in Section IV.

2) Latent feature Models. These models assume that latent groups exist for nodes and that links are determined by group memberships. Latent models extract group memberships via the low-rank decomposition of a network adjacency matrix [3] or via training which fits probabilistic models [5]. Given these models' focus on individual nodes, a central weakness arises in understanding how networks are formed. Popular methods include ranking methods [17], learning to rank methods [17], matrix factorization [16], and stochastic block methods [5], [18]. This paper implements methods from the latter two.

Weisfeiler-Leman models for link prediction are not feature-based. The ideas which motivated its implementation arose from two research areas related to graph classification: design of efficient graph kernels [14], [19], and effective graph labeling schemes [15] arising from impositions of vertex orderings. Niepert et al. [15], in particular, focus on orderings towards defining receptive fields around node pixels; the fields are then used to learn a convolutional neural network for graph classification. WL-models [40] work by instead extracting subgraphs around links instead of node pixels, and by focusing on link prediction rather than on graph classification.

WL-models are new to the link prediction landscape being only formally published in the recent Chen-Zhang paper [40] which this paper tests. Chen and Zhang, in particular, implement two key WL-Models, WLLR and WLNM, along with a third, Palette-WLNM, an extension of WLNM. Among the two, WLNM is superior. Area Under the receiver operating characteristic Curve (AUC) Results are listed in Table I, split in two parts, for five datasets and twelve non-WL methods. In short, WLNM outperforms nine state-of-the-art link-prediction methods developed by heuristic means (e.g. Katz, PageRank, SimRank, etc.), and three latent feature models (stochastic model block, and two matrix factorization methods); a full explanation of all methods will be given in Section IV. \WLLR is less successful, but nonetheless a strong adversarial method. Palette-WLNM is tested elsewhere in their paper, but not considered in this present paper given this paper's focus on WLLR.

The five datasets used above are USAir, NS, PB, Yeast, and C.ele. USAir is a network of US airlines. NS is a collaboration network of researchers who publish papers on network science. PB is a network of US political blogs. Yeast is a protein-protein interaction network in yeast. C.ele is a neural network of C. elegans. All evaluation methods (CN, Jac, AA, etc., will be described in full detail in Section IV.

TABLE I. Results from [40]

Data	CN	Jac	AA	RA	PA	Katz	RD	PR	SR	SBM	MF-c	MF-r	WLLR[10]	WLNM[10]
USAir	0.940	0.903	0.950	0.956	0.894	0.931	0.898	0.944	0.782	0.944	0.918	0.849	0.896	0.958
NS	0.938	0.938	0.938	0.938	0.682	0.940	0.582	0.940	0.940	0.920	0.636	0.720	0.862	0.984
PB	0.919	0.873	0.922	0.923	0.901	0.928	0.883	0.935	0.773	0.938	0.930	0.943	0.827	0.933
Yeast	0.891	0.890	0.891	0.892	0.824	0.921	0.880	0.927	0.914	0.914	0.831	0.881	0.854	0.956
C.ele	0.848	0.792	0.864	0.868	0.755	0.864	0.740	0.901	0.760	0.867	0.832	0.844	0.803	0.859
Rank	7.875	10.625	7.500	6.875	12.875	7.125	10.375	5.125	11.000	5.625	10.500	9.500	10.125	2.500

IV. Palette-WL and WL-Models

WL-models use a modified version of the WL-algorithm, called Palette-WL, from graph theory towards making link predictions. Additional algorithmic additional machinery is then built on top which allow for machine learning implements to operate. Three steps, in particular, flesh out an entire WL-Model.

1) Extract enclosing subgraphs: generates K-vertex neighboring subgraphs.

2) Encode subgraph patterns: via adjacency matrices with vertex ordering given by Palette-WL.

3) Training: learns nonlinear graph topological features for link prediction.

To understand each step, a review of graph labeling functions, along with the base WL-Algorithm is in order. A graph labeling function is a map $L: V \rightarrow C$ from vertices V to an ordered set of colors $C = \{1, ..., n\}$. C uniquely determines the vertex order in an adjacency matrix whenever L is one-to-one. The WL-algorithm ("WL"), then, is a color refinement algorithm which iteratively updates vertex colors on a particular graph labeling function, specified below, until a fixed point is reached. (Palette-WL, discussed later, further ensures that the converged function is one-to-one.)

WL specifically works by iteratively augmenting vertex labels using neighbors' labels. It then compresses augmented labels into new "signature" labels until convergence. At first, all vertices are set to the same color "1". Each vertex gets its new signature string by concatenating its own color and the sorted colors of its immediate neighbors. Vertices are then sorted by the ascending order of their signature strings and assigned new colors 1, 2, 3. Vertices with the same signature strings get the same color. WL is formally presented below. See Fig. 1 for an example of the process.

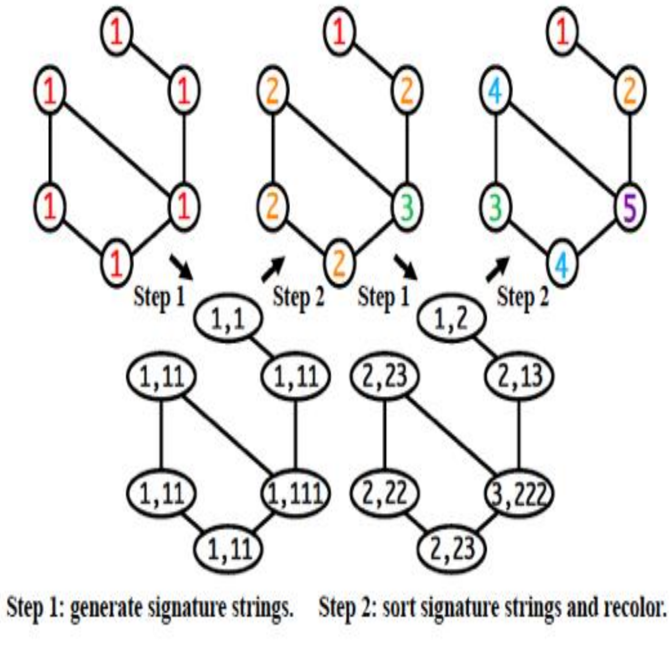


Fig. 1. An example of WL.

WL-Algorithm

```
input: graph $G = (V, E)$, initial colors $c^0(v) = 1, \forall v \in V$
output: final colors $c(v)$ for all $v \in V$
let $c(v) = c^0(v)$ for all $v \in V$
while $c(v)$ has not converged do
    for each $v \in V$ do
        collect a multiset $\{c(v') | v' \in \Gamma(v)\}$
            containing its neighbors' colors
        sort the multiset in ascending order
        concatenate the sorted multiset to $c(v)$
            to generate a signature string
                $s(v) = \langle c(v), \{c(v') | v' \in \Gamma(v)\}_{sort}\rangle$
    end for
    sort all $s(v)$ in lexicographical ascending order
    map all $s(v)$ to new colors 1, 2, 3... sequentially;
        same strings get the same color
end while
```

WL ensures that final colors encode the structural roles of vertices inside a graph. It also defines a relative ordering for vertices, with ties that is consistent across graphs. More specifically, vertices with the same final color share the same structural role within a graph.

We now fully outline the three-step (1-3) process for implementing a WL-model. Step 1 extracts K-vertex enclosing subgraphs via the "Extract Enclosing Subgraphs Algorithm" below. Step 2 then encodes subgraph patterns via adjacency matrices with vertex ordering given by the Palette-WL, also noted below. Fig. 2 gives an overview of these steps in action. Step 3, training, is via any viable machine learning algorithm. The authors use a neural network, WLNM, to achieve superior results. They also train via linear regression, in a method called WLLR.

Extract Enclosing Subgraphs Algorithm

```
input: target link (x,y), network $G$, integer $K$
output: enclosing subgraph $G(V_K)$ for (x,y)
$V_K = \{x,y\}$
fringe = {x,y}
while $|V_K| < K$ and $|fringe| > 0$ do
    fringe = $(\cup_{v \in fringe} \Gamma(v)) \setminus V_K$
    $V_K = V_K \cup$ fringe
end while
return enclosing subgraph $G(V_K)$
```

Palette-WL Algorithm

```
input: enclosing subgraph $G(V_K)$ centered at link (x, y),
    which is extracted by the EES Algorithm
output: final colors $c(v)$ for all $v \in V_K$
calculate $d(v) = sqrt[d(v,x)^* d(v, y)]$ for all v in $V_K$
```

```
get initial colors $c(v) = f(d(v))$
while $c(v)$ has not converged do
        calculate hashing values $h(v)$ for all $v \in V_K$ by (2)
        get updated colors $c(v) = f(h(v))$
end while
return $c(v)$
```

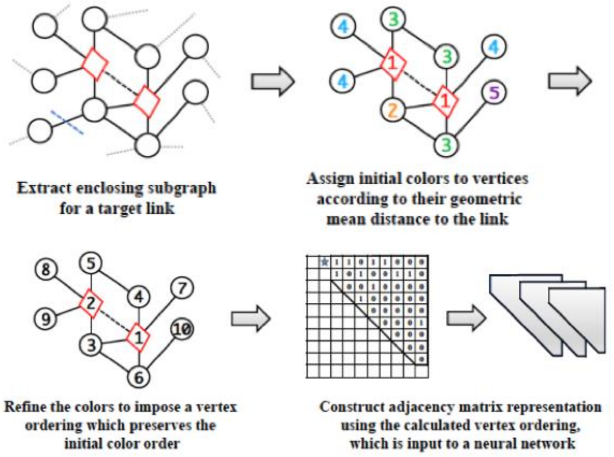

Extract enclosing subgraph for a target link
Assign initial colors to vertices according to their geometric mean distance to the link
Refine the colors to impose a vertex ordering which preserves the initial color order
Construct adjacency matrix representation using the calculated vertex ordering, which is input to a neural network
1
2
3
4
5
6
7
8
9
10

Fig. 2. An overview of a WL-model, Steps 1-2.

Chen and Zhang show, via mathematical proof that the Palette-WL graph labeling function converges in at most K iterations to a one-to-one function for a graph with K vertices. Furthermore, the function is color-order preserving: vertices' color orderings are preserved from state-to-state. Both of these facts enable WL-models to successfully predict links.

V. Assessment of WL-Models

We use Area Under the receiver operating characteristic Curve (AUC) to measure results. AUC measures on the probability that a randomly chosen missing link is given a higher score than a randomly chosen nonexistent link. More precisely, if among n independent comparisons, there are n' times the missing links having a higher score, n'' times those have the same score, the AUC value is $AUC = (n' + 0.5n'')/\text{n}$.

A. Assessment Methods

In our experiments we compare the authors WL-model implemented with linear regression (WLLR), against methods from four traditional link-based assessment areas: common neighbor-based methods, path-based methods, random walk-based methods, and latent feature-based methods. For each test, we compute the AUC value and tabulate the results. The Appendix includes tables with all of our results. We outline the algorithms used in assessment below. Our implementation is motivated largely by two recent papers [38], [39].

B. Common Neighbor (CN)-based Methods

For a node x, let Γ(x) be the set of neighbors of x. The idea is that two nodes x and y are more likely to share a link if they have many common neighbors. The most basic measure CN(x,y), defined to be $|\Gamma(x) \cap \Gamma(y)|$, asserts this. Note that if A is the adjacency matrix for the corresponding graph, then $CN(\text{x, y}) = A^2(\text{x, y})$.

Now let α be a scaler for a link measure M_α so that $M_\alpha(\text{x,y}) = CN(\text{x, y})/\alpha$, and let dx denote the degree of node x. For various choices of α, different CN-measures, noted in Table II, are obtained.

TABLE II. CN-Measures, $|\Gamma(\text{X}) \cap \Gamma(\text{Y})|$ /alpha

α =	CN	Jac	SltOna	Sora
	1	$\lvert\Gamma(x)\cup\Gamma(y)\rvert$	$sqrt(d_x \times d_y)$	$(d_x + d_y)/2$

α =	HPI[a]	HDI[a]	LHN[a]
	$\min\{d_x, d_y\}$	$\max\{d_x, d_y\}$	$d_x \times d_y$

[a] 'Measurements not provided in the KDD paper [40]

The Jaccard Index (Jac) [23] gives the probability that x and y are adjacent, given an edge of either x or y. The Salton Index (SltOn) [20] is often also called cosine similarity in the literature. Sørensen Index (Sor) [24] is primarily used for ecological data. Hub Promoted Index (HPI) [25] is used to quantify the topological overlap of pairs of substrates in metabolic network; under this scheme, links adjacent to hubs are more likely to be assigned to be assigned high scores. Hub Depressed Index (HDI) is the complementary measure to HPI. Finally, the Leicht-Holme-Newman Index (LHN) [22] assigns high similarity to node pairs that have many common neighbors compared to the expected number of such neighbors; in particular, $d_x \times d_y$ is proportional to the expected number of common neighbors of nodes x and y in the configuration model [26].

Three related CN-based methods we considered are Preferential Attachment Index (PA), Adamic-Adar Index (AA), and Resource Allocation Index (RA); these are noted in Table III. PA is motivated by a preferential attachment mechanism which ensures that the probability that a new link to be added connects x and y is proportional to $d_x \times d_y$. PA is often used to quantify the functional significance of links subject to various network-based dynamics such as percolation [27], synchronization [28], and transportation [29]. AA is a refinement of simple counting of common neighbors; it assigns less-connected neighbors more weight [2]. RA [8] is motivated by the resource allocation dynamics on complex networks. Suppose x and y are not directly connected and x can send a resource to y, with common neighbors playing the role of transmitters. If each transmitter has a unit of resource, and distributes equally to all its neighbors, then RA is the amount of resource y received from x.

TABLE III. Three Related CN-based Measures

PA	AA	RA
$d_x \times d_y$	$\sum_{z \in \Gamma(x) \cap \Gamma(y)} (1/\log(d_z))$	$\sum_{z \in \Gamma(x) \cap \Gamma(y)} (1/d_z)$

We also considered three local naïve Bayes methods with common neighbor, Adar-Adamic index, and resource allocation, respectively. These are listed as LNBCN, LNAA, and LNBRA in the tables provided in the Appendix.

In our experimental runs on the five data sets, we able to validate all of the KDD results [40] on the measures that were used in that paper: CN, Jac, AA, RA, and PA. See the Appendix, Tables VI, VII and VIII. Numerical values were rarely the same, but sufficiently close. In addition, in each of our experimental runs inclusive of all additional measures, the RA index generally performs best, while the AA, CN, and LNBAA indices follow closely behind in best overall performance. This “best performance” neglects comparisons against the WLNM test runs. Because WLNM outperforms even AA and CN, WLNM still provides superior performance according to KDD data. A caveat is that we only ran the WL-model with linear regression, called WLLR, which fared worse amongst the various CN-measures above. In fact, almost without exception, all 13 common neighbor methods implemented exhibited superior performance over WLLR on the NS, PB, and Yeast data sets. Exceptions occurred with PA, Jac, LHN, and LNBRA on certain data sets.

C. Path-based Methods

The Katz Index [9] is based on an ensemble of all paths. It is a sum over the collection of all paths with a damping factor β providing shorter paths more weight. Letting A be the adjacency matrix, $Katz(x, y) = \sum_{i \geq 1}(\beta A)^i = (I - \beta A)^{-1} - I$.

The Local Path Index (LP) [8, 33] takes local paths into additional consideration beyond CN and is defined as $LPI(x, y) = A^2 + \epsilon A^3$ where ϵ is a free parameter; note that when $\epsilon = 0$, the index is just CN. A more expanded version allows for n sum factors, and as n → ∞ the index becomes Katz. Experimental results show that the optimal n is positively correlated with the average shortest distance of the network [33].

The Leicht-Holme-Newman Index (LHNII) [22] is a variant of the Katz index and is based on the concept that two nodes are similar if their immediate neighbors are themselves similar. A self-consistent matrix formulation is $S = \beta AS + \psi I = \psi(I - \beta A)^{-1} = \psi(I + Katz(x, y))$ where ψ and β are free parameters controlling the balance between the two components of the similarity. An formulation useful for computations is $S = 2m\lambda D^{-1}(I - \beta A/\lambda)^{-1}D^{-1}$ where λ is the largest eigenvalue of A, m is the total number of edges in the network, D is the degree matrix, and β is a free parameter. The choosing of β depends on the investigated network, and smaller β assigns more weights on shorter paths.

The KDD paper [40] only implements Katz with $\beta = 0.01$. In our experiments we also run Katz with $\beta = 0.001$, as well as LocalPath, and LHNII with $\beta = 0.9$, 0.95, and 0.99. See Appendix, Tables IX and X. In our implementations, on four data sets (USAir, NS, PB, and Yeast), Katz (both versions) and LocalPath always exhibited superior performance to WLLR, with the exception of Katz ($\beta = 0.01$) on USAir. For the NS dataset, LHNII actually exhibited top performance on WLLR over all methods discussed in this section.

D. Random Walk-based Methods

Resistance Distance (RD) is often called Average Commute Time (ACT) in other contexts and equal to $s(x, y) + s(y, x)$ where $s(x, y)$ denotes the average number of steps required by a random walker starting from node x to reach node y. The pseudoinverse L^+ of the Laplacian matrix $L = D - A$, is easily computable as $m(L^+(x, x) + L^+(y, y) - 2L^+(x, y))$ [11, 30]. ACT(x,y) is defined as the reciprocal of this with m=1 in order to ensure that two nodes are more similar whenever they have a smaller average commute time.

The PageRank (PR) algorithm [13] may be directly applied using Random Walk with Restart (RWR). Consider a random walker starting from node x, who will iteratively move to a random neighbor with probability c and return to node x with probability 1 – c. Let p_{xy} denote the probability that a random walker locates at node y in the steady state. Then $\boldsymbol{p_x} = \langle p_{x1}, p_{x2}, ...\rangle$ is given by $\boldsymbol{p_x} = c \cdot P^T\boldsymbol{p_x} + (1 - c) \cdot e_x$ where e_x is the n × 1 vector in Euclidean n-space which is 1 at entry x and zero elsewhere, and P is the transition matrix with $P(x, y) = 1/d_x$ if x and y are connected and 0 otherwise. The solution, given by $\boldsymbol{p_x} = (1 - c)(I - cP^T)^{-1}e_x$, is used in the code for this project. Define RWR as $RWR(x, y)= p_{xy} + p_{yx}$.

SimRank (SR) [31] is defined in a self-consistent way similar to LHNI as $SR(x, y) = \beta\,(\sum_{z\in\Gamma(x)} \sum_{z'\in\Gamma(y)} SR(z, z'))/(d_xd_y)$ where $SR(x, x) = 1$ and $\beta \in [0, 1]$ is a decay factor. The underlying idea is that two nodes are similar if they are connected to similar nodes. From a random-walk perspective, SR measures how soon two random walkers, respectively starting from nodes x and y, are expected to meet at a certain node.

Cosine Similarity based on L^+ (Cos+) [30] is defined using pseudoinverse L^+ of the Laplacian matrix $L = D - A$ so that

$Cos+(x, y) = L^+(x, y)/\text{sqrt}(L^+(x, x) \cdot L^+(y, y))$.

Local Random Walk with step s (LRWs) [34] measures the similarity between nodes x and y when random walker is initially put on node x and proceeds for s steps. The density vector is defined by $V_x(0) = e_x$ and $V_x(t + 1) = P^T \cdot V_x(t)$ for $t \geq 0$. Then $\text{LRWs}(x, y) = init(x) \cdot V_{xy}(s) + init(y) \cdot V_{yx}(s)$ where $init$ is the initial configuration function. In [34] Liu and Lü determine $init(x)$ by node degree so that $init(x) = d_x/m$.

Superposed Random Walk with step s [34] is similar to the LRWs index and defined as $\text{SRWs}(x, y) = \sum_{1 \leq i \leq s} \text{LRWs}(x, y)$. Here a random walker is continuously released at the starting point. A higher similarity is between the target node and the nodes nearby results.

Matrix Forest Index [32] is defined by $MFI = (I + L)^{-1}$. MFI gives the ratio of the number of spanning rooted forests (such that nodes x and y belong to the same tree rooted at x) to all spanning rooted forests of the network.

Transfer Similarity with CN (TS) is defined, using a parametrized version of MFI, by $TS = (I + \lambda \cdot CN)^{-1} * CN$.

The KDD paper [40] only implements RD, PR, and SR. In our experiments we also implemented Cos+, RWR with $\beta = 0.95$, LRW with 3-5, SRW with 3-5, MFI, and TS. See Appendix, Tables XI, XII, and XIII. In our implementations, random walk-based methods could not be calculated on two datasets (NS, Yeast) due to memory limitations. On the other three sets (USAir, PB, and Celegens), every random walk method was superior to the WLLR model, with the exception of RD and SR and TS on all sets, and RWR 0.95 and MFI on USAir.

E. Latent Feature-based Methods

Latent (present participle of lateo, "lie hidden") feature-based models attempt recover hidden features which are then used to predict graphs links.

Stochastic Block Model (SBM) [5], [35]-[37] partitions nodes into groups and the probability that two nodes are connected depends solely on the groups to which they belong. Let M be a partition, Q_{ab} denote the probability that groups a and b are connected (so that $Q_{aa} = 1$ for all a), and c_{ab} denote the number of connections (i.e. edges) between groups. The likelihood $L(A|M)$ of observed structure A given M is therefore $\Pi_{a \leq b}(Q_{ab})^{c\_(ab)} \cdot (1 - Q_{ab})^{1 - c\_(ab)}$. From this $SBM(x, y)$ [21] is defined via Bayes' Theorem as

$$\int_{\Omega} L(A_{xy} = 1|M)L(A|M)p(M)dM \div \int_{\Omega} L(A|M')p(M')dM'$$

where Ω is the set of all partitions and $p(M)$ is a constant assuming no prior knowledge about the model.

With Matrix Factorization (MF) [39], some entries in A are unknown. MF attempts to approximate A, using only known entries, into two low-dimensional matrices so that $A \approx FG^T$ with F being $N \times K$, G being $N \times K$, and K being the number of latent features. The squared error is thus given by $(e_{ij})^2 = (a_{ij} - \sum_{k \leq K} f_{ik} g_{kj})^2$. A regularization technique adds a factor to avoid over-fitting so that instead $(e_{ij})^2 = (a_{ij} - \sum_{k \leq K} f_{ik} g_{kj})^2 + (\beta /2)\sum_{k \leq K}(\|F\|^2 + \|G\|^2)$. The goal is to minimize the sum of all squared errors to obtain optimal F and G. The gradient at current values is calculated via partial differentiation with respect to f_{ik} and separately with respect to g_{kj}. Weights are then updated in the direction opposite the gradient and this gives rules $f_{ik}' = f_{ij} + \alpha (2 e_{ij} g_{kj} - \beta f_{ik})$ and $g_{kj}' = g_{kj} + \alpha (2 e_{ij} f_{ik} - \beta g_{kj})$ which are then used iteratively until error converges to a minimum. Implicit in the above formulation is that squared errors must be known elements of A, so a_{ij} is in the training set.

Amongst the two methods, our results showed that SB generally provided better results with respect to link prediction. See Appdendix, Table XIV. For the USAir and Celegens data sets, SB outperformed the WLLR. For the PB and Celegens datasets, MF likewise outperformed WLLR.

F. Summary

Each of the aforementioned methods were implemented with various parameters as outlined in the Appendix. In particular, our implementation extends the KDD authors' implementation by testing all data sets on the additional common neighbor schemes (13 instead of 5), path-based methods (6 instead of 1), random walk-based methods (13 instead of 3). We also implemented all of the authors' latent feature models. On the five datasets sets tested, all author data on these methods were able to be validated. In our implementation, a WL-model is run as linear regression and noted as WLLR. WLLR did not exhibit superior performance in any run for which comparisons were possible. Amongst the comparisons possible, in 80% of these our implementations (not implemented in the KDD paper) demonstrated superior performance. In particular, WLLR AUC results for NS, PB, Yeast, and Celegens were 0.865, 0.838, 0.860, and 0.804, respectively; the corresponding results for LHRII (all cases), LRW3, LNBAA, and LRW3 were 0.9690, 0.9367, 0.8990, and 0.9197, respectively. For USAir, the WLLR AUC score was 0.930 and RA demonstrated superior performance with an AUC score of 0.9540. Common neighbor methods were superior in 40% of the cases. Random walk-based methods were superior in another 40% of the cases. Finally, the Path-based methods were superior in the remaining 20% of cases. Table IV demonstrates these results.

TABLE IV. COMPUTATIONS

Data Set	WLLR	Top Score	Models with Top Score	Top Score Class	Top Score in this Paper
USAir	0.930	0.9540	RA	Common neighbor	
NS	0.865	0.9690	LHNII B ∈{3,4,5}	Path-based	✔□
PB	0.838	0.9367	LRW3	Random-walk	✔□
Yeast	0.860	0.8990	AA, RA, LNBAA	Common neighbor	✔□[a]
C.ele	0.804	0.9197	LRW3	Random-walk	✔□

[a] AA and RA were implemented in [40]; LNBAA was implemented in this paper

VI. CONCLUSION

Chen and Zhang [40] develop novel WL-link prediction scheme which to-date best predicts links in real world graph data, according to experimental data. An innovative feature is that link formation mechanisms are learned, not assumed. WL-link prediction schemes work by encoding enclosed subgraphs as adjacency matrices. Encoding occurs via the author's modification of the Weisfeiler-Leman (WL) algorithm [12] from graph theory. The authors' modification, Palette-WL, labels vertices according to their structural roles in the subgraph and preserves subgraph intrinsic directionality. Training on adjacency matrices then learns a predictive model. When training is done on the authors' neural network, the model is called WLNM, i.e. WL Neural Machine, and this requires advanced support for parallelism or distributed computing. When training is done with linear regression, the model is called WLLR, i.e. WL Linear Regression, and this was the focus of our work; forthcoming work will tackle WLNM.

We extended the authors' implementations by testing all data sets on additional CN-, path- and random walk-based schemes. In particular, we implemented 32 such schemes compared to the nine from the KDD authors. We also implemented all of the authors' latent feature models. On five data sets, all author data on these methods were validated. The WLNM model still demonstrates superiority even when all additional schemes are implemented, according to data provided by the KDD authors. The linear regression version of the model, WLLR, on the other hand, was not superior when calculable. In fact, in 80% of the datasets where comparisons were possible, one of our added implementations proved superior.

VII. Future Work

Given the successes in validating the results from [40] for WLLR and in demonstrating a multitude of results which supplant that model, our next step will be to implement and validate WLNM. That is, the current work is limited in scope to only the WLLR model. Therefore, the goal will be to determine if the WLNM model can also be supplanted. We will also run all experiments on the three additional data sets (Power, Router, and E.coli) which the authors test in [40] and which require advanced support for parallelism or distributed computing.

Other opportunities for future work also abound. For instance, in what ways can might the graph coloring scheme be applied to other algorithms in data science (e.g. clustering)? Also, as the authors' algorithm pertains only to undirected graphs, how might the WLNM be modified in order to apply to directed graphs? It is also plausible that for specific classes of graphs, a parred set of calculations might suffice towards leading to manageable neural network computations; one direction for future work would be to identify such classes and prove optimal computational bounds. Another line of work would be to determine whether there might be a class of circumstances in which a heuristic method (or an embedding model) may provide a better result. If so, what properties, would such a class or model have, mathematically, and could an example be found in nature?

References

[1] D. Liben-Nowell and J. Kleinberg, "The link-prediction problem for social networks", Journal of the American society for information science and technology, Vol. 58, No. 7, pp. 1019–1031, 2007.

[2] L. A. Adamic and E. Adar, "Friends and neighbors on the web", Social networks, Vol. 25, No. 3, pp. 211–230, 2003.

[3] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems", Computer, Vol. 8, pp. 30–37, 2009.

[4] M. Nickel, K. Murphy, V. Tresp, and E. Gabrilovich, "A review of relational machine learning for knowledge graphs", arXiv preprint, 1503.00759, 2015.

[5] E. M. Airoldi, D. M. Blei, S. E. Fienberg, and E. P. Xing, "Mixed membership stochastic blockmodels", Journal of Machine Learning Research, Vol. 9, pp. 1981–2014, September 2008.

[6] T. Oyetunde, M. Zhang, Y. Chen, Y. Tang, and C. L. Boostgapfill, "Improving the fidelity of metabolic network reconstructions through integrated constraint and pattern-based methods", Bioinformatics, 2016.

[7] R. Salakhutdinov and A. Mnih, "Bayesian probabilistic matrix factorization using markov chain monte carlo", Proceedings of the 25th international conference on Machine learning (ICML), ACM, pp. 880–887, 2008.

[8] T. Zhou, L. Lü, and Y. Zhang, "Predicting missing links via local information", European Physical Journal B, Vol. 71, No. 4, pp. 623–630, 2009.

[9] L. Katz, "A new status index derived from sociometric analysis", Psychometrika, Vol. 18, No. 1, pp. 39–43, 1953.

[10] L. Lü and T. Zhou, "Link prediction in complex networks: A survey", Physica A: Statistical Mechanics and its Applications, Vol. 390, No. 6, pp. 1150–1170, 2011.

[11] D. J. Klein and M. Randic, "Resistance distance", Journal of Mathematical Chemistry, Vol. 12, No. 1, pp. 81–95, 1993.

[12] B. Weisfeiler and A. A. Lehman, "A reduction of a graph to a canonical form and an algebra arising during this reduction", Nauchno-Technicheskaya Informatsia, Vol. 2, No. 9, pp. 12–16, 1968.

[13] S. Brin and L. Page, Reprint of: "The anatomy of a large-scale hypertextual web search engine", Computer networks, Vol. 56, No. 18, pp. 3825–3833, 2012.

[14] N. Shervashidze, P. Schweitzer, E. J. van Leeuwen, K. Mehlhorn, and K. M. Borgwardt, "Weisfeiler-lehman graph kernels", Journal of Machine Learning Research, Vol. 12, pp. 2539–2561, September 2011.

[15] M. Niepert, M. Ahmed, and K. Kutzkov, "Learning convolutional neural networks for graphs", Proceedings of the 33rd annual international conference on machine learning, ACM, 2016.

[16] K. Miller, M. I. Jordan, and T. L. Griffiths, "Nonparametric latent feature models for link prediction", Advances in neural information processing systems, pp. 1276–1284, 2009.

[17] S. Rendle, C. Freudenthaler, Z. Gantner, and L. Schmidt-Thieme, "Bpr: Bayesian personalized ranking from implicit feedback", Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence, AUAI Press, pp. 452–461. 2009.

[18] C. Aicher, A. Z. Jacobs, and A. Clauset, "Learning latent block structure in weighted networks", Journal of Complex Networks, Vol. 3, No. 2, pp. 221–248, 2015.

[19] S. V. N. Vishwanathan, N. N. Schraudolph, R. Kondor, and K. M. Borgwardt, "Graph kernels", Journal of Machine Learning Research, Vol. 11, pp. 1201– 1242, April 2010.

[20] G. Salton, M. J. McGill, Introduction to Modern Information Retrieval, MuGraw-Hill, Auckland, 1983.

[21] R. Guimera, M. Sales-Pardo, "Missing and spurious interactions and the reconstruction of complex networks", Proc. Natl. Acad. Sci. U.S.A., Vol. 106, 2009.

[22] E. A. Leicht, P. Holme, M. E. J. Newman, "Vertex similarity in networks", Phys. Rev. E, Vol. 73, 2006.

[23] P. Jaccard, "E´tude comparative de la distribution florale dans une portion des Alpes et des Jura", Bulletin de la Societe Vaudoise des Science Naturelles, Vol. 37, No. 1901, pp. 547.

[24] T. Sørensen, "A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons", Biol. Skr., Vol. 5, pp. 1, 1948.

[25] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A. L. Barabasi, "Hierarchical organization of modularity in metabolic networks", Science, Vol. 297, No. 1551, 2002.

[26] M. Molloy and B. Reed, "A critical point for random graphs with a given degree sequence", Random Structure and Algorithms, Vol. 6, No. 161, 1995.

[27] P. Holme, B. J. Kim, C. N. Yoon, and S. K. Han, "Attack vulnerability of complex networks", Phys. Rev. E, Vol. 65, 2002.

[28] C. Y. Yin, W. X. Wang, G. R. Chen, and B. H. Wang, "Decoupling process for better synchronizability on scale-free networks", Phys. Rev. E, Vol. 74, 2006.

[29] G. Q. Zhang, D. Wang, and G. J. Li, "Enhancing the transmission efficiency by edge deletion in scale-free networks", Phys. Rev. E, No. 76, 2007.

[30] F. Fouss, A. Pirotte, J. M. Renders, and M. Saerens, "Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation", IEEE Trans. Knowl. Data. Eng., Vol. 19, No. 355, 2007.

[31] G. Jeh and J. Widom, "SimRank: A measure of structural-context similarity", Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM Press, New York, pp. 271-279, 2002.

[32] P. Chebotarev and E. V. Shamis, "The matrix-forest theorem and measuring relations in small social groups", Automation and Remote Control, Vol. 58, No. 1505, 1997.

[33] L. Lü, C. H. Jin, and T. Zhou, "Similarity index based on local paths for link prediction of complex networks", Phys. Rev. E, Vol. 80, 2009.

[34] W. Liu and L. Lü, "Link prediction based on local random walk", EPL, Vol. 89, 2010.

[35] H. C. White, S. A. Boorman, and R. L. Breiger, "Social structure from multiple networks I: Blockmodels of roles and positions", Am. J. Sociol., Vol. 81, No. 730, 1976.

[36] P. W. Holland, K. B. Laskey, and S. Leinhardt, "Stochastic blockmodels: First steps", Social Networks, Vol. 5, No. 109, 1983.

[37] P. Dorelan, V. Batagelj, and A. Ferligoj, Generalized Blockmodeling, Cambridge University Press, Cambridge, UK, 2005.

[38] L. Lü and T. Zhao, "Link Prediction in Complex Networks: A survey", "Physica A: Statistical Mechanics and its Applications", Vol. 390, No. 6, pp. 1150-1170, 2011.

[39] Z. Wu and Y. Chen, "Link prediction using matrix factorization with bagging", 15th Int. Conf. on Computer and Inf. Sci. (ICIS), Okayama, pp. 1-6, 2016.

[40] Y. Chen and M. Zhang, Weisfeiler-Lehman Neural Machine for Link Prediction, 23rd ACM SIGKDD Int. Conf. on KDD Mining, New York, NY, USA, pp. 575-583, 2017.

APPENDIX

The full set of results from this paper are presented in various tables as Area Under the receiver operating characteristic Curve (i.e. AUC) measurements ("Test"). Calculations are towards validating the results in [40] ("Paper").

Entries with '–' require a Graphics Processing Unit for calculations and are outside the scope of this study. Tables or columns marked with * contain results run exclusively in this paper; no such computations were run in [40] ("Paper").

TABLE V. WEISFEILER-LEMAN BASED METHODS

No	Data	Source	WLLR 10[a]	WLNM 10[b]
1	USAir	Paper	0.896	0.958
		Test	0.930	–
2	NS	Paper	0.862	0.984
		Test	0.865	–
3	PB	Paper	0.827	0.933
		Test	0.838	–
4	Yeast	Paper	0.854	0.956
		Test	0.860	–
5	C.ele	Paper	0.803	0.859
		Test	0.804	–

[a] Weisfeiler-Leman Linear Regression Model, K = 10 (WLLR 10)

[b] Weisfeiler-Leman Neural Machine, K = 10 (WLMN 10)

TABLE VI. COMMON NEIGHBOR METHODS I

No	Data	Source	CN[a]	Jac[b]	AA[c]	RA[d]	PA[e]
1	USAir	Paper	0.940	0.903	0.950	0.956	0.894
		Test	0.939	0.905	0.950	0.954	0.899
2	NS	Paper	0.938	0.938	0.938	0.938	0.682
		Test	0.945	0.945	0.945	0.945	0.705
3	PB	Paper	0.919	0.873	0.922	0.923	0.901
		Test	0.912	0.867	0.915	0.917	0.897
4	Yeast	Paper	0.891	0.890	0.891	0.892	0.024
		Test	0.898	0.897	0.899	0.899	0.836
5	C.ele	Paper	0.848	0.792	0.864	0.868	0.755
		Test	0.842	0.782	0.858	0.862	0.761

[a] Common Neighbor (CN)

[b] Jaccard Index (Jac)

[c] Adar-Adamic Index (AA)

[d] Resource Allocation (RA)

[e] Preferential Attachment (PA)

TABLE VII. COMMON NEIGHBOR METHODS II*

No	Data	Source	SltOn[a]	Sor[b]	HPI[c]	HDI[d]
1	USAir	Test	0.9037	0.8959	0.8621	0.8891
2	NS	Test	0.9450	0.945	0.9449	0.9449
3	PB	Test	0.8692	0.8673	0.8478	0.8637
4	Yeast	Test	0.8972	0.8972	0.8961	0.8971
5	C.ele	Test	0.7897	0.7816	0.7958	0.7685

[a] Salton Index (SltOn)

[b] Sorenson Index (Sor)

[c] Hub Promoted Index (HPI)

[d] Hub Depressed Index (HDI)

TABLE VIII. COMMON NEIGHBOR METHODS III

No	Data	Source	LHN*	LNBCN[a]	LNBAA[b]	LNBRA[c]
1	USAir	Test	0.7615	0.9434	0.9503	0.8943
2	NS	Test	0.9446	0.9452	0.9452	0.7045
3	PB	Test	0.7584	0.915	0.9165	0.8973
4	Yeast	Test	0.8932	0.8987	0.899	0.8355
5	C.ele	Test	0.716	0.858	0.8623	0.7605

[a] Leicht-Holme-Newman (LHN)

[b] Local naive bayes method with Common Neighbor (LNBCN)

[c] Local naive bayes method with Adar-Adamic Index (LNBAA)

[d] Local naive bayes method with Resource Allocation (LNBRA)

TABLE IX. PATH-BASED METHODS I

No	Data	Source	Katz with β= 0.01[a]
1	USAir	Paper	0.931
		Test	0.926
2	NS	Paper	0.940
		Test	0.947
3	PB	Paper	0.928
		Test	0.924
4	Yeast	Paper	0.921
		Test	–
5	C.ele	Paper	0.864
		Test	0.860

[a] Katz Index with damping factor β = 0.01

TABLE X. PATH-BASED METHODS II

No	Data	Source	Katz $\beta = 0.001$[a]	LocalPath[b]	LHNII 0.9[c]	LHNII 0.95[d]	LHNII 0.99[e]
1	USAir	Test	0.9279	0.9306	0.6040	0.5870	0.5712
2	NS	Test	0.9474	0.9499	0.9690	0.9690	0.9690
3	PB	Test	0.9266	0.9273	0.6363	0.5810	0.5273
4	Yeast	Test	–	–	–	–	–
5	C.ele	Test	0.8614	0.8626	0.6070	0.5551	0.5003

[a.] Katz Index with damping factor $\beta = 0.001$

[b.] Local Path Index (LocalPath)

[c.] Leicht-Holme-Newman II with 0.90

[d.] Leicht-Holme-Newman II with 0.95

[e.] Leicht-Holme-Newman II with 0.99

TABLE XI. RANDOM-WALK BASED METHODS I

No	Data	Source	RD[a]	PR[b]	SR[c]
1	USAir	Paper	0.898	0.944	0.782
		Test	0.911	0.931	0.775
2	NS	Paper	0.582	0.940	0.940
		Test	–	–	–
3	PB	Paper	0.883	0.935	0.773
		Test	0.879	0.930	0.771
4	Yeast	Paper	0.880	0.927	0.914
		Test	–	–	–
5	C.ele	Paper	0.740	0.901	0.760
		Test	0.726	0.899	0.758

[a.] Resistance Distance, or Average Commute Time (RD)

[b.] PageRank, or Random Walk with restart, with damping factor d = 0.85 (PR)

[c.] SimRank with 0.6 (SR)

TABLE XII. RANDOM-WALK BASED METHODS II*

No	Data	Source	Cos+[a]	RWR[b]	LRW 3[c]	LRW 4[d]	LRW 5[e]
1	USAir	Test	0.9342	0.914	0.9389	0.9367	0.9337
2	NS	Test	–	–	–	–	–
3	PB	Test	0.9196	0.917	0.9367	0.9293	0.9325
4	Yeast	Test	–	–	–	–	–
5	C.ele	Test	0.865	0.857	0.9197	0.9034	0.9105

[a.] Cos+ based on Laplacian matrix (Cos+)

[b.] Random walk with restart with damping factor 0.95 (RWR 0.95)

[c.] Local Random Walk with step 3 (LRW 3)

[d.] Local Random Walk with step 4 (LRW 4)

[e.] Local Random Walk with step 5 (LRW 5)

TABLE XIII. RANDOM-WALK BASED METHODS III

No	Data	Source[a]	SRW 3[a]	SRW 4[b]	SRW 5[c]	MFI[d]	TS
1	USAir	Test	0.9407	0.9389	0.9384	0.9129	0.589
2	NS	Test	–	–	–	–	–
3	PB	Test	0.9257	0.9272	0.9292	0.8959	0.4417
4	Yeast	Test	–	–	–	–	–
5	C.ele	Test	0.9009	0.9031	0.9063	0.8722	0.5076

[a.] Superposed Random Walk with step 3 (SRW 3)

[b.] Superposed Random Walk with step 4 (SRW 4)

[c.] Superposed Random Walk with step 5 (SRW 5)

[d.] Matrix Forest Index (MFI)

[e.] Transfer Similarity (TS)

TABLE XIV. LATENT FEATURE BASED METHODS

No	Data	Source	SBM[a]	MF-c[b]
1	USAir	Paper	0.944	0.918
		Test	0.932	0.914
2	NS	Paper	0.920	0.636
		Test	–	0.620
3	PB	Paper	0.938	0.930
		Test	–	0.927
4	Yeast	Paper	0.914	0.831
		Test	–	–
5	C.ele	Paper	0.867	0.832
		Test	0.878	0.837

[a.] Stochastic Block Method (SBM)

[b.] Matrix Factorization with classification loss function (MF-c)

Methodology for Selecting the Preferred Networked Computer System Solution for Dynamic Continuous Defense Missions

San Diego
Dr. Glenn S. Tolentino
Command & Control and Enterprise Engineering Department
SPAWAR Systems Center Pacific
Dallas, TX, USA

Dr. Jeff Tian
Computer Science and Engineering Department
Southern Methodist University
Dallas, TX, USA

Dr. Jerrell T. Stracener
Engineering Management, Information, and Systems Department
Southern Methodist University

***Abstract*—This paper presents a methodology for addressing the challenges and opportunities in defining and selecting the preferred Networked Computer System (NCS) solution in response to specified United States Defense mission planning requirements. The identified set of mission requirements are aligned with existing computer system capabilities allowing them to be acquired and processed as candidates to be included as part of the preferred NCS solution. In performing the proper selection process, decision making process is required in being able to properly select the preferred NCS by utilizing associated models for analysis. The models will then be applied towards NCS mission planning in analyzing an NCS solution's effectiveness in terms of operational availability, mission reliability, capability sustainment and lifecycle cost. The analysis and models were developed in response to the need to develop defense mission planning capability solutions by utilizing existing computer systems enabling the Department of Defense acquisition professionals to perform a practical approach in selecting and defining the preferred NCS for satisfying a mission.**

***Keywords*—Mission reliability; sustainment reliability; operational availability; basic reliability; networked computer system; system of systems**

I. Introduction

There's currently a methodology framework as part of this research that provides the process for modeling and analysis of an NCS's cost-effectiveness that provides the basis for ranking candidate computer systems and for selecting the preferred NCS solution. In this paper, we present models and describe the associated analysis for NCS mission reliability, sustainment reliability, operational availability and lifecycle cost that provide estimates for DoD acquisition managers to use during the decision-making process of defining the preferred NCS solution in response to defense mission planning requirements [6]. In developing this paper, there were ground rules that were determined and assumed. The following ground rules are described as part of the NCS cost-effectiveness modeling and analysis:

- The constituent computer systems are currently operational, or will be within the required acquisition time.
- The NCS solution requires existing information from each of the owners of the constituent systems as input to cost-system effectiveness analysis.
- The NCS solution executes on a computer network whose availability and reliability is not considered as part of the NCS cost-effectiveness analysis.
- Each of the computer systems assumes data required to perform their respective capability is available.

II. Methodology Factors and their Relationships

NCS systems solutions that are formulated in response to DoD defense mission planning requirements may be compared in terms of systems cost effectiveness. NCS system cost effectiveness is a function of two factors, namely, system effectiveness and system cost. While there is not a standard definition of system effectiveness, we use a generally accepted definition consisting of three factors; system availability, system performance (form, fit, and function) and system mission reliability. In terms of system availability, it depends upon two factors: system sustainment reliability and system capability sustainment (Integrated Logistics Support plus Capability Upgrades) [3]. System cost is the total cost of ownership over the system life cycles as specified by the requirements for a needed NCS for DoD mission planning. In general, system cost consists of Acquisition Cost plus the Operating and Sustainment Cost plus the Disposal Cost.

Note that all NCS system cost effectiveness factors are correlated and related variables. For example, the mission reliability for a candidate NCS solution, a System of Systems (SoS), can be increased by adding redundancy, i.e., by putting an identical system in parallel to a NCS system. In adding a redundant system, that specific change in mission reliability configuration in term will increase NCS mission reliability, and may, or may not, increase system effectiveness. But, it will increase system cost while decreasing NCS availability as a result of decreasing NCS Sustainment Reliability and increasing NCS Capability Sustainment. These relationships are indicated in Fig. 1.

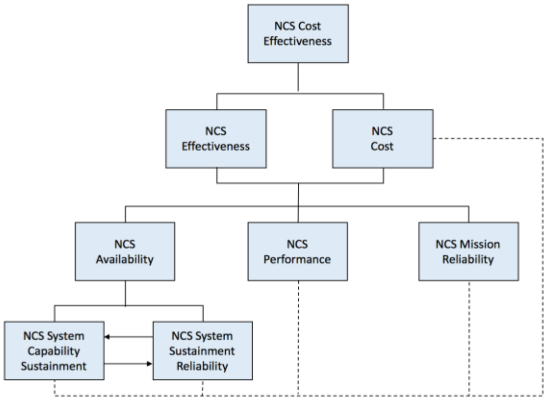


Fig. 1. NCS cost effectiveness relationships.

Two major contributions of the methodology that this study made to the current body of knowledge, and specifically to the DoD, are the capability to quantify (a) NCS cost-effectiveness for required capability so that candidate solutions may be objectively compared and (b) the impact of trades among the NCS cost-effectiveness variables (see Fig. 2).

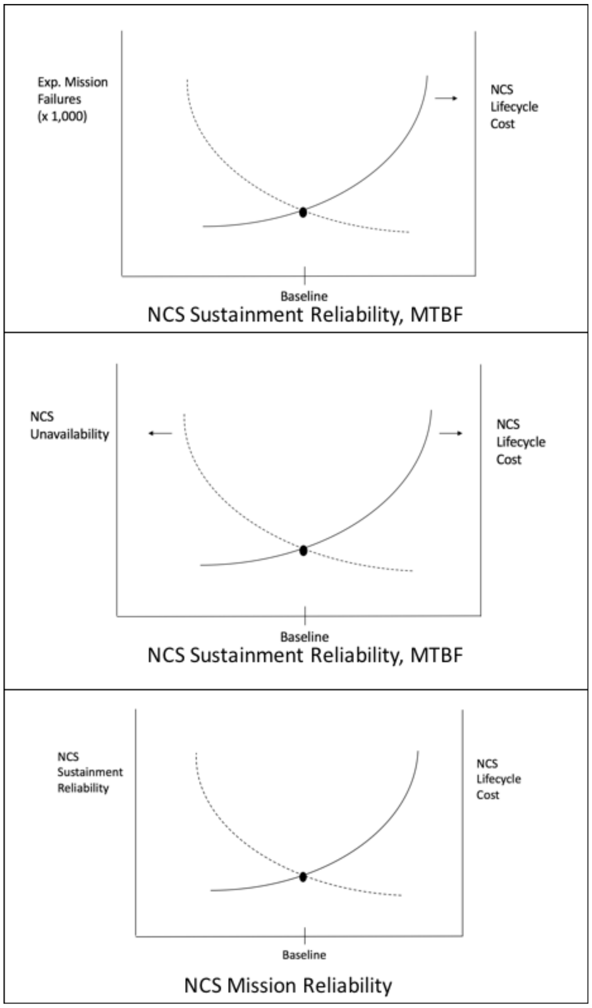


Fig. 2. NCS cost-effectiveness interrelationships.

III. NCS Form, FIT, and Function with Respect to Requirements

Form, Fit, and Function is the identification and description of specific characteristics of a part, component, and/or assembly of a component or in this case the NCS. By developing the NCS based on requirements, it would allow for the possibility of making certain changes to the NCS while keeping the form, fit, and function of the solution. This is a critical consideration in selection of the computer systems and defining preferred NCS solution such that computer systems is unique but relevant, fits into the overall NCS, and functions as required in providing its capabilities.

The “form” refers to specific dimensions, size, and physical characteristic of a NCS. In this case, the form is the computer systems and network architecture that makes up the NCS. In describing the overall characteristic of a NCS, the form is a comprehensive number of computer systems that is made up of hardware and software in performing a specific mission. The network architecture is also considered as part of the “form” as it provides the physical connectivity of the computer systems. The reference to an NCS “fit” refers to the ability for the computer systems that are parts of the NCS to be able to interconnect and interface with other computer systems. In this case of “fit”, computer systems communicate with one another through computer networking providing its capability at some time period. The ability for each of these computer systems to be able to connect, interface, interoperate, and communicate with one another satisfies the definition of the term “fit” in the case of an NCS architecture. In translating “function” of the NCS, the function refers to the purpose of each of the computer systems of the NCS in being able to perform its own capabilities. This defines that each computer system capability that is expected to perform in fulfilling its purpose. Furthermore, the capabilities performing its own function can be extended to the overall NCS mission as the overarching function.

The NCS form, fit, and function requirements are based on overall requirements for an NCS for mission planning. The form, fit, and function of each candidate NCS is evaluated to determine whether or not the NCS for mission planning are met. If not, the candidate NCS is eliminated from consideration.

IV. NCS Mission Reliability and Analysis

There are methodologies for addressing mission reliability modeling and analysis being performed for individual systems [3], [4]. However, the methodology developed for this research and dissertation has been modified to accurately model constituent computer systems that are composed of both hardware and software while interconnected through computer networking in order to function as a SoS. In addressing mission reliability of a single constituent computer system, software and hardware reliability is provided as single reliability measures [2], [7]. Future research is considered in the area of the interrelationship between software and hardware reliability with respect to mission reliability. However, in this research of NCS mission reliability, the overall system reliability is considered as one measure for both hardware and software. Furthermore, the methodology developed addresses mission reliability in terms of a single NCS solution. The analysis starts with an overview of the NCS and its mission planning to develop the mission description, Concept of Operations (CONOPS), and mission definition and description. It is followed in describing the NCS mission profile, mission essential functions, success criteria, mission essential hardware and software for each phase.

The following ground rules are described and assumed as part of the NCS Mission Reliability modeling and analysis:

- The NCS mission is deemed successful when all NCS phases are completed individually by each computer system capability.
- The NCS and its mission are considered failures when any of the constituent computer system in its respective NCS systems architecture phase fails during the mission.
- The constituent computer systems may have redundancy or failover capabilities if designed as a redundant system.
- The length of each of the NCS mission phases is determined prior to performing the mission.
- No failures are repaired at any time during the NCS mission.
- Each of the computer systems reliability measures is provided by each of the system owners.

A. *Overview and Process*

The model development and analysis overview is described and depicted with the following process (see Fig. 3).

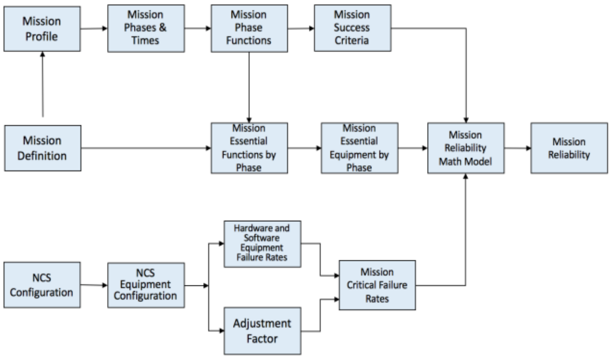

Mission Profile
Mission Phases & Times
Mission Phase Functions
Mission Success Criteria
Mission Definition
Mission Essential Functions by Phase
Mission Essential Equipment by Phase
Mission Reliability Math Model
Mission Reliability
Hardware and Software Equipment Failure Rates
NCS Configuration
NCS Equipment Configuration
Mission Critical Failure Rates
Adjustment Factor

Fig. 3. Process for NCS mission reliability analysis.

The model development and analysis of the NCS mission reliability starts with the mission definition and mission profile. The initiation of these steps in the process allows for the understating of the mission specific phases and times, functions, and success criteria of the NCS mission. In addition, these steps also characterize the various mission essential functions performed by the software and hardware equipment during each phase. The development of an NCS mission reliability mathematical model requires an understanding of the NCS equipment configuration and estimates of associated failure rates.

B. *NCS Description, CONOPS, and Mission Definition and Description*

An NCS is described as a number of constituent computer system working and associated equipment for communicating with one another in order to accomplish a defense mission over some period of time. Each of the constituent computer systems is an integral part of the overall NCS by providing essential capabilities required during a phase or phases within the mission. Each of the capabilities are provided independently as individual computer systems. However, since each of the computer systems play a key role in the success of the mission, any computer system failure may be classified as a failure of the mission of the NCS depending on the type of failure and the functional configuration.

While an NCS is described as a number of computer systems interconnected through computer networking in providing capabilities throughout the mission, there is typically one computer system that is considered to be the main system for operating and orchestrating through each of the capabilities. This command and control computer system is where an NCS user would utilize a mission staging system for collecting the results of the capability outputs and compiling the information as a final product. In the case of using an NCS during mission planning, the product would be an operational planning document that is used for a Department of Defense (DoD) operational purpose in both strategic and tactical execution. A concept of operation is described using the following notional CONOPS diagram (see Fig. 4).

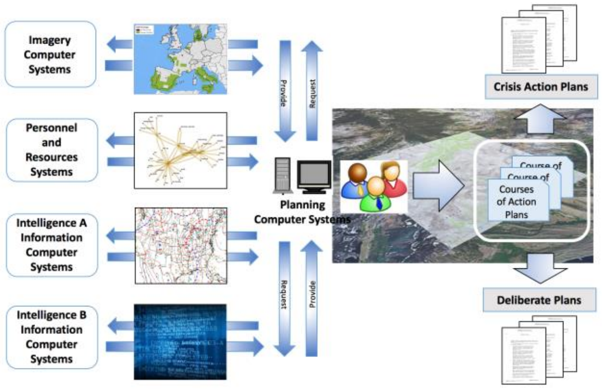

Imagery Computer Systems
Personnel and Resources Systems
Intelligence A Information Computer Systems
Intelligence B Information Computer Systems
Provide
Request
Request
Provide
Planning Computer Systems
Course of
Course of
Courses of Action Plans
Crisis Action Plans
Deliberate Plans

Fig. 4. NCS notional CONOPS.

Fig. 4 describes the characteristics of an NCS system by a user who will interact with the NCS during mission planning. This interaction with the NCS by the user is specified by the CONOPS and is the viewpoint on how the NCS will be utilized by that user. The NCS supports the DoD mission planning requirement to be responsive to any mission planning efforts given by the DoD. The NCS enhances the DoD's ability to support rapid and emerging response to capabilities necessary for crisis and deliberate planning. The NCS provides the interconnection between varying capabilities that enables mission planning to accomplish critical tasks in areas of intelligence information, operational capabilities, imagery integration, and command and control that is vital in executing search and rescue, combat readiness, physical disruption, and cyber operations. The NCS provides the means to continue the time sensitive synthesis of information allowing mission planning products to be developed with accurate information and performed in a timely manner. In order to perform mission reliability modeling and analysis of a NCS, a mission definition and description must be developed.

In defining a mission plan, the NCS mission follows the format of the Joint Operation Plan (JOP) publication process. Joint operation planning begins when an appropriate authority recognizes potential for military capability to be employed in response to a potential or actual crisis. At the strategic level, that authority—the President, Secretary of Defense, or Chairman of the Joint Chief of Staff—initiates planning by deciding to develop military options. The JOP published a process consisting of activities associated with joint military operation planning performed by military leadership in response to deliberate and crisis action activities. The publication process is a well-defined process in order to address challenging military operations and activities. The JOP provided a framework in which an NCS is developed in order to execute a defense mission plan.

Table I describes an NCS mission planning process in in leveraging the JOP.

In determining the capabilities required by the NCS, it is important to understand the mission through the mission analysis phase. During the mission analysis phase, the process determines what systems are required to be able to provide the required capabilities in planning the right mission with the best capabilities. In this phase, mission analysis is utilized to research and study the assigned tasks and their objectives along with identifying all the required tasks necessary to accomplish the mission. Therefore, the mission analysis phase clearly defines the mission, the objectives the mission is trying to satisfy during each of the phases.

In developing a generalized NCS mission, the following table (see Table II) depicts a mission characterized with i number of phases with an n number of systems. The table describes how each of the systems is utilized throughout the various phases of the mission.

The notional mission description provides a generalization of an NCS mission and provides a holistic approach in describing a process to model and analyze the NCS. Furthermore, the NCS mission can be characterized as a number of phases that are performed sequentially to satisfy the mission objectives. In order to measure the NCS mission quantitatively, one must understand when each of the computer systems will be used and at what phase(s) of the NCS mission. This method will result in a mission reliability mathematical model given an NCS mission, the essential computer system capabilities, and the phases requiring each of the capabilities over some time phased period.

The failure of the NCS to be able to perform its mission means that capability performed by one or more of the constituent system was not successful or failed to perform. In an NCS architecture, if a capability in any of the phases is not performed, then the reliability of the overall NCS mission is jeopardized. A mission description includes elements in each phase, the length of each of the phases, essential functions, and success criteria for each of the critical systems involved in the overall mission of the NCS.

TABLE I. GENERAL NCS MISSION PLANNING PROCESS

Planning Initiation	Mission Analysis	Course of Action Development	Course of Action Analysis	Course of Action Comparison	Course of Action Approval	Plan/Order Development

TABLE II. GENERAL NCS MISSION DESCRIPTION WITH I PHASES

Phase No.			**1**	**2**	**3**	**…**	**i**
Mission Phase Titles			**P1**	**P2**	**P3**	**…**	**P_i**
Phase Length			**T_1**	**T_2**	**T_3**	**…**	**T_i**
Constituent Systems of NCS	**System 1**	**S_1**	***Required***	***Required***	***Required***	**…**	***Required***
	System 2	**S_2**	***Required***	**Not Required**	**Not Required**	**…**	**Not Required**
	: **:**	**:**	**Not Required**	***Required***	**Not Required**	**…**	**Not Required**
	: **:**	**:**	**Not Required**	**Not Required**	***Required***	**…**	**Not Required**
	System N	**S_N**	**Not Required**	**Not Required**	**Not Required**	**…**	***Required***

C. *NCS Mission Profile, Mission Essential Functions by Mission Phase, Success Criteria, and Mission Essential Equipment (Hardware and Software) by Phase*

The NCS mission profile must be developed with respect to the duration of the mission, mission phases, and the duration of each of the phases. In addition, a success criteria must be defined in ensuring the mission essential functions at each of the phases. The NCS mission profile involves both hardware and software mission essential functions that must be considered individually and overall holistically. This functional dependency implies a level of coupling between software and hardware, making the two components to be highly dependent on each other. This in fact is the systems architecture typical of computer systems. Therefore, computer systems functionality should take in for account that hardware and software reliability should both be considered when quantifying the systems reliability in addressing the overall mission reliability [8]. Therefore, in this methodology application, it will be assumed that each of the computer system reliability has been calculated such that each computer system addresses both hardware and software reliability as one complete system reliability.

Based on the mission description, a mission profile is developed for the NCS mission (see Fig. 5) and is represented by phase in terms of the constituent computer systems delivering their required capabilities during the specified phased time period.

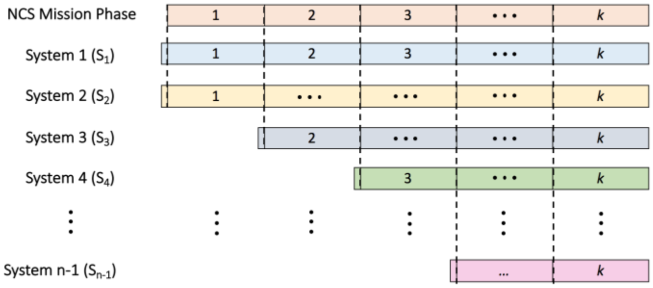
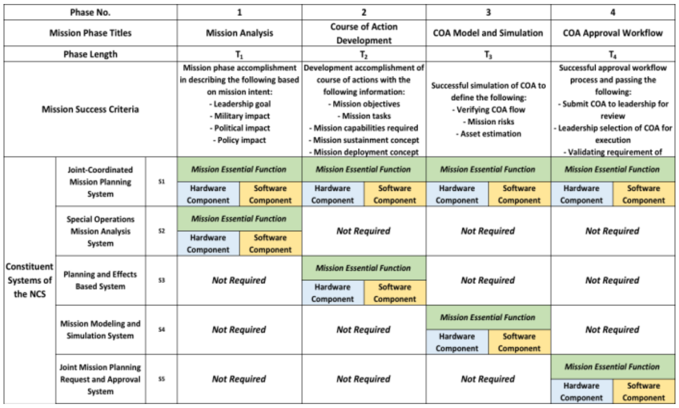


Fig. 5. Mission profile for a general NCS.

Fig. 5 depicts a mission profile of an NCS where each of the constituent systems is used in sequence to perform its function during a phase of the mission. In ensuring that each computer system is ready to provide its intended function, an initial pre-check occurs prior to the start of the mission for that system. The purpose of the system pre-check is a formal verification that the computer system is “all up”, including any redundant elements prior to the start of the mission.

In providing a specific capability, each constituent computer system is required to provide its Mission Essential Functions (MEF) during each phase of the mission in order to be successful. In an NCS mission planning scenario, the mission phases have success criteria during each of the phases, including mission analysis, course of action development, model and simulation, and approval process [9].

The success criteria describe that each of the computer systems was able to perform its capabilities during its phase. In order for these MEFs to satisfy the defined success criteria successfully, they depend on mission essential equipment of an NCS solution (see Fig. 6). In the case of an NCS, the mission essential equipment consists of both hardware and software elements. At a minimum, the software includes the operating system and the software application to provide the defined capability, whereas the hardware components consists of computer hardware elements such as a computer desktop, laptop, rack server, and tower, to name a few.



Phase No.			1	2	3	4
Mission Phase Titles			Mission Analysis	Course of Action Development	COA Model and Simulation	COA Approval Workflow
Phase Length			T_1	T_2	T_3	T_4
Mission Success Criteria			Mission phase accomplishment in describing the following based on mission intent: - Leadership goal - Military impact - Political impact - Policy impact	Development accomplishment of course of actions with the following information: - Mission objectives - Mission tasks - Mission capabilities required - Mission sustainment concept - Mission deployment concept	Successful simulation of COA to define the following: - Verifying COA flow - Mission risks - Asset estimation	Successful approval workflow process and passing the following: - Submit COA to leadership for review - Leadership selection of COA for execution - Validating requirement of
Constituent Systems of the NCS	Joint-Coordinated Mission Planning System	S1	*Mission Essential Function* Hardware Component Software Component	*Mission Essential Function* Hardware Component Software Component	*Mission Essential Function* Hardware Component Software Component	*Mission Essential Function* Hardware Component Software Component
	Special Operations Mission Analysis System	S2	*Mission Essential Function* Hardware Component Software Component	*Not Required*	*Not Required*	*Not Required*
	Planning and Effects Based System	S3	*Not Required*	*Mission Essential Function* Hardware Component Software Component	*Not Required*	*Not Required*
	Mission Modeling and Simulation System	S4	*Not Required*	*Not Required*	*Mission Essential Function* Hardware Component Software Component	*Not Required*
	Joint Mission Planning Request and Approval System	S5	*Not Required*	*Not Required*	*Not Required*	*Mission Essential Function* Hardware Component Software Component

Fig. 6. Mission phases, mission essential functions, and success criteria.

D. *Mission Reliability Block Diagram*

In developing an NCS mission reliability block diagram (RBD), the mission as a whole is analyzed in depicting mission success utilizing mission essential equipment as well as any alternate modes of operation and redundancy. Therefore, in order to properly develop a RBD, the following must be considered; system functional block diagrams, mission definition and profile, mission success criteria, and mission essential functions and equipment. The mission definition and profile along with the mission essential functions and equipment were described in the previous sections. In this section, a system functional block diagram and well defined mission are developed for use in developing the RBD.

The NCS functional block diagram describes the interrelationships between the computer systems during the mission (see Fig. 7). In this diagram, it depicts which computer system provides commands and variables to another specific computer system providing output variables in order for the next computer to accept the output variables as input variables. This interrelationship continues during each phase of the mission requiring for the computer systems to successfully perform their objectives and eventually satisfying the criteria successfully.

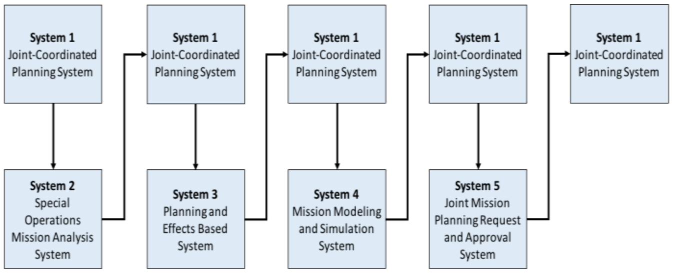

System 1 Joint-Coordinated Planning System
System 1 Joint-Coordinated Planning System
System 1 Joint-Coordinated Planning System
System 1 Joint-Coordinated Planning System
System 1 Joint-Coordinated Planning System
System 2 Special Operations Mission Analysis System
System 3 Planning and Effects Based System
System 4 Mission Modeling and Simulation System
System 5 Joint Mission Planning Request and Approval System

Fig. 7. NCS functional block diagram.

The functional block diagram depicts the interrelationship of the computer systems of the NCS and how they are connected together in being able to receive input variables and provide output variables. In order to fully depict the successful satisfaction of the objectives of the computer systems during each phase, a comprehensive success criteria must be clearly defined up front. This will provide a means to measure whether the computer system or systems were able to provide the necessary capabilities as part of the mission.

Table III presents success criteria during each phase of the mission. This table shows which systems are required for each of the phases and the success criteria to satisfy mission success. Based on the NCS configuration, each phase requires two computer systems collaboratively performing their capabilities satisfy the mission objectives.

TABLE III. NCS MISSION SUCCESS CRITERIA

System Name	System	Phase	Mssion Success Criteria
Joint-Coordinated Planning System	S1	1	Mission phase accomplishment in describing the following based on mission intent:- Leadership goal- Military impact - Political impact- Policy impact
Special Operations Mission Analysis System	S2		
Joint-Coordinated Planning System	S1	2	Development accomplishment of course of actions with the following information:- Mission objectives- Mission tasks- Mission capabilities required- Mission sustainment concept- Mission deployment concept- Mission interagency task
Planning and Effects Based System	S3		
Joint-Coordinated Planning System	S1	3	Successful simulation of COA to define the following:- Verifying COA flow- Mission risks- Asset estimation
Mission Modeling and Simulation System	S4		
Joint-Coordinated Planning System	S1	4	Successful approval workflow process and passing the following:- Submit COA to leadership for review- Leadership selection of COA for execution- Validating requirement of Leadership Intent with COA- Receive Leadership recommendation
Joint Mission Planning Request and Approval System	S5		

In understanding the mission as a whole, the system functional block diagrams, mission definition and profile, mission success criteria, and mission essential functions, mission essential hardware and software, provided inputs in developing a diagram for depicting how the computer systems contributes to the success or failure of the NCS solution [1]. The following NCS RBD is developed to better understand the mission phases, capability needs, and the notional redundancy that must be performed throughout the mission to accomplish mission success through mission reliability (see Fig. 8).

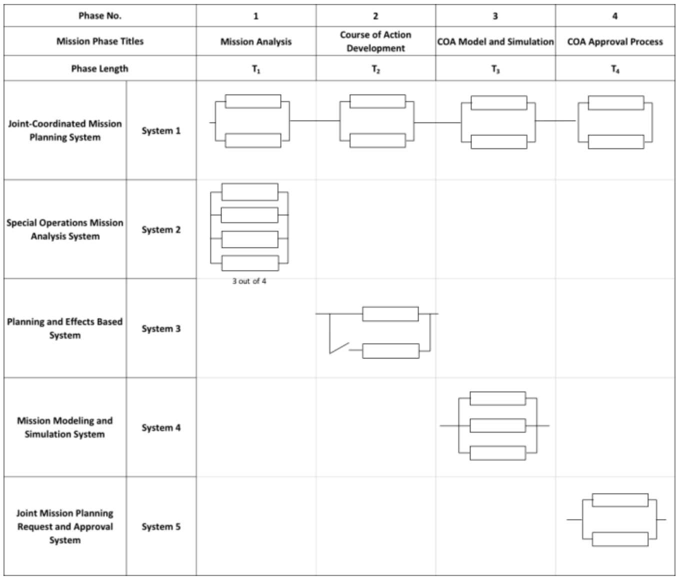

Phase No.
1
2
3
4
Mission Phase Titles
Mission Analysis
Course of Action Development
COA Model and Simulation
COA Approval Process
Phase Length
T_1
T_2
T_3
T_4
Joint-Coordinated Mission Planning System
System 1
Special Operations Mission Analysis System
System 2
3 out of 4
Planning and Effects Based System
System 3
Mission Modeling and Simulation System
System 4
Joint Mission Planning Request and Approval System
System 5

Fig. 8. NCS reliability block diagram.

E. Mission Reliability Mathematical Model

In general, mission reliability, R_m , is the conditional probability that a specific system will successfully perform its essential functions during a specified mission, given that all mission essential equipment is *up* at mission start. An NCS mission reliability mathematical model is used to evaluate system reliability in terms of:

- Probability of mission success
- Expected number of mission losses per 1,000 missions

An NCS mission reliability mathematical model describes the mathematical relationship between probability of mission success and mission critical failure rates and configuration of mission essential equipment, including software, and success criteria [4]. The development of an executable math model requires the type of time to failure probability distributions.

F. Time to Failure Probability Distribution

In developing an executable NCS mission reliability model, the model requires a time to failure probability distribution. The default probability distribution for reliability analysis of computer systems and NCSs is the Exponential Probability Distribution, i.e., the random variable Time to Failure, T, has the Exponential Probability Distribution with parameter λ . In this application, λ is the failure rate. Therefore, the probability of an NCS element surviving a period of time t during a mission is

$$R(t) = e^{-\lambda t},$$

and, λ is the mission critical failure rate.

If all NCS elements, i.e. computer systems, time between failures can be characterized by the exponential probability distribution with failure rate λ , then the NCS mission reliability model was developed from the RBD in Fig. 8.

The mission Reliability Block Diagram by mission phase is as follows (see Fig. 9):

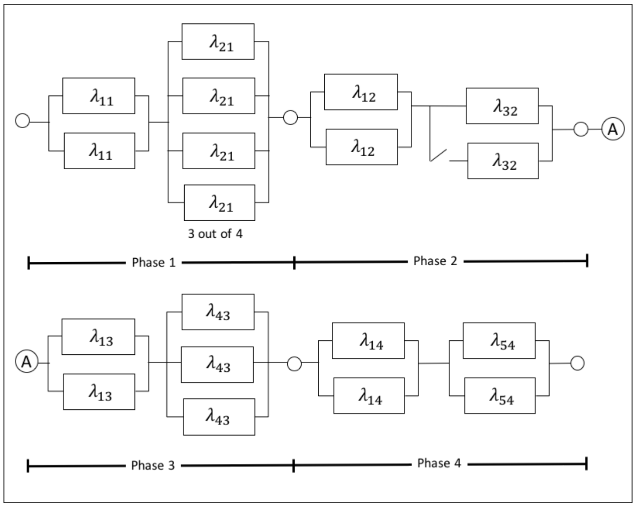


Fig. 9. NCS mission reliability block diagram.

where λ_{ij} is the mission critical failure for i=system number=1,2,3,4,5 and j= phase number =1,2,3,4.

We derived the following NCS Mission Reliability Model from the above RBD as follows:

$$R_{NCS}(t_m) = \prod_{k=1}^{4} R_k(t_k),$$

where

$R_{NCS}(t_m)$ is the probability of NCS mission success for the defined mission of duration t_m , and $R_k(t_k)$ is the probability of NCS success for mission phase t_k, for k=1,2,3,4

where

$R_1(t_1) = \left(2e^{-\lambda_{11}t_1} - e^{-2\lambda_{11}t_1}\right) * \left(4e^{-3\lambda_{21}t_1} - 3e^{-4\lambda_{21}t_1}\right),$

$R_2(t_2) = \left(2e^{-\lambda_{12}t_2} - e^{-2\lambda_{12}t_2}\right) * (1 + \lambda_{32}t_2)e^{-\lambda_{32}t_2},$

$R_3(t_3) = \left(2e^{-\lambda_{13}t_3} - e^{-2\lambda_{13}t_3}\right) * (e^{-3\lambda_{43}t_3} - 3e^{-2\lambda_{43}t_3} + 3e^{-\lambda_{43}t_3})$,

and

$R_4(t_4) = \left(2e^{-\lambda_{14}t_4} - e^{-2\lambda_{14}t_4}\right) * (2e^{-3\lambda_{54}t_4} - e^{-2\lambda_{54}t_4})$

Note that use of the above formula for NCS mission reliability overstates the true value since the per phase reliabilities are conditional probabilities of success, given that are mission essential systems are all “up” at the beginning of a phase.

The mission reliability block diagram and mission reliability model may be developed for a general NCS having a well-define mission by utilizing the above methodology.

G. Estimation of Failure Rates for Mission Essential Equipment – Hardware and Software

If the time to failure probability distribution for a NCS element, *E*, is Exponential with parameter, λ, its failure rate, then the failure rate of element *E* is estimated as follows

$$\hat{\lambda} = \frac{r_c}{t_c} \quad ,$$

where r_c is the cumulative number of mission critical failures, associated with both hardware and software, that occurred in cumulative time t_c.

H. Mission Reliability Analysis

NCS mission reliability analysis provides an estimate of the probability successfully completing a well-defined specified NCS mission, given that a successful permission NCS check is completed prior to mission start.

NCS mission reliability for a specified mission depends on the NCS reliability configuration, mission time and mission critical failure rates of mission essential hardware and software. Since the methodology that we developed is applicable to NCS solutions utilizing existing systems to the extent feasible, the failure rates of mission essential equipment must be provided by the computer system owners as estimates. And for a specified mission, the mission time, t_m , is a specified value. Therefore, the only variable in NCS mission reliability is the reliability configuration. Redundant reliability configurations, such as active parallel and standby, can be considered as alternatives in NCS mission reliability analysis. In general, use of redundancy will increase mission reliability. But with an associated penalty in $MTBF_{SR}, A_o$, and Lifecycle Cost [6].

V. NCS Sustainment Reliability Model and Analysis

System sustainment reliability, $MTBF_{SR}$, in general, is the mean time between failures, over a specified period of calendar time. Time, T, and failure, F, must be defined for a particular system type. Time, T, is often defined to be system operating time for a specified system or for all systems of that type. Failure, F, is usually defined to be an event that results in a response to an indicated malfunction. Under this definition, failure can range from “no fault found” to loss of system function due to a physical failure of a part. Failure may be associated with systems hardware, software and hardware-software interaction.

The formula for calculating $MTBF_{SR}$ depends on the type of probability distribution of the random variable T, the time between failures. The default distribution for NCSs is exponential. In which case,

$$MTBF_{SR} = \frac{1}{\lambda_{SR}}$$

then

$$\lambda_{SR} = \sum_{i=1}^{n} \lambda_i$$

where λ_i is the failure rate of the ith element of the jth computer system for j=1,2,…,k and

$$\lambda_{NCS} = \sum_{j=1}^{k} \lambda_{sys_j}$$

The sustainment reliability of the NCS is:

$$MTBF_{NCS} = \frac{1}{\lambda_{NCS}}$$

$MTBF_{SR}$ provides a mean time, a mathematical expectation of operating time between indicated malfunctions. The $MTBF_{SR}$ is a factor in calculating NCS operational availability and life cycle cost. It is not a factor in calculating mission reliability since only mission critical failures of mission essential equipment (hardware and software) may cause mission failure.

VI. NCS Operational Availability Model and Analysis

System operational availability, in general, is defined to be the likelihood or probability that the system is capable of initiating its required mission at any given point in calendar time. Operational availability is a measure of system readiness. For a NCS for mission planning, the operational availability, A_o, is defined as

$$A_o = \frac{MTBF_{SR}}{MTBF_{SR}+MDT} \quad ,$$

where $MTBF_{SR}$ is NCS sustainment reliability and MDT is the mean down time of the NCS for all causes including time associated with preventative unscheduled and scheduled maintenance, downtime due to supply, administrative downtime, and support equipment downtime.

MDT is also considered in the equation as a basic measure of maintaining and repairable items. In a case of NCS hardware requiring repair time, the hardware is either repaired or replaced which requires some down time as far as operations. Since the NCS also has software components, MDT also applies to software if required in the form of reboot, rejuvenation, or installation time requiring down time of the software. MDT is also an important consideration as average time a system is awaiting maintenance which requires time in troubleshooting, remediation, while following process and procedures to address any maintenance requirements. In using the obtained parameters, the $MTBF_{SR}$ and MDT estimates the availability of the NCS capable of its assigned mission at a given time.

Since the methodology that we developed is based on using existing systems to define candidate NCS solutions, $MTBF_{SR}$, and MDT are estimated from data provided by their owners.

VII. NCS Life Cycle Cost Model and Analysis

NCS life cycle cost is the total cost of a candidate NCS over its lifecycle as defined by DoD in Fig. 10.

Since the methodology that we have developed is applicable only to NCS development using existing computer systems to the extent feasible in meeting NCS requirements, the activity in the first four phases involves planning, architecture development, identification, of existing computer systems and definition of candidate systems, acquisition planning and integration of existing systems. This activity is generally minimal comparted to new development NCS with relative short calendar time.

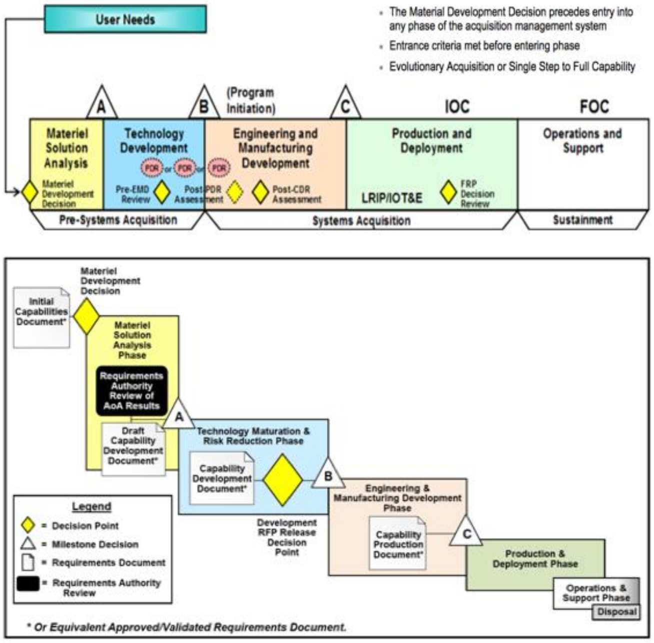

User Needs
The Material Development Decision precedes entry into any phase of the acquisition management system
Entrance criteria met before entering phase
Evolutionary Acquisition or Single Step to Full Capability
A
B
(Program Initiation)
C
IOC
FOC
Materiel Solution Analysis
Technology Development
Engineering and Manufacturing Development
Production and Deployment
Operations and Support
Materiel Development Decision
PDR
or
PDR
or
PDR
Pre-EMD Review
Post-PDR Assessment
Post-CDR Assessment
LRIP/IOT&E
FRP Decision Review
Pre-Systems Acquisition
Systems Acquisition
Sustainment
Materiel Development Decision
Initial Capabilities Document*
Materiel Solution Analysis Phase
Requirements Authority Review of AoA Results
A
Draft Capability Development Document*
Technology Maturation & Risk Reduction Phase
Capability Development Document*
B
Development RFP Release Decision Point
Engineering & Manufacturing Development Phase
Capability Production Document*
C
Production & Deployment Phase
Operations & Support Phase
Disposal
Legend
= Decision Point
= Milestone Decision
= Requirements Document
= Requirements Authority Review
* Or Equivalent Approved/Validated Requirements Document.

Fig. 10. DoD NCS life cycle.

Therefore, the first four life cycle (LC) phase is associated with the preferred NCS acquisition and its cost is nonrecurring. The NCS operations and support phase consists of two activities, namely, (1) providing the mission planning capability required and (2) supporting the NCS to sustain the required capability [5]. The cost associated with this phase is recurring.

VIII. Life Cycle Cost Model

The lifecycle cost model for a specific DoD required NCS capability is

$$C_{LCC} = C_A + C_{OS}$$

where

C_A is the NCS acquisition cost, nonrecurring, that is required (estimated) to accomplish the following activities by DoD life cycle phases (see Table IV):

and C_{OS} is the NCS operations and support cost, recurring, that is required (estimated) to provide NCS operations and the associated support cost over this DoD life cycle phase consisting of the following activities (see Table V):

The NCS support cost, C_S, depends on maintenance cost and logistics support cost, and can be estimated as follows:

$$C_S = \left(\frac{Total\ Operating\ Hours}{MTBF_{SUS}}\right)C_{MF} \quad ,$$

where

$MTBF_{SUS}$ is the NCS sustainment reliability defined in 5.4 and C_{MF} is the average cost of an indicated malfunction of the NCS, including cost of manpower, material and other logistics support resources.

TABLE IV. NCS ACQUISITION PHASES

Material Solution Analysis Phase
Analyze and define mission description and goals
Analyze and define mission objectives
Analyze and define mission requirements
Analyze contractual cost of existing computer systems capabilities
Technology Development Phase
Determine requirements can be met by existing computer systems capabilities
Develop preliminary design of the NCS
Develop configuration items from existing computer systems capabilities as initial candidates for considerations
Finalize preliminary design based on configuration items
Engineering & Manufacturing Phase
Develop master integration plan for developing an NCS
Identify candidate computer systems
Develop NCS candidate systems
Select the preferred NCS solution
Develop operations and support plan
Production & Development Phase
Develop initial working NCS prototype to ensure communication and computer system capabilities are functional across the computer network
Integrate existing computer systems to obtain the full working NCS
Perform operational test and evaluation
Acquire NCS operations and support resources

TABLE V. NCS OPERATIONS AND SUPPORT PHASES

NCS Capability Sustainment
Manage the system overall to ensure its operating and providing users with their needs
Operate the system in terms of support with a help desk center (tier 1 to tier 3 help desk engineering support)
Ensure the computer systems are working properly by replacing and upgrading equipment as required and ensure the computer network is functioning
Support the NCS in terms of required training, maintenance, and logistics support
NCS Disposal
Develop plan for NCS disassemble
Approve contractual disposal agreement
Disassemble NCS capabilities
Complete with engineering efforts in ensuring computer system capabilities are returned or disposed by system owner plan

A. *Life Cycle Cost Analysis*

An analysis is performed to determine (estimate) the life cycle cost of each candidate NCS for a specified DoD mission planning NCS [6], [9]. The life cycle cost may be used in conjunction with corresponding $MTBF_{SUS}, MR, A_o$ to characterize the cost –effectiveness for each candidate NCS [6].

NCS life cycle cost analysis utilizes input date from various sources, ranging from actual recorded costs to forecasts based on extrapolation, to obtain a “best” estimate. Since the uncertainty associated with the estimated NCS life cycle cost may be large, sensitivity analysis may be utilized to identify life cycle cost drivers and to assess the effect of variation of input data on NCS life cycle cost.

B. *Life Cycle Cost Summary*

NCS life cycle cost analysis must be performed to ensure that a candidate NCS solution meets the cost requirement for a specific DoD required capability for mission planning. In addition, life cycle cost analysis results in an estimate, that when used in conjunction with $MTBF_{SUS}, A_o$ and MR provides an estimate of the cost effectiveness of each candidate NCS solution.

IX. SUMMARY AND CONCLUSION

In the development of this paper, we discussed specific factors and their relationships with respect to NCS system cost effectiveness. These specific factors in cost effectiveness helped identify the critical elements in producing an NCS solution that could be modeled and analyzed. In addition, the research modeled specific factors in dealing with mission reliability, sustainment reliability, operational availability, and lifecycle cost that are essential drivers in producing a cost-effective NCS solution. Also, during the mission reliability modeling, the scope of mission description, profile, and success criteria, along with reliability block diagrams was produced in understanding NCS mission as whole. This process provided a complete overall picture of the NCS mission and thus providing a detailed model to be used for analysis.

The development of the mission reliability model and analysis provided an overarching perspective on the understanding of NCS mission and the factors that are important and critical for mission success. This is essential in establishing a measurement for mission success and is the contributing factor as part of the decision making process of NCS stakeholders in determining the preferred NCS in support of U.S. DoD Mission Planning missions.

REFERENCES

[1] Chelson, P. O. and Eckstein, R. E., "Reliability Computation from Reliability Block Diagrams," Jet Propulsion Laboratory, California Institute of Technology, Pasadena, California1971.

[2] Council, N. R., *Reliability Growth: Enhancing Defense System Reliability*. Washington, D.C.: National Academy Press, 2015.

[3] Defense, D. o., "ELECTRONIC RELIABILITY DESIGN HANDBOOK," D. o. Defense, Ed., MIL-HDBK-338B ed. Washington, DC, 1998.

[4] Defense, D. o., "Reliability Modeling and Prediction," vol. MIL-STD-756B, D. o. Defense, Ed., ed. Washington, DC: Department of Defense, 1981.

[5] Defense, D. o., "Reliability Program for Systems and Equipment Development and Production," vol. MIL-STD-785B, D. o. Defense, Ed., ed. Washington, DC: Department of Defense, 1980.

[6] Defense, O. o. t. S. o., "Operating and Support Cost-Estimation Guide," O. o. t. S. o. Defense, Ed., ed. Washington, DC: Cost Assessment and Program Evaluation (CAPE), 2014.

[7] Friedman, M. A., Tran, P. Y., and Goddard, P. I., "Hardware/Software System Reliability Modeling," in *Reliability of Software Intensive Systems (Advanced Computing and Telecommunications Series)*, 1st Edition ed: William Andrew, 1995.

[8] Prasad, M., Flowrence, L., and Srikrishna, C. V., "Overview of Software Reliability Models," *International Journal of Engineering and Management Research*, vol. 3, pp. 11-15, October 2013 2013.

[9] Staff, J. C. o., "Joint Operation Planning," D. o. Defense, Ed., 5-0 ed. Washington DC, 2011.

Multi Focus Image Fusion using Combined Median and Average Filter based Hybrid Stationary Wavelet Transform and Principal Component Analysis

Tian Lianfang
School of Automation Science and Engineering, South China University of Technology; Research Institute of Modern Industrial Innovation, South China University of Technology; Key Laboratory of Autonomous Systems and Network Control of Ministry of Education, Guangzhou, China

Du Qiliang
School of Automation Science and Engineering, South China University of Technology; Research Institute of Modern Industrial Innovation, South China University of Technology; Key Laboratory of Autonomous Systems and Network Control of Ministry of Education, China

Jameel Ahmed Bhutto
School of Automation Science and Engineering, South China University of Technology, Guangzhou, China

Bhawani Shankar
IEEE Senior member, Dean Faculty of Electrical, Electronic and Computer Engineering, Mehran University of Engineering and Technology, Jamshoro, Sindh, Pakistan

Saifullah Adnan
School of Electronic and Information Engineering, South China University of Technology, Guangzhou, China

***Abstract*—Poor illumination, less background contrast and blurring effects makes the medical, satellite and camera images difficult to visualize. Image fusion plays the vital role to enhance image quality by resolving the above issues and reducing the image quantity. The combination of spatial and spectral technique Discrete Wavelet Transform and Principal Component Analysis (DWT-PCA) decrease processing time and reduce number of dimensions but down sampling causes lack of shift invariance that results in poor quality final fused image. At first this work uses combined median and average filter that eliminates noise in the image which is caused by illumination, camera circuitry and sensor at preprocessing stage. Then, hybrid Stationary Wavelet Transform and Principal Component Analysis (SWT-PCA) technique is implemented to increase output image accuracy by eliminating down sampling and is not influenced by artifacts and blurring effects. Further, it can overcome the trade-off of Heisenberg's uncertainty principle by improving accuracy in both domains, time (spatial) as well as frequency (spectral). The proposed combined median and average filter with hybrid SWT-PCA algorithm measures quality parameters, such as peak signal to noise ratio (PSNR), mean squared error (MSE) and normalized cross correlation (NCC) and improved results depict the superiority of the algorithm than existing techniques.**

Keywords*—Image fusion; Heisenberg's uncertainty principle; combined median and average filter; Haar wavelet; Stationary Wavelet Transform and Principal Component Analysis (SWT-PCA)

I. Introduction

Image fusion is the process of extracting high quality, more informative single image out of multiple images by removing artifact, noise and blurring effects [1]-[3]. Image fusion is an interdisciplinary area of research and has received a lot of interest in academic, industrial, hospitals, manufacturing, robotics, military and computer vision [4], [5]. The spatial domain techniques such as PCA, averaging method have low spectral resolution whereas spectral domain such as DWT, Curve-let, SWT techniques have low spatial resolution therefore it degrades the quality of output image [6].

In [7], Moris has developed image fusion technique based on Maximum method in which blurring effects limit the contrast of fused image. Simple averaging method is used for image fusion in [8]. However, better quality image is not produced due to artifacts and noise issues. Spectral domain technique named Curve-let transform is implemented by Choi in [9]. It produces better results than discrete wavelet transform but it suffers performance degradation if image is not curved shape. A good data compression technique named PCA developed in [10] that reduces the number of dimensions, however it suffers spectral degradation. In [11], author has proposed a Discrete Wavelet transform technique that achieves better and precise results with fast computation but it suffers with spatial degradation that reduces the overall performance of output image. In [12], author has implemented the Discrete Stationary Wavelet transform that minimizes spectral degradation in comparison to DWT. The shortcoming of this technique is less spatial resolution. The DWT-PCA has been implemented in [13] to achieve better results in both domains, spatial and spectral. Though this technique achieves good quality image than existing algorithms, but the final fused image is still affected by shift-invariance due to down sampling.

In this paper, combined median and average filter at pre-processing stage has been used. It does not only reduce the complexity but also preserve the edges by reducing the image noise. At the same time, statistical histogram is applied to improve the searching speed of median value. The proposed technique achieve fast computation by PCA and high quality image by eliminating down sampling in SWT. In addition, Haar wavelet is used in our work because it has number of attractive features such as, orthogonality, compact support, and infinite support in frequency domain, symmetric in scaling function and anti-symmetric in wavelet function. These all characteristics make this work more accurate, robust and memory efficient. This proposed method for multi-focus data finds applications in various fields such as recognition of objects, feature extraction in military surveillance as well as in aeronautical observations.

The rest of the paper is arranged as follows: Image fusion techniques and proposed hybrid model has been elaborated in Sections 2 and 3, respectively. Finally simulation results and conclusion will be discussed in Sections 4 and 5.

II. Techniques of Image Fusion

Image fusion techniques are classified in two domains; spatial (time) and Spectral (frequency) which are discussed below:

A. PCA

PCA transforms the number of correlated variables into uncorrelated variables that makes it more accurate and reliable. Moreover, it reduces the number of dimensions by choosing the highest eigenvalue vector as the principle component that results in fast computation [14]. It is quantitatively rigorous method that generates new set of variables called principal components and all these components are orthogonal to each other which mitigate redundant information. PCA is widely used in many applications such as image processing, machine learning, wireless communication, pattern matching and so on [7], [14]-[16].

The flow diagram for PCA technique is shown in Fig. 1. The input images are arranged in column vectors and their resulting vector has dimension of $n \times 2$ where n is length of each vector. The eigenvectors and eigenvalues are computed for both images. Finally the principle components pc_1 and pc_2 are obtained from eigenvectors corresponding to largest eigenvalue [16], [17].

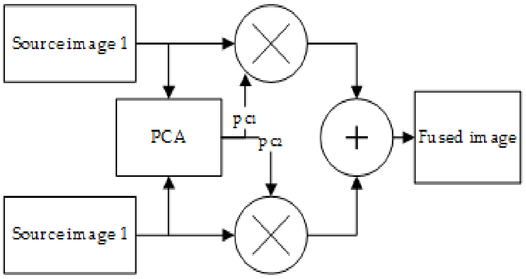


Fig. 1. Diagram of PCA.

The analysis of PCA involves five main steps:

- Produce the column vector from source images.
- Get covariance /correlation matrix A from data sets of input images. Variance and covariance are obtained from data set of input images; $x_1 \; and \; x_2$.

$$A = \begin{bmatrix} v_1 & c_{\text{var}}(1,2) \\ c_{\text{var}}(1,2) & v_2 \end{bmatrix} \tag{1}$$

$$c_{\text{var}} = \frac{\left(\sum\left(x_1 - \overline{x_1}\right) \times \left(x_2 - \overline{x_2}\right)\right)}{N-1} \tag{2}$$

v_1, v_2, $c_{\text{var}}(1,2)$ and N represents the variance, covariance and number of terms respectively.

- The Eigen values can be obtained by solving:

$$Det\left|A - \lambda I\right| = 0 \tag{3}$$

λ represents the eigenvalues.

$$\begin{bmatrix} v_1 & c_{\text{var}}(1,2) \\ c_{\text{var}}(1,2) & v_2 \end{bmatrix} - \lambda I = 0 \tag{4}$$

$$\left(\left(v_1 - \lambda\right) \times \left(v_2 - \lambda\right)\right) \times c_{\text{var}}(1,2)^2 = 0 \tag{5}$$

- Eigen vectors are solved as follow:

$$\left|A - \lambda I\right| \times \left[X\right] = 0 \tag{6}$$

- The co-ordinates of each data point in the direction of principal component are obtained by:

$$pc_j = a_{i1}Y_1 + a_{i2}Y_2 \ldots\ldots\ldots\ldots a_{in}Y_n \tag{7}$$

Where

a_i is coefficient of factor i, pc_j is the j_{th} principle component and $Y_1, Y_2 \ldots Y_n$ represents coordinates of each data.

B. DWT

It provides a compact representation of signal's frequency component by achieving better frequency information and good time resolution. It decomposes image into frequency sub-bands at different scale by splitting into high and low frequency [18]. The low frequency contains average intensity of image whereas high frequency provides edges information. Down sampling in DWT decreases the computation time and speed up the algorithm [13], [17]-[19].

It is depicted in Fig. 2, the DWT separately filters and down samples 2D image in horizontal and vertical direction. The source image is filtered by low pass (L) and high pass (H) filter in horizontal direction and down-sampled by factor 2 to get the coefficients matrices Im(L) and Im(H). Same filters with down sampling are applied to coefficient matrices in vertical direction to obtain sub-bands, Im(LL), Im(LH), Im(HL) and Im(HH) that are images with low-low, low-high, high-low and high-high frequency respectively [15], [20].

The wavelet decomposition is given by following equation:

$$G_{j+1} = HG_j H^{'} \tag{8}$$

$$D^h_{j+1} = GG_j H^{'} \tag{9}$$

$$D^v_{j+1} = HG_j G^{'} \tag{10}$$

$$D^d_{j+1} = GG_j G^{'} \tag{11}$$

Where $j = 0, 1, 2, ..., j-1$ is for decomposition level. H for low-pass filtering; G for High-pass filtering; $H^{'}$ and $G^{'}$ are conjugate of the H and G; G_{j+1} , D^h_{j+1} , D^v_{j+1} and D^d_{j+1} are approximate, horizontal, vertical and diagonal details of images respectively.

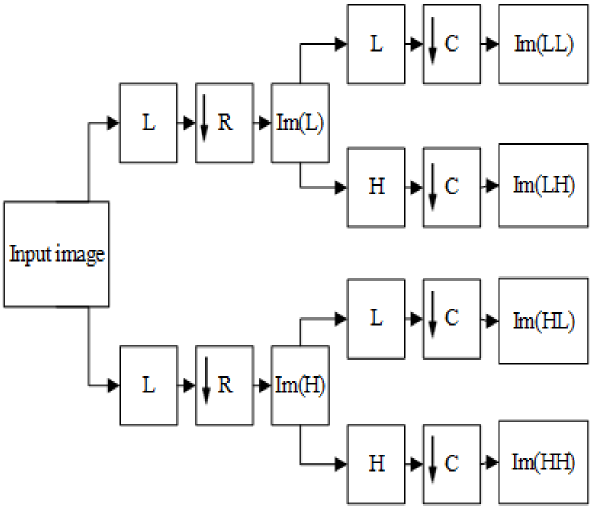


Fig. 2. Image decomposition by DWT.

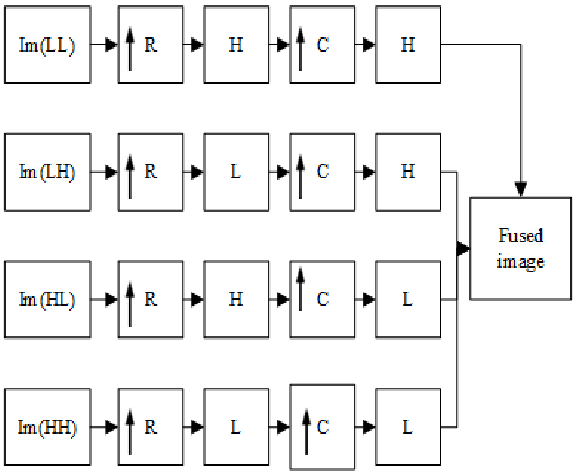


Fig. 3. Image reconstruction by DWT.

Inverse DWT is applied to reconstruct 2D image from sub-bands Im(LL), Im(LH), Im(HL) and Im(HH) as depicted in Fig. 3. This involves column up sampling the both; low and high pass filters for each sub-band images. Same filters are applied with row-up sampling to get the fused image. Image reconstruction equation is given as:

$$G_j = H^{'} G_{j+1} H + G^{'} D^h_{j+1} H + H^{'} D^v_{j+1} G + G^{'} D^d_{j+1} G \tag{12}$$

Here $j = j-1, j-2, ...0$, G_{j+1} represents low frequency component (approximate detail) and D^h_{j+1} , D^v_{j+1} and D^d_{j+1} denote high frequency components respectively (horizontal, vertical and diagonal detail).

C. SWT

Down sampling in DWT causes lack of shift-invariance that adds distortion, artifacts and blurring effects which results poor quality output image. These issues can be resolved by eliminating the down-sampling in SWT. Therefore, preserving a high quality and more informative output image [16], [21].

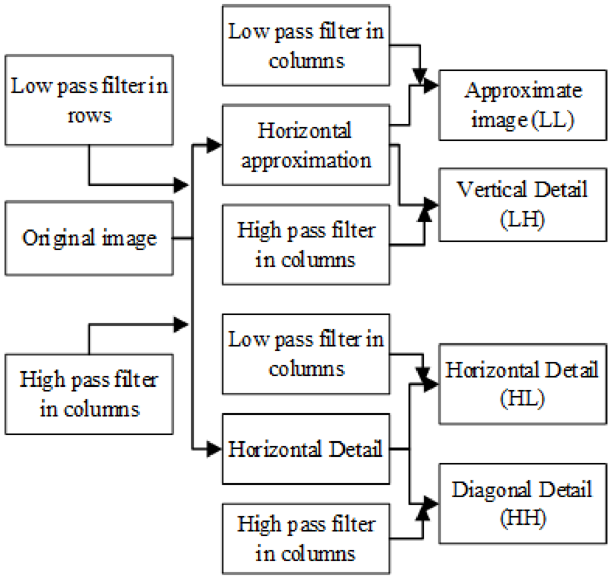


Fig. 4. Diagram of SWT.

The original image is decomposed into horizontal and vertical approximation by applying column wise and row wise low pass and high pass filters [21]. Same filtration is applied to decomposed parts in rows wise and column wise to obtain Approximate, vertical, horizontal and diagonal detail as elaborated in Fig. 4. The low pass and high pass filters preserve low and high frequencies and provides detailed information at respective frequencies.

The wavelet decomposition equation for SWT is given as:

$$A_{_j}, k_1, k_2 = \sum_{n1}\sum_{n2} h_0^{\uparrow 2j}\left(n_1 - 2k_1\right) h_0^{\uparrow 2j}\left(n_2 - 2k_2\right) A_{j-1}, n_1, n_2 \tag{13}$$

$$D^1_{\ _j}, k_1, k_2 = \sum_{n1}\sum_{n2} h_0^{\uparrow 2j}\left(n_1 - 2k_1\right) g_0^{\uparrow 2j}\left(n_2 - 2k_2\right) A_{j-1}, n_1, n_2 \tag{14}$$

$$D^{2}{}_{j},k_1,k_2=\sum_{n1}\sum_{n2}g_0^{\uparrow 2j}\left(n_1-2k_1\right)h_0^{\uparrow 2j}\left(n_2-2k_2\right)A_{j-1},n_1,n_2 \quad (15)$$

$$D^{3}{}_{j},k_1,k_2=\sum_{n1}\sum_{n2}g_0^{\uparrow 2j}\left(n_1-2k_1\right)g_0^{\uparrow 2j}\left(n_2-2k_2\right)A_{j-1},n_1,n_2 \quad (16)$$

Where A_j,k_1,k_2 ; D_j^1,k_1,k_2 ; D_j^2,k_1,k_2 and D_j^3,k_1,k_2 are low frequency (LL), horizontal high frequency (LH), vertical high frequency (HL), and diagonal component (HH) respectively for SWT. $h_0^{\uparrow 2j},g_0^{\uparrow 2j}$; indicate that 2^j-1 zeros are added between h_0 and g_0 .

$$\begin{aligned}A_{j-1},n_1,n_2=\frac{1}{4}\sum_{i=0}^{3}(&\sum_{k_1}\sum_{k_2}h_1\left(n_1-2k_1-i\right)h_1\left(n_2-2k_2-i\right)A_j,k_1,k_2\\&+\sum_{k_1}\sum_{k_2}h_1\left(n_1-2k_1-i\right)g_1\left(n_2-2k_2-i\right)D_j^1,k_1,k_2\\&+\sum_{k_1}\sum_{k_2}g_1\left(n_1-2k_1-i\right)h_1\left(n_2-2k_2-i\right)D_j^1,k_1,k_2\\&+\sum_{k_1}\sum_{k_2}g_1\left(n_1-2k_1-i\right)g_1\left(n_2-2k_2-i\right)D_j^2,k_1,k_2)\end{aligned} \quad (17)$$

The SWT can be reconstructed by (17) shown above.

D. *Combined Median and Average Filter*

The image noise is random variation that is produced by illumination, sensor and camera circuitry. It does not only add distortion in image but highly affects the visual effects. This noise can be reduced by a non-linear median filter that preserves the edges and sharpen the contrast as well. The noise reducing effects depend on the shape and size of filtering mask whereas algorithm complexity depends on searching speed to get the median value. Authors have proposed fast algorithms that improve the searching speed of median filter and reduced the complexity while preserves the edges without being affected by noise [22].

The combined median and average filter achieves better performance for noise reduction that adaptively resizes the mask filter according to level of noise mask. This technique reduces the noise and retain the better image detail by preserving the edges. We sequentially check each pixel for noise reduction. If the pixel value is greater than the average value, it means the pixel is affected by noise and we replace that pixel with median value of the mask; otherwise, we un-change that value of the pixel [22].

III. Proposed Algorithm

We have implemented a combined median and average filter based hybrid SWT-PCA which is discussed briefly.

A. *Combined Median and Average Filter based Hybrid SWT-PCA Algorithm*

The proposed method involves two steps: At preprocessing stage, combined median and average filter is applied on input images then image fusion is done by hybrid SWT-PCA. Prepressing stage involves the following steps:

Step 1: Adaptively resize the mask

- Initialize the filter by $n=3$
- Compute $A_1=med-\min;A_2=med-\max$
- Verify, if $A_1\succ 0$ and $A_2\prec 0$ then return to step 1 otherwise enlarge the mask size by $n=n+2$.

Where med , min , med , A_1 and n represent the median, minimum, maximum, average values and mask size, respectively.

Step 2: Apply median filtering to find median values.

The output images O_1 and O_2 are obtained by applying filters to the input images. Furthermore, using hybrid SWT-PCA, O_1 and O_2 images are decomposed into four sub bands LL_1,LH_1,HL_1,HH_1 and LL_2,LH_2,HL_2,HH_2 by SWT. We compute the PCA for these sub band and eigen vectors with maximum values are selected. Each sub band images are multiplied and summed to combine mentioned sub bands. The new sub-bands $LL_{new},LH_{new},HL_{new}$ and HH_{new} are calculated by following equations:

$$LL_{new}=pc_1\times LL_1+pc_2\times LL_2 \quad (18)$$

$$LH_{new}=pc_3\times LH_1+pc_4\times LH_2 \quad (19)$$

$$HL_{new}=pc_5\times HL_1+pc_6\times HL_2 \quad (20)$$

$$HH_{new}=pc_7\times HH_1+pc_8\times HH_2 \quad (21)$$

$pc_1,pc_2.......pc_8$ are principal components for sub bands, respectively.

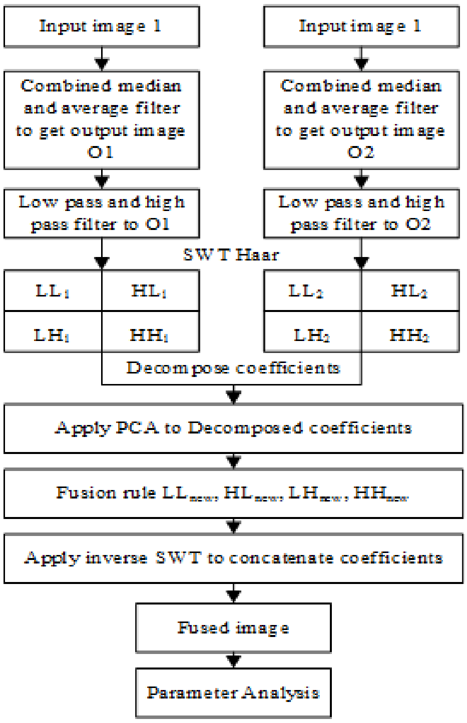

Input image 1
Input image 1
Combined median and average filter to get output image O1
Combined median and average filter to get output image O2
Low pass and high pass filter to O1
Low pass and high pass filter to O2
SWT Haar
LL1
HL1
LH1
HH1
LL2
HL2
LH2
HH2
Decompose coefficients
Apply PCA to Decomposed coefficients
Fusion rule LLnew, HLnew, LHnew, HHnew
Apply inverse SWT to concatenate coefficients
Fused image
Parameter Analysis

Fig. 5. Proposed SWT-PCA algorithm.

The proposed hybrid SWT-PCA algorithm is depicted in Fig. 5. Combined median and average filter is applied to input images. The output of this filter is feed to low pass and high pass filters in row wise and column wise. The decomposed sub-bands LL_1, LH_1, HL_1, HH_1 and LL_2, LH_2, HL_2, HH_2 are obtained by applying the Haar wavelet. SWT does not use down sampling that makes the fusion output stable and consistent with original input sequence. Resulting coefficients are evaluated by PCA and each coefficient of source image is applied to corresponding coefficient of PCA. Fusion rule merges the PCA components with new decomposed coefficient ($LL_{new}, LH_{new}, HL_{new}$ and HH_{new}). Finally, more informative and high quality fused image is obtained by applying inverse SWT on coefficients of combined transform.

The flowchart of combined median and average filter based SWT-PCA technique is shown in Fig. 6.

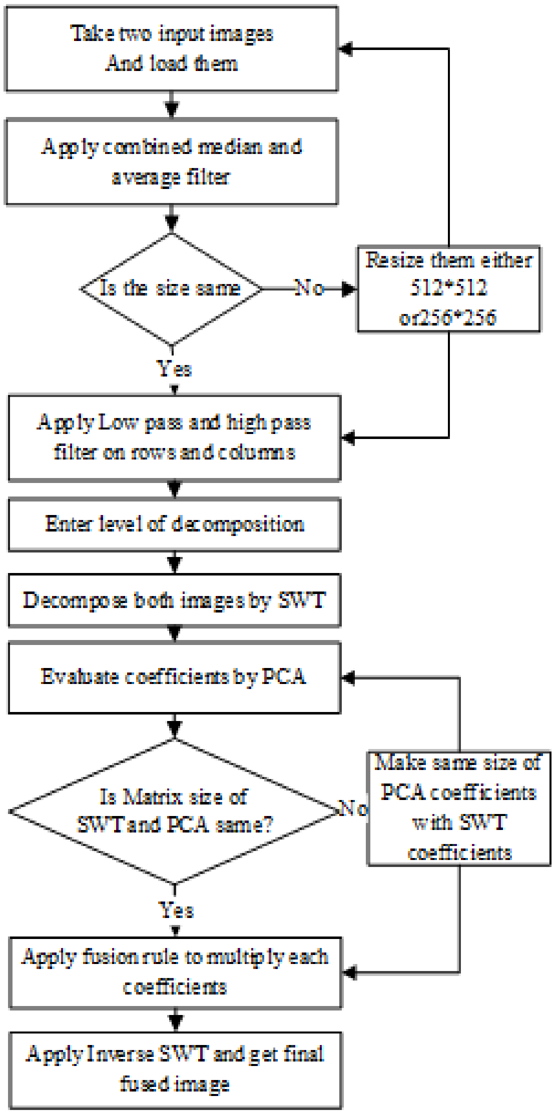

Take two input images And load them
Apply combined median and average filter
Is the size same
No
Resize them either 512*512 or256*256
Yes
Apply Low pass and high pass filter on rows and columns
Enter level of decomposition
Decompose both images by SWT
Evaluate coefficients by PCA
Is Matrix size of SWT and PCA same?
No
Make same size of PCA coefficients with SWT coefficients
Yes
Apply fusion rule to multiply each coefficients
Apply Inverse SWT and get final fused image

Fig. 6. Flowchart of proposed algorithm.

IV. Simulation Results and Discussion

In this paper, we have used MATLAB tool for our simulation and the images have been taken from the website http://imagefusion.org and paper [5].

A. Simulation Results

The simulation results given below show the comparison among DWT, SWT, DWT-PCA and combined median and average filter based SWT-PCA and their performance is measured on different quality parameters; PSNR, MSE and NCC.

1) Scenario-01

It can be seen from Fig. 7 that source images with different foci are combined by different image fusion techniques to produce better output image. Though DWT-PCA achieves better results than individual DWT and SWT but it is still affected by random variation in image. The proposed SWT-PCA with combined median and average filter achieves better results than all existing fusion techniques without being affected by noise or random variation. Furthermore, it can be analyzed from Table I that there is only a little improvement of PSNR among DWT, SWT and DWT-PCA. Consequently, in proposed technique, PSNR jumps from 40.321 to 45.154 that shows its superiority.

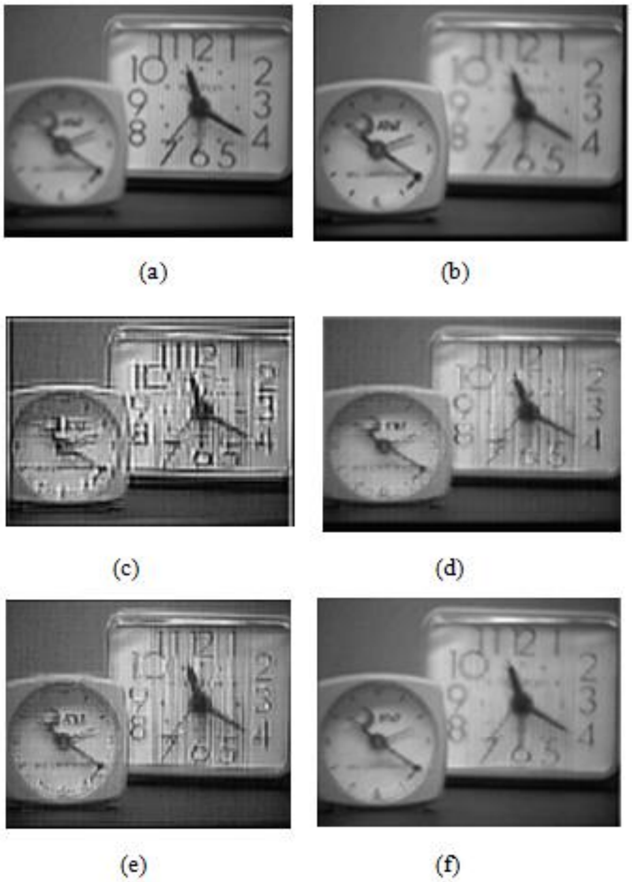

(a) (b) (c) (d) (e) (f)

Fig. 7. The input focused images and output fused images: (a) Right focused input image; (b) Left focused input image; (c) DWT fused image; (d) SWT fused image; (e) DWT-PCA fused image; (f) Proposed method (SWT-PCA fused image).

TABLE I. COMPARISON ANALYSIS OF DIFFERENT IMAGE FUSION TECHNIQUES

	DWT	SWT	DWT-PCA	Proposed method
PSNR	36.732	38.297	40.321	45.154
MSE	3.83	3.76	3.42	3.02
NCC	0.35	0.41	0.47	0.57

2) Scenario-02

Background focused and rope focused images are shown in Fig. 8. It can be clearly seen from Table II that PSNR of proposed technique jumps from 34.321 to 40.154 that depicts significant improvement. Similarly, MSE of proposed method reduces from 3.73 to 3.22 which shows better improvement as compared to existing techniques.

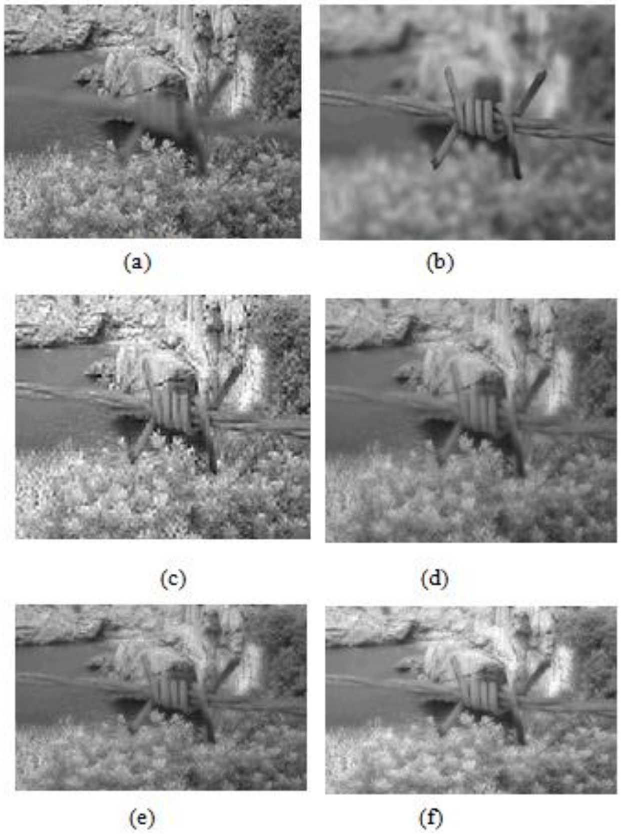


Fig. 8. The input focused images and output fused images: (a) Background focused input image; (b) Rope focused input image; (c) DWT fused image; (d) SWT fused image; (e) DWT-PCA fused image; (f) Proposed method (SWT-PCA fused image).

TABLE II. COMPARISON ANALYSIS OF DIFFERENT IMAGE FUSION TECHNIQUES

	DWT	SWT	DWT-PCA	Proposed method
PSNR	31.432	32.597	34.321	40.154
MSE	4.24	4.01	3.73	3.22
NCC	0.28	0.37	0.43	0.54

3) Scenario-03

It is shown in Fig. 9 that output fused image of DWT, SWT, DWT-PCA produce poor quality output image, whereas proposed method (SWT-PCA with combined median and average filter) eliminates the noise effect and produce better quality output image. Furthermore, it can be analyzed from Table III that there is less PSNR and NCC improvement in existing DWT, SWT and DWT-PCA. The proposed method achieves better improvement in PSNR and NCC with more reduction in MSE value. This shows the effectiveness of combined median and average filter with SWT and PCA.

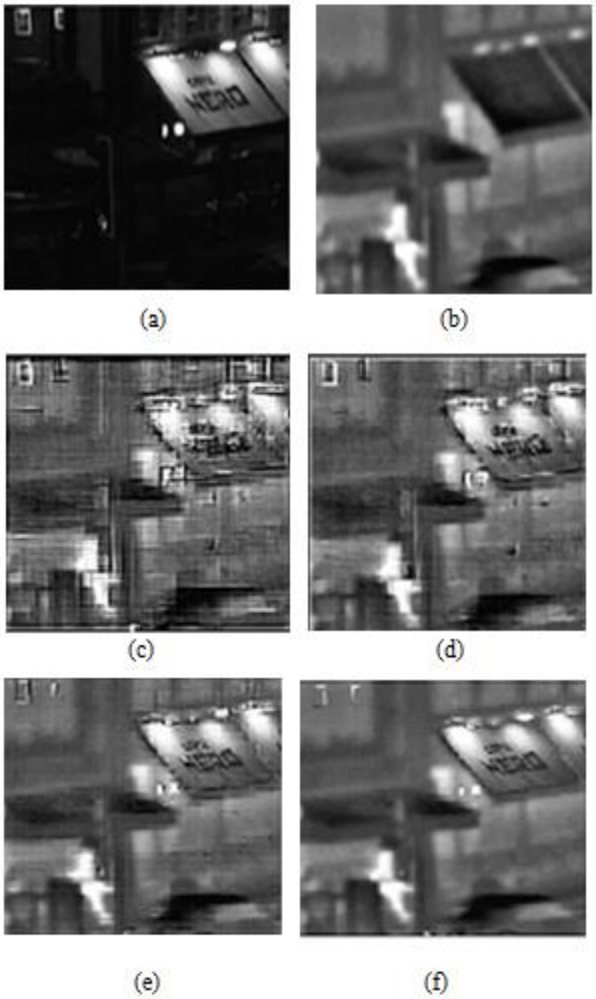


Fig. 9. The input focused images and output fused images: (a) Character focused input image; (b) Background focused input image; (c) DWT fused image; (d) SWT fused image; (e) DWT-PCA fused image; (f) Proposed method (SWT-PCA fused image).

TABLE III. COMPARISON ANALYSIS OF DIFFERENT IMAGE FUSION TECHNIQUES

	DWT	SWT	DWT-PCA	Proposed method
PSNR	35.621	36.78	38.21	44.10
MSE	3.21	3.04	2.87	2.14
NCC	0.43	0.51	0.56	0.68

4) Scenario-04

Fig. 10 shows that input images are affected by noise due to camera circuitry. The output fused image of DWT, DSWT, DWT-PCA is of poor quality due to noise. Proposed method eliminates the noise effect, produces better quality output image and achieves higher PSNR and NCC with less MSE than existing fusion techniques as shown in Table IV.

PSNR, MSE and NCC have been used as evaluation criteria in this paper. Higher the value of PSNR and NCC; better will be the final fused image. Similarly lower the MSE value corresponds to better output image. It can be clearly seen from above results that proposed technique (combined median and average filter based SWT-PCA) outperforms than current techniques by achieving lower MSE, higher PSNR and NCC values. Furthermore, it can be depicted from above figures that our proposed method remove the noise from input by combined median and average filter and obtain better accuracy by hybrid SWT-PCA.

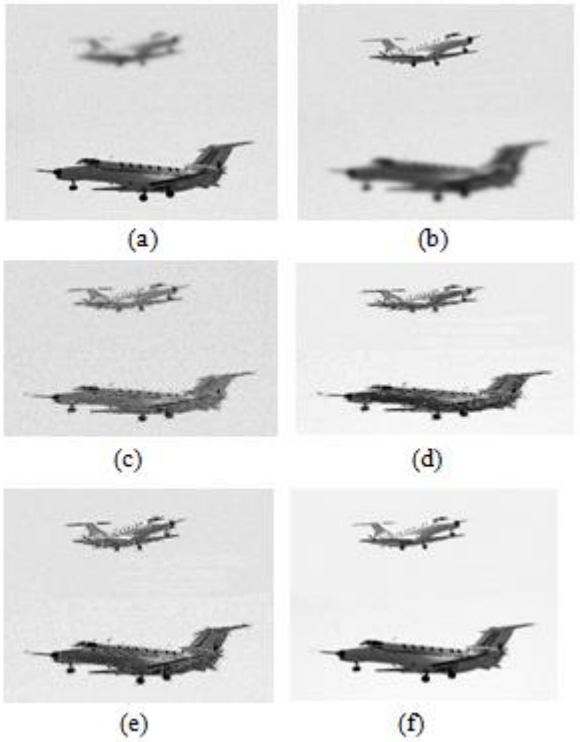


Fig. 10. The input focused images and output fused images: (a) Bottom focused input image; (b) Top focused input image; (c) DWT Fused image; (d) SWT fused image; (e) DWT-PCA fused image; (f) Proposed method (SWT-PCA Fused image).

TABLE IV. COMPARISON ANALYSIS OF DIFFERENT IMAGE FUSION TECHNIQUES

	DWT	SWT	DWT-PCA	Proposed method
PSNR	32.61	35.38	37.15	44.24
MSE	4.53	4.21	4.02	3.26
NCC	0.31	0.39	0.44	0.51

V. CONCLUSION

Recently, image fusion techniques are being considered as the most prominent and has been used in many areas such as images for medical diagnosis, military and law enforcement, robotics, manufacturing, computer vision and so on. A lot of research has been carried out regarding improving the image quality but it suffers with issues like noise, artifacts, blurring effects, lack of shift-invariance and Heisenberg uncertainty principle. However, the simulation results of proposed algorithm for multi-focus images resolve the mentioned issues and evaluate the performance on the basis of PSNR, MSE and NCC. Lower values of MSE, higher values of PSNR and NCC show the superiority of proposed algorithm than existing conventional and combined DWT-PCA techniques. Combined median and average filtering with Haar wavelet in SWT-PCA excludes down sampling that makes the hybrid algorithm shift invariant and robust while preserving the image data by reducing noise and complexity. Triple modality is also hot research topic for future work and this idea can be implemented for opaque and semi-transparent images.

ACKNOWLEDGMENT

This research work is supported by:

The Guangzhou Research Project “Development of Intelligent Escalator System with Automatic Identification Ability” (201604010114).

2016 Provincial Specialized Fund for Frontier and Key Technical Innovation “R&D and Application Demonstration of Fully Automated Equipping Production Line after Aluminum Extrusion Process” (2016B090912001).

National Ministry of Science and Technology Hai Phong Public Welfare Project “Sea Vessel Star-Machine-Island Stereoscopic Surveillance and Monitoring Technology System” (201505002).

REFERENCES

[1] A. Dogra, B. Goyal, and S. Agrawal, "From multi-scale decomposition to non-multi-scale decomposition methods: A comprehensive survey of image fusion techniques and its applications," IEEE Access, vol. 5, pp. 16040-16067, 2017.

[2] J. Luo and W. Kong, "The Infrared and Visible Light Image Fusion Based on the Non-subsample Shearlet Transform and Heat Source Concentration Ratio," in 2016 International Conference on Intelligent Networking and Collaborative Systems (INCoS), 2016, pp. 544-547.

[3] L. N. H. Hoa, L. D. Cuong, and L. C. Ke, "Enhanced spatial resolution for VNREDSat-1 multispectral images using IHS fusion technique based on sensor spectral response function," in 2016 Eighth International Conference on Knowledge and Systems Engineering (KSE), 2016, pp. 304-308.

[4] A. Galande and R. Patil, "The art of medical image fusion: A survey," in 2013 International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2013, pp. 400-405.

[5] L. Jian, X. Yang, Z. Zhou, K. Zhou, and K. Liu, "Multi-scale image fusion through rolling guidance filter," Future Generation Computer Systems, vol. 83, pp. 310-325, 2018/06/01/ 2018.

[6] P. Kaur and M. Kaur, "A comparative study of various digital image fusion techniques: A review," International Journal of Computer Applications, vol. 114, 2015.

[7] C. Morris and R. S. Rajesh, "Two Stage Spatial Domain Image Fusion Techniques," ICTACT Journal On Image And Video Processing: Special Issue On Video Processing For Multimedia Systems, vol. 5, 2014.

[8] A. Malviya and S. G. Bhirud, "Image fusion of digital images," Entropy, vol. 7, pp. 7-4955, 2009.

[9] M. Choi, R. Y. Kim, and M.-G. Kim, "The curvelet transform for image fusion," International Society for Photogrammetry and Remote Sensing, ISPRS 2004, vol. 35, pp. 59-64, 2004.

[10] A. Purushotham, G. U. Rani, and S. Naik, "Image fusion using DWT & PCA," International Journal of Advanced Research in Computer Science and Software Engineering, vol. 5, pp. 800-4, 2015.

[11] D. Mistry and A. Banerjee, "Discrete wavelet transform using matlab," International Journal Of Computer Engineering & Technology, vol. 4, pp. 252-259, 2013.

[12] M. Pradnya and S. D. Ruikar, "Image fusion based on stationary wavelet transform," International journal of Advanced Engineering Research and Studies, II/IV/July-Sept, pp. 99-101, 2013.

[13] S. Mane and S. D. Sawant, "Image fusion of CT/MRI using DWT, PCA methods and analog DSP processor," Int. Journal of Engineering Research and Applications, vol. 4, pp. 557-563, 2014.

[14] M. S. A. M. a. N. K. a. T. A. a. M. A. a. F. R. a. M. Awais, "Pakistan Sign Language Detection using PCA and KNN," International Journal of Advanced Computer Science and Applications, vol. 9, 2018.

[15] S. A. Panwar and S. Malwadkar, "A Review: Image Fusion Techniques for Multisensor Images," ed: IJAREEIE, 2015.

[16] S. Baraiya and L. P. Gagnani, "An introduction of image fusion techniques," International Journal for Innovative Research in Science and Technology, vol. 1, pp. 86-89, 2014.

[17] S. S. Bedi, J. Agarwal, and P. Agarwal, "Image fusion techniques and quality assessment parameters for clinical diagnosis: a review," International journal of advanced research in computer and communication engineering, vol. 2, pp. 2319-5940, 2013.

[18] V. T. H. T. a. N. T. Binh, "Object Contour in Low Quality Medical Images in Curvelet Domain," International Journal of Advanced Computer Science and Applications, vol. 9, 2018.

[19] X. Zhou, H. Zhang, and C. Wang, "A Robust Image Watermarking Technique Based on DWT, APDCBT, and SVD," Symmetry, vol. 10, 2018.

[20] K. Harpreet and R. Rachna, "A combined approach using DWT & PCA on image fusion," Int J Adv Res Comput Commun Eng, vol. 4, pp. 294-296, 2015.

[21] S. K. Shah and D. U. Shah, "Comparative study of image fusion techniques based on spatial and transform domain," International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET), vol. 3, 2014.

[22] Y. Zhu and C. Huang, "An Improved Median Filtering Algorithm for Image Noise Reduction," Physics Procedia, vol. 25, pp. 609-616, 2012/01/01/ 2012.

Study of Face Recognition Techniques: A Survey

Madan Lal, Kamlesh Kumar
Department of Computer Science
Sindh Madressatul Islam University,
Karachi, Sindh, Pakistan

Rafaqat Hussain Arain, Abdullah Maitlo,
Sadaquat Ali Ruk, Hidayatullah Shaikh
Department of Computer Science, Shah Abdul University,
Khairpur, Sindh, Pakistan

Abstract—**With the rapid growth in multimedia contents, among such content face recognition has got much attention especially in past few years. Face as an object consists of distinct features for detection; therefore, it remains most challenging research area for scholars in the field of computer vision and image processing. In this survey paper, we have tried to address most endeavoring face features such as pose invariance, aging, illuminations and partial occlusion. They are considered to be indispensable factors in face recognition system when realized over facial images. This paper also studies state of the art face detection techniques, approaches, viz. Eigen face, Artificial Neural Networks (ANN), Support Vector Machines (SVM), Principal Component Analysis (PCA), Independent Component Analysis (ICA), Gabor Wavelets, Elastic Bunch Graph Matching, 3D morphable Model and Hidden Markov Models. In addition to the aforementioned works, we have mentioned different testing face databases which include AT & T (ORL), AR, FERET, LFW, YTF, and Yale, respectively for results analysis. However, aim of this research is to provide comprehensive literature review over face recognition along with its applications. And after in depth discussion, some of the major findings are given in conclusion.**

Keywords—Face recognition; illuminations; partial occlusion; pose invariance

I. Introduction

The 21st century is a modern and scientific era in which a lot of progress has been achieved as to expedite humans for accomplishing their tasks. In support of above statement, nowadays use of computer technology has been an integral part of life. Computers are being used in pyramids of applications, which range from simple to complex problem solving methods. Among such contributions face recognition technology has emerged as useful tool to recognize features of faces through their inherent traits. And it has been one of the most researched areas in the field of pattern recognition and computer vision. However, due to its wide use in multitude of applications such as in biometrics, information security, law enforcement access control, surveillance system and smart cards. But it possesses many challenges for researcher that needs to be addressed. Face an object depends on facial expressions, which constitute meaningful features. For instance, pose invariance, illuminations and aging which are potential areas that require further investigation over previous work. The result of previous researches reveals that facial expressions are changing with respect to aging; therefore, they could not be permanently modeled in face recognition. The face recognition problem can be categorized into two main phases: 1) face verification and 2) face identification. For example, in real time system, face verification identifies the same person in the scene, and face identification who is this person in that scene. In the first phase it locates a face in an image. Similarly, in the second stage, it extracts features from an image for discrimination. After that they are matched with face database images in order to recognize correct face image as shown in Fig. 1. However, some existing recognition methods for authentication suffer from lack of reliability. For instance, smart cards, wallets, keys, tokens use PINs and passwords which are very difficult to remember. In addition, these passwords and codes can easily be forgotten; also these magnetic cards can be misplaced, even robbed or reproduced. As a result this makes them illegible. Contrary to biological characteristics and traits of an individual person they cannot be stolen, forgotten or misplaced [1].

Biometric recognition system can be built through various techniques. However, most commonly used are fingertips and iris methods. These require individual’s participation or involvement to access the system. Moreover, latest systems provide participant access without its intervention. Among such methods face recognition is one of the most viable technique in which an individual can easily be captured and monitored through the system. Face recognition databases vary with controllable photos to uncontrollable videos, where LFT is used for controllable photos and YTM is used for videos. Face recognition system comprises of three main modules: 1) pre-processing, 2) feature selection, and 3) Classification.

Human beings possess natural ability to recognize hundreds of faces by their visual system and cognition. That makes them recognize familiar faces even after a long period of time. Building an intelligent system similar to human perception system is still an active area of research. The researchers have suggested numerous algorithms and methodologies for recognizing a face in an effective and efficient manner. For this purpose, they have focused on detection and recognition of traits and features for individuals such as nose, eyes, mouth, face shape position, size, and beside relationship among traits and features. Furthermore, ongoing research in face recognition tries to develop such systems that could work well in an effective and efficient manner in multitude of real-world applications. In addition, many scholars have proposed and concluded that the accuracy of face recognition can further be improved through the use of 3D faces [6]. The rest of this paper is arranged as follows: Section II describes challenges related to the face recognition. Section III explains face recognition databases. Section IV elaborates methods and techniques in face recognition. Section V illustrates widely used applications in face recognition. Finally, Section VI gives conclusion and future directions.

II. Challenging Areas in Face Recognition

A. Aging

Aging is an inevitable natural process during the lifetime of a person as compared to other facial variations. Aging effect can be observed under main three unique characteristics:

1) The aging is uncontrollable: It cannot be advanced or even delayed and it is slow and irreversible.

2) Personalized aging Signs: Every human passes through different aging patterns. And these rely on his or her genes and many other factors, such as health, food, region, and weather conditions.

3) The aging signs depend on time: The face of person at a specific age will affect all older faces, but unaffected in younger age.

B. Partial Occlusion

Occlusion refers to natural or artificial obstacles in an image. It can be a local region of the face along with different objects such as sunglasses, scarf, hands, and hair. They are generally called partial occlusions. Partial occlusions correspond to any occluding object. And the occlusion less than 50% of the face is considered to be a partial occlusion. The approaches to face recognition with partial occlusion are classified into following three categories: 1) Part Based Methods, 2) Feature based methods and 3) Fractal-Based Methods [4]. Many areas of image processing have been impacted by partial occlusion such as recognition by ear is occluded due to earrings. Occlusion affects the performance of a system when people deceive it either by the use of sunglasses, scarves, veil or by placing mobile phones or hands in front of faces. In some cases, other factors like shadows due to extreme illumination also act as occluding factors. Further, local approaches are used to deal with the problem of partially occluded faces which divide the faces into different parts [5]. However, this problem can be overcome by eliminating some of the features which create trouble while accurate recognition in the image. Mostly local methods are based on feature analysis, in which best possible features are detected and then they are combined. Another approach that can be applied for this purpose is near holistic approach in which occlude features, traits and characters are eradicated and rest of the face is used as valuable information. Different techniques are being developed by the researchers to cope up with this problem [7], [8].

C. Pose Invariance

Pose variance is yet another hurdle in achieving a successful face recognition system. People pose differently every time they take a picture. There is no standardized rule for taking a pose. Therefore, it makes more difficult to distinguish and recognize the faces from images with varying poses. Pose variations degrade the performance of the facial features. In addition, many systems work under inflexible imaging conditions and as a result it affects the quality of gallery images. The methods dealing with variation in pose can be divided into two kinds i.e. multi-view face recognition and face recognition across pose. Multi-view face recognition can be considered as an annexure of frontal face recognition in which gallery image of every pose is considered. On the other hand, across a pose in face recognition, yield face with a pose which has never been exposed before to a recognition system. A good face recognition approach should provide good pose tolerance and capability to recognize different poses. Several issues in this regard are still open such as lack of perceptive subspace pose variant images. And many of research have been devoted to deal with this issue [9]-[13]. However, none of them has achieved 100% accuracy yet. There are some other methods and approaches that are being used to tackle similar problem of face recognition. Furthermore variance and changes in pose can be divided into three classes, namely: 1) general algorithms, 2) two dimensional methods for face recognition, and 3) three dimensional models [14].

D. Illuminations

Illumination is an observable property and effect of light. It may also refer to lightning effect or the use of light sources. Global illuminations are algorithms which have been used in 3D computer graphics. Illumination variation also badly affects the face recognition system. Thus it has been turned an area of attention for many researchers. However, it becomes tedious task to recognize one or more persons from still or video images. But it can be quite easy to extract desired information from images when they are taken under a controlled environment along with uniform background. Also there are three methods that can be implemented to deal with illumination problem. They are gradient, gray level and face reflection field estimation techniques. Gray level transformation technique carries out in-depth mapping with a non-linear or linear function. Gradient extraction approaches are used to extract edges of an image in gray level. As illumination is a factor that heavily affects the performance of recognition system obtained via face images or videos. These techniques are developed to suppress the effect of illumination [9]-[13].

III. Face Recognition Databases

This database of faces was previously called The ORL Database of Faces and it has a set of face images taken at the AT & T lab. This database was used for face recognition project which was carried out with the support of Speech, Vision and Robotics Group of the Cambridge University, Department of Engineering. It includes 10 different images each having 40 diverse subjects. However, for some subjects, the images were acquired under various conditions, i.e. variable light, facial expressions: smiling/sad, open/closed eyes and facial details (glasses/without glasses). These images were obtained with a dark consistent background having position in upright, frontal. Details of the databases used in face recognition are described as under.

A. AR Database

AR database was created by Computer Vision Center (CVC), University of Alabama at Birmingham. It comprises of over 4000 colour images of 126 people’s faces. And they are divided into 70, 56 man women, respectively. Images feature frontal view faces with different facial expressions, illumination conditions and occlusions (sun glasses, hair styles and scarves). The images of a single person were collected on two different days with a difference of 14 days. This database is available online and can be accessed without any cost for research and academic purposes.

B. FERET Database

The **FERET** database is being used in facial recognition for system evaluation. The Face Recognition Technology (FERET) program is executed by joint collaboration between the Defense Advanced Research Projects Agency (DARPA) and the National Institute of Standards and Technology (NIST). DARPA released a high-resolution, 24-bit color version of these images in 2003. And it was tested over 2,413 still face images, representing 856 individuals. However, the main motive behind development of FERET database was to facilitate algorithm development and evaluation. Thus initially, it requires a common database of facial images in order to develop and test for the purpose evaluation. After that, complications in image mentioned by the images should enhance.

C. LFW Database

Labeled Faces in the Wild (LFW) is a database of face photographs which was mainly developed for the comprehension of unconstrained face recognition problem. The data set has over 13,000 images of faces obtained via web. And each face is labeled with the name of the person whose picture was captured. However, roughly 1680 of the pictured people contained two or more distinct photos in the data set. However, the main constraint on these face images is that they were detected by the Viola Jones face detector. They are organized into four sets of LFW images, in which one is an original and other three are different types of aligned images. The aligned images have "funneled images" LFW-a, which includes an unpublished method of alignment, and "deep funneled" images. Among them, LFW-a, and the deep funneled images provide higher results over most of the face verification algorithms for original images and funneled images.

D. YouTube Face Database (YTF)

YTF database consists of face videos which were developed for unconstrained face recognition. In this database shortest and longest clips are 48 frames and 6,070 frames respectively. And the average length of a video clip is 181.3 frames. And all the videos were taken from YouTube. An average of 2.15 videos is available for different subjects.

E. Yale Database

The Yale Face Database includes 165 grayscale images in Graphical interchange Format (GIF) of 15 individuals. They are divided into 11 images/subject, each having different facial expression or configuration: center-light, with/glasses, no glasses, happy, sad, sleepy, normal, surprised, and wink. Left-light, right-light.

Yale face database is available in two volumes Yale face A known as yalefaces and Extended Yale face database B. In this database there are 15 different subjects (14 males and 01 female). These comprises of different conditions in facial images such as variations in an expression like sad or normal and happy, etc. This also depends upon other lighting conditions which consists of left, right or center light, and picture having glasses and non-glasses were included. Extended Yale face database is a dataset of 2414 images of 38 subjects. No variation in expression and no occlusions are found in the images but more focus is on extracting feature apt to illumination and they are available in cropped version.

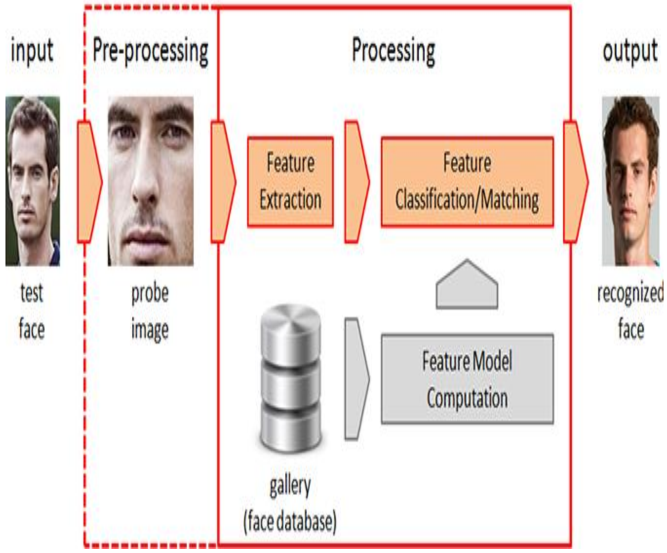


Fig. 1. Preprocessing steps of face recognition.

IV. Methods and Techniques of Face Recognition

A. Eigen Faces

The word eigenface coined by German “Eigen wert” The “Eigen” literally mean characteristic and “wert” mean value. Eigenface is well establish algorithms that was used to recognize a feature in a face image. It is based on Principle Component Analysis (PCA) [14]. In this method the fundamental concept is to recognize the face by taking its unique information about the face in question. Then encode it to compare with the decode result of previously taken image as shown in Fig. 2. In eigenface method, decoding is performed with the calculation of eigenvector and then it is represented as a matrix. However, Eigenface based face recognition systems is only suitable for images having the frontal faces but some researches identify a face with different poses have also been made [1]. Analyzing different results drawn from the researchers the accuracy ratio has been much improved in recent years as compared previous results. It is expected to have an effective and efficient output in upcoming years. A comparative study of the analysis results obtained by different researchers by applying face recognition techniques on the basis of Eigen Faces is given in Table I.

TABLE I. COMPARATIVE STUDY OF FACE RECOGNITION TECHNIQUES BASED ON PCA

S #	Year	Database	Technique	Accuracy	Reference
1	2012	ORL Faces	PCA	70.0 %	Slavković et al. [20]
2	2012	Face94	PCA	100.0 %	Abdullah et al. [33]
3	2013	FRAV Face DB	Eigen Face	96.0 %	Saha, Rajib et al. [51]
4	2014	-	PCA Eigen Faces	70.0 %	Rahman, ArmanadurniAbd, et al. [21]
5	2014	Yale Database	PCA	92% to 93%	MuzammilAbdulrahman et al. [37]
6	2014	AT & T	PCA		Johannes Reschke et al. [36]
7	2016	Computer Vision Research Projects dataset	PCA	93.6 %	Md. Al-Amin Bhuiyan [34]
8	2017	EmguCV library	PCA + RMF	93.0 %	Jacky Efendi et al. [35]
9	2017	Yale Database	PCA	98.18	Riddhi A. & S.M. Shah [46]

B. Artificial Neural Networks (ANN)

ANN provides an effective feature recognition technique, and it has been widely used after emergence of Artificial Intelligence. This consists of network, where neurons are arranged in the form of layers. Accuracy of face recognition has been boosted with the aid of better deep network architectures and supervisory methods. And recently few remarkable face representation learning techniques are evolved [15]. Using these techniques, deep learning (Fig. 3) has got much closer to human performance. For evaluation LFW face verification dataset has been used on tightly cropped face images [15]. However, the learned face representation could also add significant intrapersonal variations. One of the most viable feature of Neural Networks is it lessens the complexity. It learns from the training samples and then works fine on the images with changes in lighting conditions and increases accuracy [1]. The main drawback of the neural network is a more time is needed for its training. Initially Training is precursor step to get the desired results from the system as user point of view.

After feature extraction, classifiers for face recognition such as the Radial Basis Function and Feed Forward Neural Network (FFNN) are the implemented. Moreover, study reveals that ANNs achieves improvement over face recognition [2]. The following comparative study in Table II shows an accuracy ratio obtained through the use of ANNs.

TABLE II. COMPARATIVE STUDY OF FACE RECOGNITION METHODS USING ARTIFICIAL NEURAL NETWORS(ANN)

S #	Year	Database	Technique	Accuracy	Reference
1	2012	IIT-Dehli Database	NN Based SOM for Face recognition	88.25% to 98.3%	Raja, A. S. et al. [31]
2	2013	-	BPC and RBC Network	96.66% & 98.88%	Nandini, M. et al. [32]
3	2015	Deep ID 3		99.53%	Yi Sun, et al. [19]
4	2015	AFLW		99.00%	Haoxiang Li et al. [29]
5	2015	Multi PIE dataset	CPF	99.50%	JunhoYim et al. [52]
6	2015	AFLW		90.00%	Sachin Sudhakar Farfade [30]

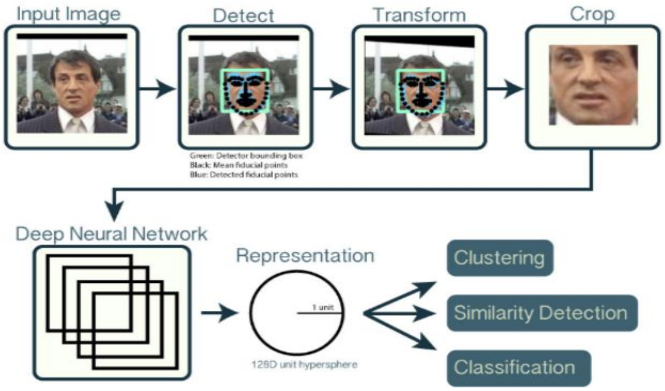


Fig. 2. Face recognition using neural network.

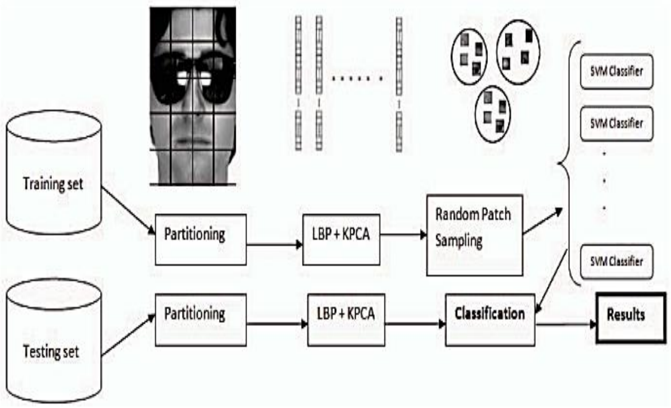


Fig. 3. Face recognition by using SVM [38].

C. Support Vector Machine (SVM)

SVM is the kind of supervised learning algorithm that uses data for classification and regression analysis. SVM provides advantages of being effective in high dimensions. SVM can be implemented to recognize the faces after facial feature extraction [3]. SVM can yield better outcomes when the large quantity of data set is selected directly with training (Fig. 2). However, Least Square Support Vector Machine (LS-SVM) [50], [51] is among the popular one in SVM types that is being successfully utilized for face recognition task.

This provides advantage of fast computation, speed along with high recognition rate [1]. Component-based SVM classifier [16] is another variant of SVM in face recognition. The Support Vector Machine (SVM) classifier is a most widely used technique that is being implemented on a wide range of classification problems. Mostly these problems are in high dimensions and they are not linearly separable. SVM is useful in the advent of dealing with very high dimensional data. Researchers worked on SVM for classification of face recognition and got better results as shown in the below Table III.

TABLE III. COMPARATIVE STUDY OF FACE RECOGNITION METHODS BASED ON SVM

S #	Ye ar	Database	Technique	Accu racy	Reference
1	009	ORL Face Database	Least Square SVM	96%	Xie, Jianhong et al. [25]
2	20 11	ORL Face Database	ICA, SVM	96%	Kong, Rui et al. [26]
3	20 11	FERET Database, AT&T Database	2D-Principal Component Analysis, SVM	95.10 %	Le, Thai Hoang et al. [27]
4	20 16	Yale Faces	SVM	97.78 %	Bhaskar Anand & Prashant K Shah [24]

D. Gabor Wavelet

Dennis Gabor in 1946 introduced a tool for signal processing in noise removal and named as Gabor filter. Gabor wavelets technique is being widely used for face tracking and position estimation in face recognition. While an image representation using the Gabor wavelet transform provides both the spatial relations and spatial frequency structure. As shown in Fig. 4, it has a characteristic that allow it to delineate the properties of spatial localization, spatial frequency selectivity, and orientation [17], [18]. Gabor Wavelets works well over extraction of edge and shape information and it represents the faces in a compact way which is more similar to the feature based methods [19].

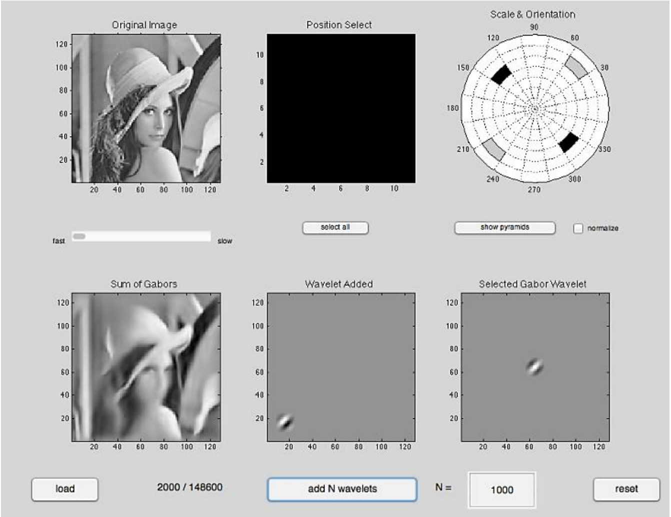


Fig. 4. Gabor wavelet process of recognition.

The main advantages of Gabor Wavelets Transform are face feature reduction and its global feature representation in face recognition [21], [29], [30]. Table IV shows comparative analysis of Gabor filters by different researchers.

TABLE IV. COMPARATIVE STUDY OF FACE RECOGNITOIN METHODS BASED ON GABOR WAVELETS

S#	Year	Database	Technique	Accuracy	Reference
1	2012	FRGC & CASIA	3D GPSR	95.80%	Ming et al. [44]
2	2014	Yale Face DB	PCA, LGBPHS & DPL	98.30%	Hyunjong Cho et al. [43]
				97.30%	
				99.20%	
				99.70%	
3	2014	RBM	PCA	99.5%	Dong Yi et al. [42]
4	2013	ORL Database, FRGCv2	Magnitude Phase of Gabor, PCA, SVM	99.90%	Bellakhdhar et al. [45]
5	2015	T1-w dataset	SVM RBF	93.8% +0.3+-	Nooshin et al. [38]
6	2015	ORL	MAHCOS Distance	97.50%	Vinay.A et al. [39]
7	2015	IIT-Dehli Database	PCA	99.20%	Shervin Minaee et al. [40]
8	2015	FERET & CMU-PIE		92.80%	Leonardo A. Camenta [41]

E. Hidden Markov Models

Hiddel Markov Model (HMM) is another statistical modelling technique in which the system undergoes in Markov process with hidden states (Fig. 5). This model was proposed in 1960 and provided a significant contribution towards speech recognition. HMM is well established method in reinforcement learning, temporal pattern recognition, and bioinformatics applications. Currently it is being implemented to recognize face expressions. Also it can be applied to the video sequences for face recognition. It needs a sequence of 1D and 2D images for experimental purposes [21]; but firstly these images should be converted into a chronological sequence of 1D or spatial. However, model consists of two processes, in which first Markov Chain process having a finite number of states is not viewed explicitly. While in other process each state constitutes a set of probability density function connected with it [1]. Although for research, generally 5-state HMM is designed for face recognition system. 5-state HMM is grouped into five facial features such as eyes, nose, mouth, chin and forehead for frontal view face images as shown in Fig. 6 [47]. But the number of states can be added or removed which depends upon system's requirement.

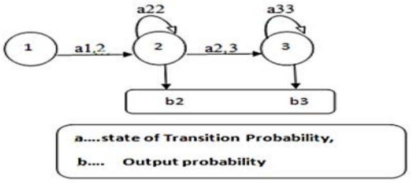


Fig. 5. Three states of transition from left to right for HMM [25].

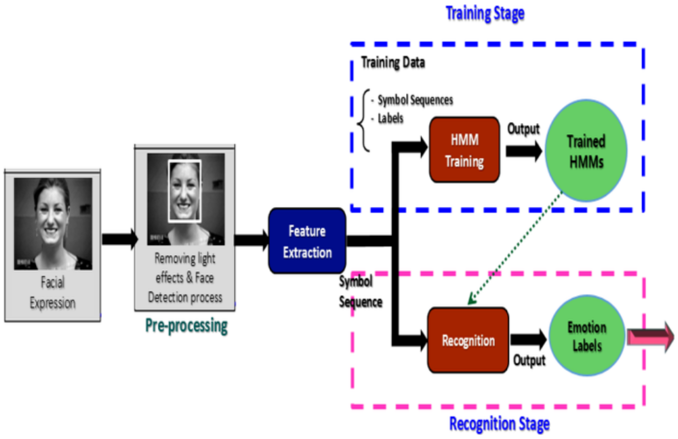


Fig. 6. Hidden Markov Model process of recognition.

In other case, 7-State HMM [48] provide more details which boosts the performance of the face recognition system. Various researchers have worked to get satisfactory outcomes by applying different algorithms in this model. However, a new enhanced model The Adaptive Hidden Markov Model (AHMM) [49] is proposed by authors to find out the issues of identifying the faces via a video sequence. The comparative result in Table V depicts the accuracy ratio of HMM.

TABLE V. COMPARATIVE STUDY OF FACE RECOGNITION METHODS BASED ON HIDDEN MARKOV MODELS (HMM)

S#	Year	Database	Technique	Accuracy	Reference
1	2013	ORL FaceDB, Yale Face DB	Sub-Holistic HMM	95.25% & 94.45%	Sharif, Muhammad et al. [28]
2	2015	CK+ UNBC-McMaster	MIL - HMM	85.23%	Chongliang Wu et al. [22]
3	2015	UMIST	MS-HMM	93.66%	Samiksha Agarwal et al. [23]
4	2013	MSR-Action3D	DMM-HMM	90.50%	Chen Chen et al. [24]
5	2015	SCOP	HHblits	93.80%	James Lyons et al. [25]

V. APPLICATIONS OF FACE RECOGNITION

There are many applications where face recognition techniques are successfully used to perform a specific task. Few of them are described as under:

A. *Access Control*

Access control allows the authorized group of users to access the personal account by logon through their email account using computer accessing bank account through ATM machine. But using face recognition system face pictures are taken under natural conditions such as frontal face images. Such kind of systems yields optimal accuracy without any intervention from the user. These automatic face recognition systems are also used to view and control a user activity on PC or ATM machine; for example, when users leave the PC without proper closing their files and folders for a predetermined time. Then system halts until user again logon and is recognized. In this case, only legitimate persons are allowed to access account.

B. *Security*

Security is a most important precursor at all places. Computer security is being carried out with use of face recognition application. In this regard, image database is being used for investigation purposes [1]; for instance, searching image for authentication of licensed drivers to search missing peoples, immigrants in law enforcement agencies, General identity verification [1], Electoral registration, banking, electronic commerce, identifying newborns, national IDs, passports, employee IDs.

C. *Surveillance*

The word surveillance has been derived from a French phrase which means "watching over". Here (*sur* means "from above” and *veiller* means "to watch"). Surveillance is used to monitor the individual’s behavior, activities, or beside other related information for ensuring the people safety. This can be achieved by means of electronic equipment i.e. closed-circuit television (CCTV) cameras) or interception of electronically transmitted information. Surveillance system offer no. of benefits to different organizations. For example, it is being used by governments for intelligence gathering, control the crime, monitoring the process, person, crowd or object, or the inquiry of crime. However, on the other side, surveillance is often considered as a violation of privacy, and in such cases it is often criticized by civil societies, groups and activists. Liberal democracies have laws which bounds local governments and law agencies to use surveillance, usually restricting them in those circumstances where public safety is compromised. Legitimate organizations have often been imposed such domestic restrictions. However, international surveillance is similar among all types of countries. Moreover researchers are trying to achieve more improved and contented results through the use of latest algorithms in face recognition.

D. *Time & Attendance*

Biometric Time Attendance technologies have been used for Access Control solutions and these are among the latest solutions over traditional systems [23]. In this technology, users are required to expose their face into the machine’s camera by making a certain distance and remove any physical contact with the device. This eliminates any possibility of being tempering or machinery alteration through its non-contact method procedure. Face Recognition system captures specific features from a human face and record it in the form of mathematical template as depicted in Fig. 7. In order to recognize the face, facial image is normalized as to line-up eyes and mouth. Then it performs matching with mathematical vectors from database. Finally face recognition system verifies face and allows for marking attendance or access transaction. These machines could also be implemented for other solutions, where biometric identification/verification is required; such as canteen management, salary distribution, and social services.

Fig. 7. Facial recognition.

E. Pervasive Computing

The aim of pervasive computing is to create a sensor based network as to make smart devices. Hence, sensor network is used to collect, process and send data, and eventually, it can understand its surroundings and improves the human capability and quality of life. However, pervasive computing uses wireless communication and networking technologies, mobile devices, wearable computers, embedded systems, Radio Frequency Identification Devices (RFID) tags, middleware and software agents. Pervasive computing is being widely used in number of applications, for instance in energy, consumer, healthcare, production, military, safety, and logistics.

One of the examples of pervasive computing is a smart Watch developed by Apple Watch. It informs a user for incoming phone call and allows him to complete the call using watch [1].

VI. Conclusion and Future Directions

Study of face recognition has remained a striving area for researchers for many years. In this paper, a comprehensive study was performed over different face recognition methods. After detailed analysis it revealed that PCA is best suited technique when dimension of features is higher for original face images, whereas eigen faces image features method work well for frontal face recognition. Among face recognition methods, the most popular are Neural Networks, Support Vector Machine, Sparse Representation based Classification (SRC), Linear Regression Classification (LRC), Regularized Robust Coding (RRC) and Nearest Feature Line ((NFL). These methods provide better results when the image dimension is under 150 or more. Furthermore, it is suggested that PCA, SVM, NN and Eigen methods still need to be researched so that more satisfactory results could be achieved for face recognition. Moreover, in this paper we also mentioned state of the art face recognition image database and face technology benefits in various applications. However, main findings of this research are highlighted as under:

- The development trends and achievements in the realm of face recognition shows that a lot of researchers have been carried out in last four decades.
- Currently, face recognition system has been implemented for many real-time applications, but still it suffers from several challenges that need to be addressed in order to design a well-established face recognition system.
- Developed face recognition techniques could be analyzed over varying facial expression i.e. under varying lighting conditions and pose. And evaluation could be performed using benchmark and latest face databases.
- Similarly to the face image recognition, the video image recognition is more complicated that needs to be researched.

Further, it is suggested that for recognition of video images, YouTube Faces could be analyzed for evaluation. Furthermore, recognition of emotional human behavior has emerged recently as a promising [52] research area for scholars that should be exploited in future. Finally, it is concluded that still there remains a gap in terms of study in face recognition system that requires to be filled in order to improve its accuracy and efficiency.

References

[1] Muhammad Sharif et al.: “Face Recognition: A Survey”, Journal of Engineering Science and Technology Review 10 (2) (2017) 166- 177

[2] Lacey et al.: “A Longitudinal Study of AutomaticFace Recognition”, IEEE ICB, 2015 pp: 1- 8

[3] Ms. Snehal Houshiram Gorde1, et al.” A Review on Face Recognition Algorithms” Volume III, Issue I Issn No.:2350-1146, I.F-2.71

[4] Azeem, Aisha, et al. "A survey: face recognitiontechniques under partial occlusion." Int. Arab J. Inf. Technol. 11.1 (2014): 1-10.

[5] Tarrés, Francesc, Antonio Rama, and L. Torres. "Anovel method for face recognition under partial occlusion or facial expression variations." Proc. 47th Int’l Symp. ELMAR. 2005.

[6] Sharif M., Mohsin S., Hanan R., Javed M. and Raza M., "3D Face Recognition using Horizontal and verticalMarked Strips", Sindh University Research Journal (SURJ), 43(01-A), (2011)

[7] Jia, Hongjun, and Aleix M. Martinez. "Face recognitionwith occlusions in the training and testing sets." AutomaticFace & Gesture Recognition, 2008. FG'08. 8th IEEEInternational Conference on. IEEE, 2008.

[8] Zhou, Zihan, et al. "Face recognition with contiguousocclusion using markov random fields." Computer Vision, 2009 IEEE 12th International Conference on. IEEE, 2009.

[9] Huang, Fu Jie, et al. "Pose invariant face recognition." Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on. IEEE, 2000.

[10] Chai, Xiujuan, Shiguang Shan, and Wen Gao. "Pose normalization for robust face recognition based on statistical affine transformation." Information,Communications and Signal Processing, 2003 and FourthPacific Rim Conferenceon Multimedia. Proceedings of the2003 Joint Conference of the Fourth InternationalConference on. Vol. 3. IEEE, 2003.

[11] Wright, John, and Gang Hua. "Implicit elastic matching with random projections for pose-variant facerecognition." Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. IEEE, 2009.

[12] Zhang, Wuming, et al. "3D aided face ecognitionacrosspose variations."Biometric Recognition. Springer Berlin Heidelberg, 2012. 58-66.

[13] Shah J. H., Sharif M., Raza M. and Azeem A.," Facerecognition across pose variation and 3S problem", In TÜBİTAK Academic Journals (2012)

[14] Zhang, Xiaozheng, and YongshengGao. "Facerecognition across pose: A review." Pattern Recognition42.11 (2009): 2876-2896.

[15] Wang, Jizeng, and Hongmei Yang. "Face detectionbased on template matching and 2DPCA algorithm." Imageand Signal Processing, 2008. CISP'08. Congress on. Vol. 4. IEEE, 2008.

[16] Huang, Jennifer, Volker Blanz, and Bernd Heisele."Face recognition using component-based SVMclassification and morphable models." Pattern Recognitionwith Support Vector Machines. Springer Berlin Heidelberg,2002.334-34

[17] Jin, Yi, and Qiu Qi Ruan. "Face recognition using gabor-based improved supervised locality preservingprojections." Computing and Informatics 28.1 (2012): 81-95.

[18] Bellakhdhar, Faten, Kais Loukil, and Mohamed Abid."Face recognition approach using Gabor Wavelets, PCA and SVM." IJCSI International Journal of Computer Science Issues 10.2 (2013): 201-206.

[19] Kar, Arindam, et al. "Classification of high-energizedgabor responses using bayesian PCA for human face recognition." 2.2(09)

[20] Dong Yi, Zhen Lei, Shengcai Liao and Stan Z. Li "SharedRepresentation Learning for Heterogeneous FaceRecognition" 2015 pp: 1-15

[21] Chongliang Wu1, Shangfei Wang _1 and Qiang Ji"Multi-Instance Hidden Markov Model For Facial Expression Recognition 2015 IEEE

[22] Samiksha Agrawal, "Facial Expression Detection Techniques: Based on Viola and Jones algorithm and Principal Component Analysis 2015 IEEE

[23] Chen Chen et al., "Real-time human action recognitionbased on depth motion maps" 2013 Springer

[24] Bhaskar Anand & Prashant Face Recognition using SURF Features and SVM Classifier" ISSN 0975- 6450 Volume 8, Number 1 (2016) pp. 1-8

[25] Xie, Jianhong. "Face recognition based on Curvelettransform & LS-SVM." Proceedings of the 2009 International Symposium on Information Processing (ISIP'09), Huangshan, PR China. 2009.

[26] Kong, Rui, and Bing Zhang. "A New Face Recognition Method Based on Fast Least Squares Support Vector Machine." Physics Procedia 22 (2011): 616-621.

[27] Le, Thai Hoang, and Len Bui. "Face recognition based on SVM and 2DPCA."arXiv preprint arXiv:1110.5404 (2011).

[28] Sharif M., Ayub K., Sattar D. and RAZA M., "Real TimeFace Detection", Sindh Univ. Res. Jour. (Sci. Ser.) Vol. 44(4), 597- 600, (2012)

[29] KalavdekarPrakash, N. "Face Detection using NeuralNetwork." International Journal of Computer Applications(0975–8887) 1.14 (2010).

[30] Li, Yongmin, et al. "Multi-view face detection usingsupport vector machines and eigenspacemodelling." Knowledge-Based Intelligent Engineering Systems and Allied Technologies, 2000. Proceedings. Fourth International

[31] Raja, A. S., and V. JosephRaj. "Neural network basedsupervised self-organizing maps for face recognition." International Journal on Soft Computing 3.3 (2012).

[32] Nandini, M., P. Bhargavi, and G. Raja Sekhar. "FaceRecognition Using Neural Networks." International JournalOf Scientific and Research Publications 3.3 (2013): 1. Conference on. Vol. 1. IEEE, 2000.

[33] Abdullah, Manal, MajdaWazzan, and Sahar Bo-Saeed."Optimizing Face Recognition Using PCA." arXiv preprintarXiv:1206.1515 (2012).

[34] Shah J. H., Sharif M., Raza M. and Azeem A.," Facerecognition across pose variation and 3S problem", In TÜBİTAK Academic Journals (2012)

[35] Zhang, Xiaozheng, and YongshengGao. "Face recognition across pose: A review." Pattern Recognition 42.11 (2009): 2876-2896.

[36] Azeem, Aisha, et al. "A survey: face recognition techniques under partial occlusion." Int. Arab J. Inf. Technol. 11.1 (2014): 1-10.

[37] Chai, Xiujuan, Shiguang Shan, and Wen Gao. "Posenormalization for robust face recognition based on statistical affine transformation." Information, Communications and Signal Processing, 2003 and Fourth Pacific Rim Conference on Multimedia. Proceedings of the 2003 Joint Conference of the Fourth International Conference on. Vol. 3. IEEE, 2003.

[38] Wei, Xingjie, Chang-Tsun Li, and Yongjian Hu. "Robustface recognition under varying illumination and occlusion considering structured sparsity." Digital Image Computing Techniques and Applications (DICTA), 2012 International Conference on. IEEE, 2012.

[39] Zhao, Wenyi, and Rama Chellappa. "Illumination-sensitive face recognition using symmetric shape-from-shading."Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on. Vol. 1. IEEE, 2000.

[40] Sharif M., Mohsin S., Jamal M. J. and Raza M.,"Illumination Normalization Preprocessing for face recognition", IEEE International Conference on Environmental Science and Information ApplicationTechnology (ESIAT), , 44-47 (2010)

[41] Sharif M., Raza M. and Mohsin S., "Face RecognitionUsing Edge Information and DCT", Sindh Univ. Res. Jour. (Sci. Ser.) Vol.43 (2) 209-214, (2011)

[42] Sharif M. and Saad A., "Enhanced SVD Based FaceRecognition", Journal of Applied Computer Science & Mathematics, Suceava, Vol. 12(6) ,49-53,(2012)

[43] Cho, Hyunjong, Rodney Roberts, Bowon Jung, Okkyung Choi, and Seungbin Moon. "An efficient hybrid face recognition algorithm using pca & gabor wavelets" 2014

[44] Ming, Yue, QiuqiRuan, and Xueqiao Wang. "Efficient 3dface recognition with Gabor patched spectral regression." Computing and Informatics 31.4 (2012): 779-803.

[45] Bellakhdhar, et al. "Face recognition approach using Gabor Wavelets, PCA and SVM."IJCSI International Journal of Computer Science Issues 10.2 (2013): 201-206.

[46] Riddhi A. Vyas1, Dr.S.M. Shah "Comparision of PCA and LDA Techniques for Face Recognition Feature BasedExtraction with Accuracy Enhancement" 2017, IRJET pages:3332-3336

[47] Nefian, Ara V., and Monson H. Hayes III. "Hidden markov models for face recognition." choice 1 (1998): 6.

[48] Miar-Naimi, H., and P. Davari. "A new fast and efficient HMM-based face recognition system using a 7-state HMM

[49] Liu, Xiaoming, and Tsuhan Chen. "Video-based face recognition using adaptive hidden markov models." Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on. Vol. 1. IEEE along with SVD coefficients." (2008).

[50] Xie, Jianhong. "Face recognition based on Curvelet transform and LS-SVM." Proceedings of the 2009 (ISIP'09)Huangshan,

[51] Zhang, Xinming, and JianZou. "Face recognition based on sub-image feature extraction and LS-SVM." Computer Science and Software Engineering, 2008 International Conference on. Vol. 1. IEEE, 2008.

[52] Maha Jazouli, Aicha Majda, Arsalane Zarghili. "A $PRecognizer for Automatic Facial Emotion Recognition using Kinect Sensor." 978-1-5090-4062-9/17/$31.00 ©2017 IEEE.

MapReduce Programs Simplification using a Query Criteria API

Boulchahoub Hassan, Khalil Namir, Amina Rachiq, Labriji Elhoussin, Benabbou Fouzia
Department of Mathematics and Computer Science
Faculty of Sciences Ben M'SIK
Casablanca, Morocco

Abstract—**A Hadoop HDFS is an organized and distributed collection of files. It is created to store a huge part of data and then retrieve it and analyze it efficiently in a less amount of time. To retrieve and analyze data from the Hadoop HDFS, MapReduce Jobs must be created directly using some programming languages like Java or indirectly using some high level languages like HiveQL and PigLatin. Everyone knows that creating MapReduce programs using programming languages is a difficult task that requires a remarkable effort for their creation and also for their maintenance. Writing MapReduce code by hand needs a lot of time, introduce bugs, harm readability, and impede optimizations. Profiles working in the field of big data always try to avoid hard and long programs in their work. They are always looking for much simpler alternatives like graphical interfaces or reduced scripts like PIG Latin or even SQL queries. This article proposes to use a MapReduce Query API inspired from Hibernate Criteria to simplify the code of MapReduce programs. This API proposes a set of predefined methods for making restrictions, projections, logical conditions and so on. An implementation of the Word Count example using the Query Criteria API is illustrated in this paper.**

Keywords—Hadoop; HDFS; MapReduce

I. Introduction

Big data analysis has become a priority for all companies and organizations that want to maintain a high level of competition. To accomplish this task, companies use several frameworks like Hadoop ecosystem which ensures both storage and processing despite the huge volume of data. Hadoop [1], [2] contains mainly a distributed File System HDFS [3] and a distributed computation framework MapReduce [4].

To analyse data stored in HDFS and according to the user's profile and competence, several programming languages are used such as java to create MapReduce programs directly [5]. Some high level languages are also used like PIG Latin scripts or HiveQL queries [6]. In this domain, several research projects have attempted to simplify the code of MapReduce programs to make them readable and easily maintainable.

This article suggests using an API called MapReduce Criteria inspired from the Hibernate Criteria API to hide the code of the restrictions and projections made on the data stored in the HDFS. Thus, the number of lines in MapReduce programs will be reduced to facilitate readability and maintenance.

This paper is organized as follows. Section 2 describes Hadoop ecosystem. Section 3 presents the motivation of using MapReduce Criteria. As for Section 4, it talks about Pig, Hive and Sqoop as related work. Section 5 briefly describes the Query Criteria API. The last section contains final conclusions and points to further work.

II. Hadoop Ecosystem

Hadoop ecosystem is a set of popular frameworks [7] that provides distributed processing over a huge amount of data. Hadoop is designed to solve data storage problems caused by the large amount of data generated each second. The data managed by this framework is processed in a parallel way by exploiting thousands of machines. Fig. 1 shows a simple architecture of Hadoop ecosystem.

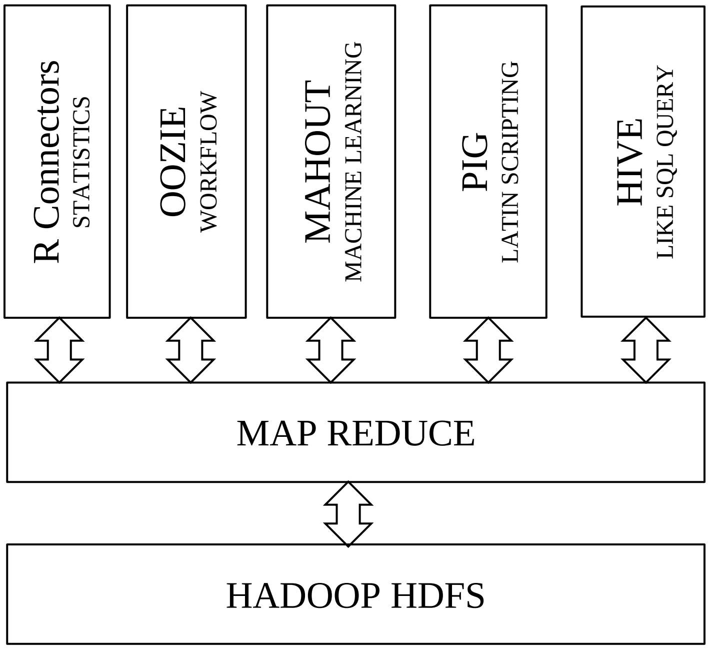


Fig. 1. Simple architecture of Hadoop ecosystem.

Along with the market trends and the diversity of profiles that intervene on the data, several additional technical components have been emerged; components for people who are used to work in declarative SQL language and other components that are based on procedural languages.

With the multitude of solutions that currently exist in the market, each profile must carefully choose the Big Data solution that aligns with its skills. The analysts will likely find that they can ramp up on Hadoop faster by using Hadoop data warehouses such as Hive [8], Impala [9] and HAWQ now frequently deployed at customer sites. Developers who want better control of the data flow process and those who come from a procedural language context will choose to work in PIG Latin. Despite the diversity of existing solutions, they all use the same HDFS for cluster storage and the same MapReduce model for distributed processing (Fig. 2).

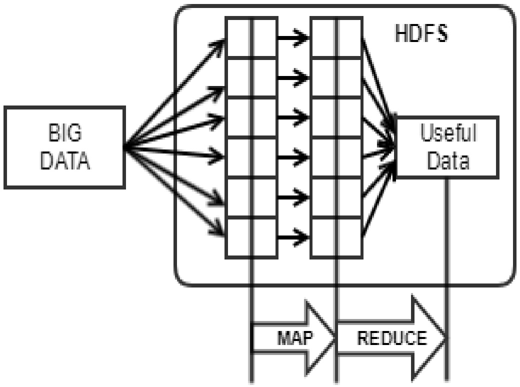


Fig. 2. Conceptual overview of HDFS and MapReduce.

A. Hadoop Distributed File System (HDFS)

To meet the ever-changing volumes of data processed every day, The Hadoop Distributed File System (HDFS) is designed to be highly fault-tolerant and to be deployed on low-cost hardware. HDFS is based on a master / slave architecture. It offers a master server (NameNode) and slaves (DataNodes) per node of the cluster [3]. The NameNode manages the namespace of the file system and also orchestrates access to the files by the clients. The DataNodes manages the storage associated with the nodes on which they run. A simple HDFS architecture is given at Fig. 3.

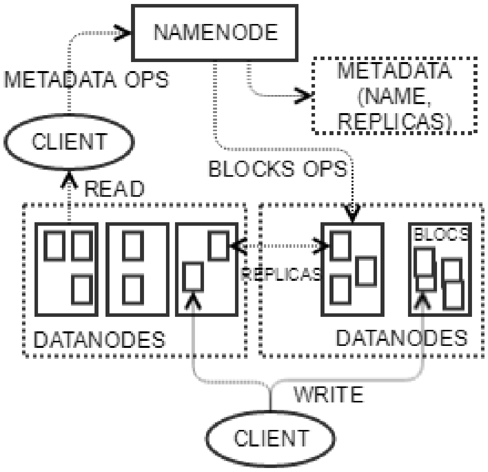


Fig. 3. Architecture of the HDFS.

B. MapReduce

The basis of the MapReduce framework was defined by Dean and Ghemawat at their paper in 2004 [5]. MapReduce orchestrates the processing of a large data sets using parallel computing on a cluster. It manages all issues related to partitioning the input data, scheduling the program's execution and data transfers. Several research papers are focused on the MapReduce model to apply it to some business domains [10], [11] to resolve some algorithms issues [12], [13] or to search for some optimization leads [14], [15]. The user of the MapReduce library expresses the computation as two functions: Map and Reduce.

Map function takes an input pair and produces a set of intermediating key/value pairs. It gathers together all intermediate values associated with the same intermediate key and passes them to the Reduce function. Reduce function written by the user accepts an intermediate key and a set of values for that key. It merges these values to form a possibly smaller set of values (Fig. 4).

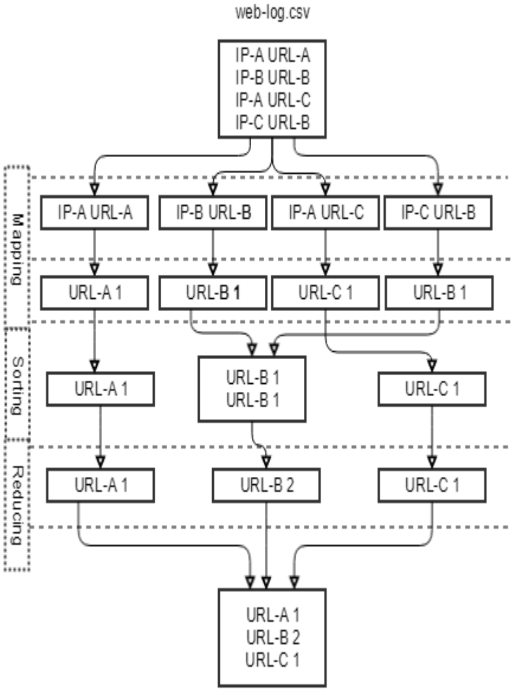


Fig. 4. Map function showing values to form a possibly smaller set of values.

III. Motivation

MapReduce programs are positioned in the core of all BigData systems. Unfortunately MapReduce programs have been criticized for several disadvantages including the large number of instructions, the lack of readability and also the difficulty of maintenance.

In order to simplify the number of instructions, the readability and the maintenance of the MapReduce programs, we propose to use the MapReduce Query API which will hide all the instructions related to:

- **Restrictions** like equal, not equal, less than, more than, etc.

- **Logical expressions like AND, OR, XOR etc.**
- **Relations between data** like “inner_join”, “left_join”
- **Projections** like group, maximum, minimum, average, etc.
- **Orders** ascendant and descendant.

We propose to apply all methods defined in Hibernate Criteria [16] to MapReduce programs. Among the major concerns of big data solutions today, we can mention the optimization of execution times and also the simplification of creating MapReduce programs. Solutions mentioned at the next section try to hide the complexity of MapReduce programs by generating MapReduce plans automatically.

IV. Related Work

Different tools and sub-projects have been created to simplify the task for users who are not so good at programming languages. Many frameworks have been implemented to help users who are struggling with Hadoop, especially while performing any MapReduce tasks. Among these solutions, we find Pig, Hive and Sqoop described briefly in the following paragraphs.

A. Pig

Pig is a procedural language platform used to develop a script Pig Latin [17]: a sequence of steps, much like in a programming language, each of which carries out a single high-level data transformation e.g., filtering, grouping, or aggregation.

B. Hive

A data warehouse solution that allows users to write SQL like Query (HiveQL) and translate them into physical plans of MapReduce jobs using the Thrift Server. Hive proposes many external interfaces (Command Line, Web UI, JDBC...) to challenge with its database [18]-[20]. The latest version of Hive (since version 2.0) allows also procedural SQL on Hadoop [21].

C. Sqoop

This solution is also adopted by the Apache Foundation in order to achieve bulk data transfers between Hadoop and structured databases such as relational databases. Sqoop hides and simplifies the complexity of MapReduce programs to users [22], [23].

V. Proposed work

The MapReduce Query API inspired from Hibernate Criteria API will represent a query against a particular file stored at the HDFS. The interface will provide the same powerful mechanism of hibernate criteria API and will allow a programmatic creation of queries against the HDFS (Fig. 5). It’s an alternate way to manipulate objects generated from data stored at the HDFS. Specifying the structure of the data to be loaded from the HDFS is required for using MapReduce Query API. It is the equivalent of Relational Object Mapping in the Hibernate Framework. Any program based on MapReduce Query API will be automatically translated to MapReduce programs according to a previously defined plan.

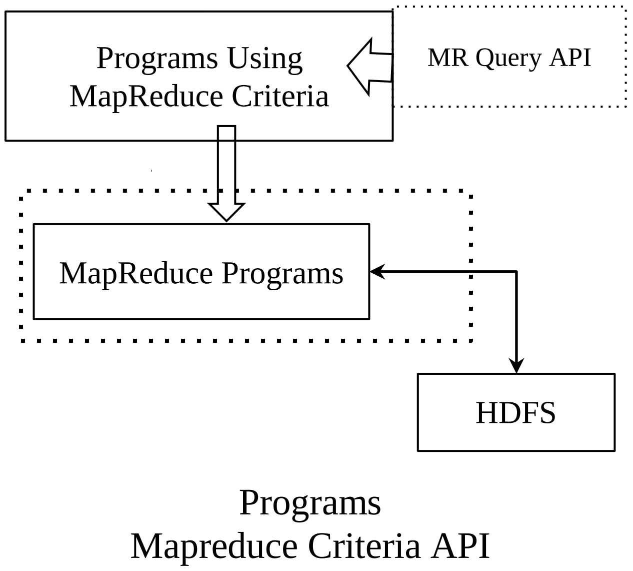


Fig. 5. MapReduce processing.

VI. Implementation

WordCount is a famous application that counts the number of occurrences for each word in a given set of files. The input for this implementation is a file of comments as detailed below:

A. WordCount Example without MapReduce Criteria

To develop a simple MapReduce example in the current model, it is necessary to create at least three classes: A class "Mapper" as shown in Table I, a "Reduce" class as shown in Table II and a "Main" class as shown in Table III.

TABLE I. WordCount Mapper Class

```
package com.hadoop.mapreduce.example;
import java.io.IOException;
import java.util.StringTokenizer;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.LongWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapred.MapReduceBase;
import org.apache.hadoop.mapred.Mapper;
import org.apache.hadoop.mapred.OutputCollector;
import org.apache.hadoop.mapred.Reporter;
public class WordCountMapper extends MapReduceBase
implements Mapper<LongWritable, Text, Text,
IntWritable> {
        private final static IntWritable one = new
IntWritable(1);
private Text word = new Text();
public void map(LongWritable key, Text value,
OutputCollector<Text, IntWritable> collector, Reporter
reporter) throws IOException {
String line = value.toString();
StringTokenizer st = new StringTokenizer(line, " ");
while (st.hasMoreTokens()) {
```

```
            word.set(st.nextToken().trim());
if (!"Apache".equals(word))
{
   collector.collect(word, one);
                          }
            }
}
}
```

TABLE II. WORDCOUNT REDUCER CLASS

```
package com.hadoop.mapreduce.example;
import java.io.IOException;
import java.util.Iterator;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapred.MapReduceBase;
import org.apache.hadoop.mapred.OutputCollector;
import org.apache.hadoop.mapred.Reducer;
import org.apache.hadoop.mapred.Reporter;
public class WordCountReducer extends MapReduceBase
implements Reducer<Text, IntWritable, Text, IntWritable>
{
public void reduce(Text key, Iterator<IntWritable> values,
OutputCollector<Text, IntWritable> outputCollector,
Reporter reporter) throws IOException {
         int sum = 0;
          while (values.hasNext()) {
            sum = sum + values.next().get();
          }
outputCollector.collect(key, new IntWritable(sum));
}
}
```

TABLE III. WORDCOUNT MAIN CLASS

```
package com.hadoop.mapreduce.example;

import java.net.URI;
import org.apache.hadoop.conf.Configuration;
import org.apache.hadoop.fs.FileSystem;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapred.FileInputFormat;
import org.apache.hadoop.mapred.FileOutputFormat;
import org.apache.hadoop.mapred.JobClient;
import org.apache.hadoop.mapred.JobConf;
import org.apache.hadoop.mapred.RunningJob;
import org.apache.hadoop.mapred.TextOutputFormat;
public class WordCount {
public static void main(String[] args) throws Exception {
Configuration conf = new Configuration();
Path inputPath = new Path(
            "hdfs://127.0.0.1:9000/input/comments.txt");
Path outputPath = new
Path("hdfs://127.0.0.1:9000/output/");
JobConf job = new JobConf(conf, WordCount.class);
job.setJarByClass(WordCount.class);
job.setJobName("WordCounterJob");
FileInputFormat.setInputPaths(job, inputPath);
FileOutputFormat.setOutputPath(job, outputPath);
job.setOutputKeyClass(Text.class);
job.setOutputValueClass(IntWritable.class);
job.setOutputFormat(TextOutputFormat.class);
job.setMapperClass(WordCountMapper.class);
job.setReducerClass(WordCountReducer.class);
FileSystem hdfs =
FileSystem.get(URI.create("hdfs://127.0.0.1:9000"),
                                        conf);
if (hdfs.exists(outputPath))
hdfs.delete(outputPath, true);
RunningJob runningJob = JobClient.runJob(job);
System.out.println("job.isSuccessfull: " +
runningJob.isComplete());
}
}
```

B. WordCount Example using MapReduce Criteria

The objective of MapReduce Criteria is to reduce the number of rows and classes for developers using MapReduce; it will facilitate the creation and also the maintenance of their programs. Each program based on MapReduce Criteria will be automatically translated into Mappers and Reducers classes. The example "WordCount" in MapReduce Criteria is given in Table IV.

TABLE IV. WORDCOUNT EXAMPLE USING MAPREDUCE CRITERIA

```
package com.hadoop.mapreduce.example;
public class WordCountMrCriteria extends MrQueryScript
{
public static void main(String[] args) {
String hdfs = "hdfs://127.0.0.1:9000";
DataList dlist1 = load(hdfs + "/input/comments.txt",
new String[] {"line:String"});
DataList dlist2 = null;
for (DataObject dobj : dlist1.getDataObjectList()) {
dlist2 = tokenize(dobj,"line", " ", new String[]
{"word:String"});
}
MapReduceCriteria c1 = dlist2.getMrCriteria()
.add(Restrictions.ne("word", "Apache"))
.add(Projections.groupBy("word"))
.add(Projections.rowCount());

DataList dlist3 = c1.dataList();
dlist3.store(hdfs + "/output/countword.txt");
}
}
```

VII. Conclusion

MapReduce is a programming model created to perform distributed processing of a large datasets stored in a distributed file system HDFS. It is well-known that MapReduce programs are difficult to create, to read and to maintain. Therefore, it is necessary to simplify them using some frameworks or APIs.

This paper has suggested using an API called MapReduce Criteria in order to reduce the number of MapReduce instructions and also to hide Mappers and Reducers classes for developers. In our future work we will compare MapReduce programs that use the Query Criteria API with existing languages that also simplify the use of MapReduce as Pig Latin and Hive.

Acknowledgment

We would like to thank Mr Labriji and Mr Rachiq for their useful comments and their dedicated work.

References

[1] D. Keulen, Hadoop: The Definitive Guide. CreateSpace, 2014.

[2] “Welcome to ApacheTM Hadoop®!” [Online]. Available: http://hadoop.apache.org/.

[3] “HDFS Architecture Guide.” [Online]. Available: https://hadoop.apache.org/docs/r1.2.1/hdfs_design.html.

[4] “Apache Hadoop 2.9.0 – MapReduce Tutorial.” [Online]. Available: https://hadoop.apache.org/docs/stable/hadoop-mapreduce-client/hadoop-mapreduce-client-core/MapReduceTutorial.html

[5] J. Dean and S. Ghemawat, “MapReduce,” Commun. ACM, vol. 51, no. 1, 2008, p. 107.

[6] S. Stewart, The Hive. University of Georgia Press, 2008.

[7] “Hadoop Ecosystem: An Introduction,” International Journal of Science and Research (IJSR), vol. 5, no. 6, 2016, pp. 557–562.

[8] E. Capriolo, D. Wampler, and J. Rutherglen, Programming Hive. “O’Reilly Media, Inc.,” 2012.

[9] J. Russell, Getting Started with Impala: Interactive SQL for Apache Hadoop. “O’Reilly Media, Inc.,” 2014.

[10] U. D., D. Umesh, and B. Ramachandra, “Big Data Analytics to Predict Breast Cancer Recurrence on SEER Dataset using MapReduce Approach,” Int. J. Comput. Appl. Technol., vol. 150, no. 7, 2016, pp. 7–11.

[11] Z. Wu, B. Mao, and J. Cao, “MRGIR: Open geographical information retrieval using MapReduce,” in 2011 19th International Conference on Geoinformatics, 2011.

[12] Q. Guo, B. Palanisamy, and H. Karimi, “A Distributed Polygon Retrieval Algorithm using MapReduce,” in Proceedings of the 10th IEEE International Conference on Collaborative Computing: Networking, Applications and Worksharing, 2014.

[13] M. A. Alshammari and E.-S. M. El-Alfy, “MapReduce implementation for minimum reduct using parallel genetic algorithm,” in 2015 6th International Conference on Information and Communication Systems (ICICS), 2015.

[14] Q. Liu, W. Cai, B. Wang, Z. Fu, and N. Linge, “An Optimization Scheme in MapReduce for Reduce Stage,” International Journal of Grid and Distributed Computing, vol. 9, no. 8, 2016, pp. 197–208.

[15] Y. Tao, Q. Zhang, L. Shi, and P. Chen, “Job Scheduling Optimization for Multi-user MapReduce Clusters,” in 2011 Fourth International Symposium on Parallel Architectures, Algorithms and Programming, 2011.

[16] “Chapter 15. Criteria Queries.” [Online]. Available: https://docs.jboss.org/hibernate/orm/3.3/reference/en/html/querycriteria.html.

[17] C. Olston, B. Reed, U. Srivastava, R. Kumar, and A. Tomkins, “Pig latin,” in Proceedings of the 2008 ACM SIGMOD international conference on Management of data - SIGMOD ’08, 2008.

[18] A. Thusoo et al., “Hive,” Proceedings VLDB Endowment, vol. 2, no. 2, 2009, pp. 1626–1629.

[19] E. Capriolo, D. Wampler, and J. Rutherglen, Programming Hive. “O’Reilly Media, Inc.,” 2012.

[20] J. Venner, Pro Hadoop. 2009.

[21] “HPL/SQL Reference - HPL/SQL - Procedural SQL on Hadoop, NoSQL and RDBMS,” 04-Aug-2017. [Online]. Available: http://www.hplsql.org/doc.

[22] “Sqoop User Guide (v1.4.0-incubating).” [Online]. Available: https://sqoop.apache.org/docs/1.4.0-incubating/SqoopUserGuide.html.

[23] K. Ting and J. J. Cecho, *Apache Sqoop Cookbook: Unlocking Hadoop for Your Relational Database*. “O’Reilly Media, Inc.,” 2013.

Modelling of Thermal Storage in Damaged Composite Structures using Time Displaced Gradient Field Technique (TDGF)

Mahmoud Zaki Iskandarani
Department of Electronics and Communications Engineering
Al-Ahliyya Amman University
Amman, Jordan

***Abstract*—This paper presents a new approach to composite surface characterization using Gradient Field time displacement. The new technique employs calculation of thermally charged regions within a composite structure as a result of each area gradient and then correlates the regions (storage areas) using a time displaced (Lag) model. The resulting data show that a rate-dependent model is fit to describe the behavior of damaged areas within a composite structure, which act as energy storage elements. The rate of dissipation of stored energy per region contributes to the shape and area of the resulting correlated Lag curve.**

***Keywords*—Gradient norm; edge detection; gray level mapping; segmentation; rate-dependent; lag; thermal images**

I. Introduction

Image enhancement is critical in image processing. Normally image enhancement is basically achieved using histogram equalization or one of its related techniques with no provision of proper mathematical model for the intensity variation associated with histogram equalization. Also, histogram equalization uses Probability Distribution Function (PDF) and Cumulative Distribution Function (CDF) and both assume uniform distribution, which is not applicable in many cases.

This problem can be resolved using gradient analysis or correlated gradient with histogram analysis. Such a unique concept has been employed for specific applications, such as, deblurring and image restoration [1]-[7].

Edge detection is an extremely important technique in image processing and image analysis. Edge detection process preserve important image structural features. Edge Detection involves taking steps in the process of locating the sharp edges which are discontinuous that result in variation in pixels intensities, thus, defining boundaries within the image [8]-[11].

Composite structures continue to be used widely in aerospace and automotive applications due to its light weight coupled with high strength. There is a growing interest in the application of thermal methods for nondestructive testing (NDT) of composite components. There are many gains using thermal NDT as compared to other methods, as its capable of covering large and complex areas without direct contact, and it produces results in a reasonably short time. Thermography has a good potential for detection of various abnormalities, such as delamination, and disbands.

Non-uniform composite structures and/or damaged composites will most of the time result in the creation of edges within the fiber-matrix system. Consequently, segmentation will occur and subdivides the composite image into regions. The number of regions formed as a result of segmentation depends on the type of damage a structure suffered. The segmented image is a function of both discontinuity and uniformity. Both factors can be used to establish similarity and level of damage as a result of abrupt change or edges [12]-[16].

Detecting damage using thermal images produced as a result of testing of composite structures is a challenging task owing to their variable appearance and the wide range of shapes and orientations that a damage can cause. The first need is a robust feature set that allows for damaged areas to be discriminated efficiently, even in cluttered backgrounds under difficult illumination. The second need is for a good mathematical model that covers effect of thermal energy on image properties due to the existence of non-uniform or damaged composite structure [17]-[21]. In addition, Knowledge of the thermal conductivity of the composite structure is essential in assessing and modelling heat transfer through fiber and matrix.

At present, active thermal imaging is regularly used for inspection during manufacturing and in service to inspect composite parts and units. The technique can be used to specify damaged areas and their boundaries. The technique measures the difference between the temperature of a defect area and a defect-free one. Processing of the results of can be achieved using dynamic thermal tomography technique fast Fourier transform.

In this work, a new approach to using gradient field analysis is applied to PVT images of composite structures, which results in a new approach in detecting and analysing composite structure damage, which is comparable in its approach to the active thermal imaging technique. The presented technique in the work is based on correlative displacement of gradient fields associated with damaged areas of composite structures identified as a result of energy based segmentation due to energy storage within the composite structure that results in thermal edges caused by structural damage.

II. Materials and Methods

The main objective is to use segmentation of thermal charged regions to detect damage. Damaged areas will vary in their energy storage in comparison to undamaged areas, resulting in thermal edges. Such a phenomenon can be realized through gradient application to captured images over periods of time. Pulse Video Thermography (PVT) is employed in the testing. The used equipment comprised a heating source and thermal imaging system. Pulses obtained by discharging energy pulses through flash tubes, which are directed at the tested structure. The tested structure consisted of 5mm thick Reaction Injection Molding (RIM) composites. The obtained thermal images converted to gray levels before applying segmentation algorithm as a function of Gradient Field (GF). Different levels of threshold used to obtain an optimum boundary isolation and region segmentation. Fig. 1 shows the experimental arrangement.

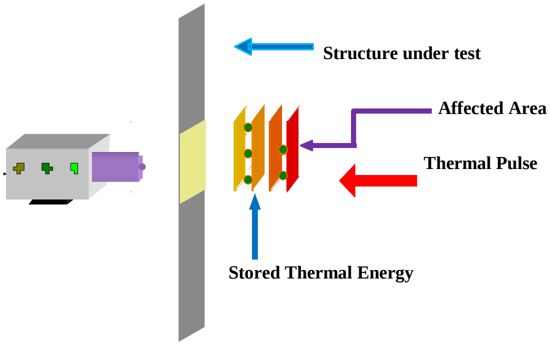


Fig. 1. Experimental setup.

III. Results

Fig. 2 to 4 show the tested component, at sampled intervals {13, 18, 23} minutes.

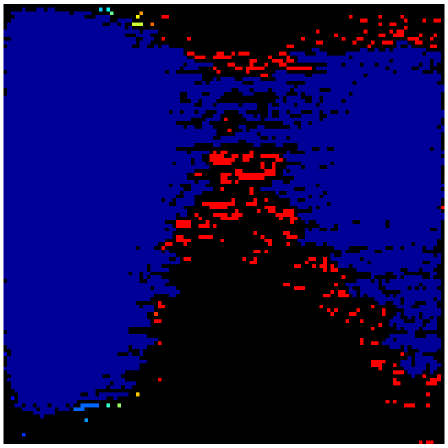


Fig. 2. Jet coloring of thermal image after 13 minutes.

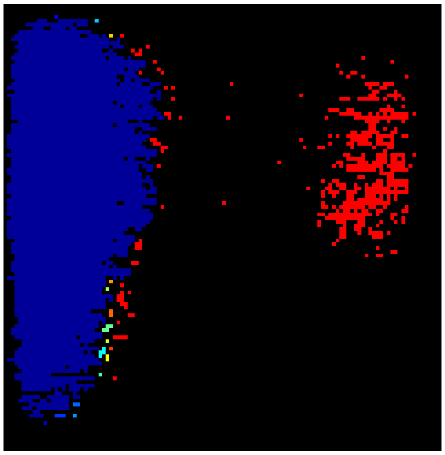


Fig. 3. Jet coloring of thermal image after 18 minutes.

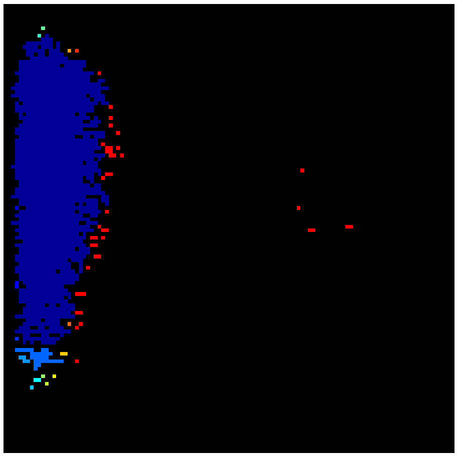


Fig. 4. Jet coloring of thermal image after 23 minutes.

Fig. 5 to 7 show the Gradient Field for each of the time sampled thermal images.

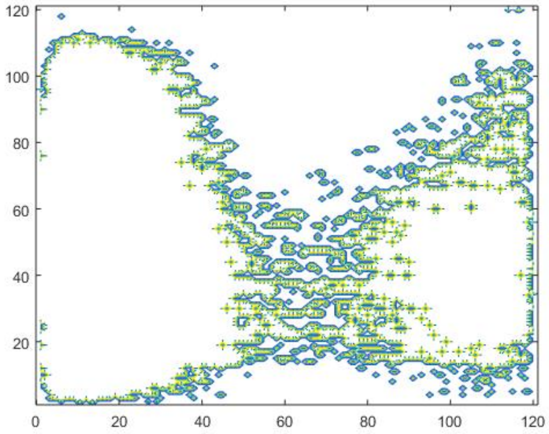


Fig. 5. Gradient field of thermal image after 13 minutes (coordinates represent pixels with image area of 120 by 120 by 3).

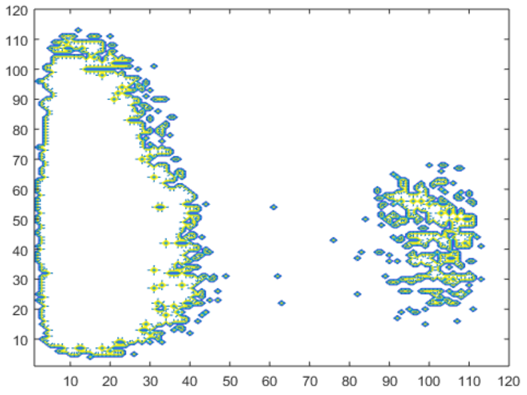


Fig. 6. Gradient field of thermal image after 18 minutes coordinates represent pixels with image area of 120 by 120 by 3).

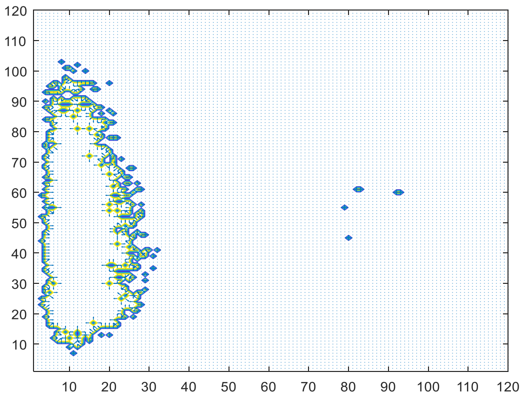


Fig. 7. Gradient field of thermal image after 23 minutes coordinates represent pixels with image area of 120 by 120 by 3).

IV. Analysis and Discussion

The observed Gradient field indicates four stages of thermal interaction between the composite sample under test and the thermally applied pulse:

A. *Initial absorption (t=t+)*: The whole structure is subjected to the applied pulse of energy with approximately equal distribution of intensity levels and pixel population over the structural area.

B. *Second stage (t=13 minutes)*: Two interconnected and thermally charged areas appear, with the rest of the areas discharged. The two touching areas (inter-thermal convergence) exchange thermal energy with each other and with the rest of the composite structure.

C. *Third stage (t=18 minutes)*: Two thermally charged areas, with well-defined boundaries (inter-thermal divergence and intra-thermal convergence) observed as a result of the areas in the second stage relinquished. This is due to thermal discharge of stored energy.

D. *Fourth stage (t = 23 minutes)*: Mainly one area is left, which is associated with the highest level of surface non-uniformity(damage), as other parts have come to a total thermal discharge and heat dissipation.

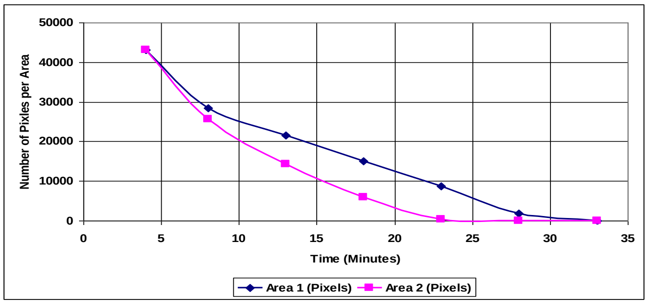


Fig. 8. Relationship between the two thermally charged regions in the tested composite.

Fig. 8 shows the relationship between the time responses of the two thermally charged arras within the tested composite structure. From the plots the relationship between the two charged areas representation can be modelled using a Rate-Dependent model as it is evident that during the thermal discharge process one thermal energy is exchanged between the two mainly charged regions (damaged areas) of tested composite structure, thus introducing a time delay or lag between the discharging activities of the two main regions. This introduce two main cycles considered in the proposed model:

1) Primary thermal charge-discharge cycle as a function of the applied PVT pulse.

2) Secondary thermal charge-discharge cycle as a function of damaged area size.

Both processes follow an outward direction. In general, the model is represented as in (1).

$$A_1(t) = R_{initial} - \int_0^{\infty} A_1(t-\tau)\,d\tau \tag{1}$$

$$A_2(t) = R_{initial} - \int_0^{\infty} A_2(t-\tau)\,d\tau \tag{2}$$

$$Lag(t) = \left|A_1(t) - A_2(t)\right| \tag{3}$$

$$Lag(t) = \left|\int_0^T A_1(t-\tau)\,d\tau - \int_0^T A_2(t-\tau)\,d\tau\right| \tag{4}$$

From the results, it is observed that one area will discharge at a faster rate and hence thermal decay will occur within the integration interval of the second area. Thus, (4) becomes (assume here that A_2 discharges at a faster rate):

$$Lag(t) = \left|\left[\int_0^{t_1} A_1(t-\tau)\,d\tau + \int_{t_1}^{T} A_1(t-\tau)\,d\tau\right] - \int_0^{t_1} A_2(t-\tau)\,d\tau\right| \tag{5}$$

where T is the maximum testing time. Thus:

$$Lag(t) = \left|\left[\int_0^{t_1} A_1(t-\tau)\,d\tau - \int_0^{t_1} A_2(t-\tau)\,d\tau\right] + \int_{t_1}^{T} A_1(t-\tau)\,d\tau\right| \tag{6}$$

which gives:

$$Lag(t) = \left| \left[\left(\int_0^{t_1} A_1(t-\tau) - A_2(t-\tau) \right) d\tau \right] + \int_{t_1}^{T} A_1(t-\tau) d\tau \right| \tag{7}$$

From (7), both A_2 and A_1 can be obtained as a function of each other:

$$A_1(t) = lag(t) - A_2(t) \tag{8}$$

$$A_2(t) = lag(t) - A_1(t) \tag{9}$$

Applying observations to (7) and noting that one region will undergo decay at a faster rate than the other (depending on which time limit is critical to show critical damage), results in the following:

$$Lag(t) = \left| \left[\left(\int_0^{t_1} A_1(t-\tau) - 0 \right) d\tau \right] + \int_{t_1}^{T} A_1(t-\tau) d\tau \right| \tag{10}$$

Which gives:

$$Lag(t) = \left| \left[\left(\int_0^{t_1} A_1(t-\tau) \right) d\tau \right] + \int_{t_1}^{T} A_1(t-\tau) d\tau \right| \tag{11}$$

Hence:

$$Lag(t) = \left| \int_0^{t_1} A_1(t-\tau) + \int_{t_1}^{T} A_1(t-\tau) d\tau \right| \tag{12}$$

Thus

$$Lag(t) = \left| \int_0^{T} A_1(t-\tau) d\tau \right| \tag{13}$$

Equation (13) indicates that a main area of damage is present in the tested composite structure.

The previous set of equations can be applied to any number of adjacently charged regions by extending (7) to include a set of time limits. Thus, allowing for various levels of charged regions to discharge their thermal energy and resulting in a tabulation of regions that contain fiber-matrix problems. This will allow for detailed analysis of causing effects and an enhanced process of manufacturing.

Fig. 9 show another dimension to the kind of relationship between the observed charged regions plotted over additional time intervals. From the figure, the following is observed:

1) An almost symmetrical difference curve as a function of time.

2) The maximum Area (pixel) difference occurs at t=18 minutes. Thus indicating a total separation between the two observed regions and the start of each region specified by an area of pixel population to diminish over a period of time.

3) The difference curve highlights what the Lag function describes and could alternatively be represented as in (14):

$$Diff = \left| \left(\frac{dA_2}{dt} \right) - \left(\frac{dA_1}{dt} \right) \right| \tag{14}$$

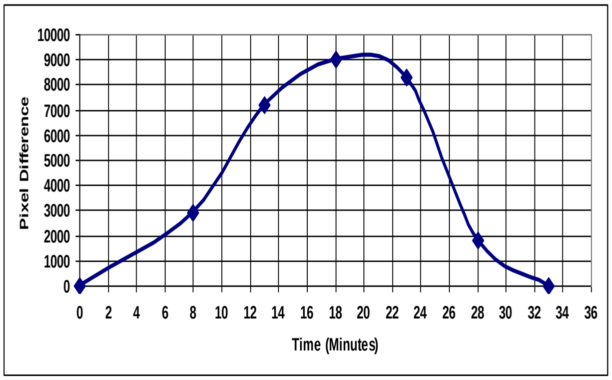


Fig. 9. Relationship (difference) between the two thermally charged regions in the tested composite.

Equation (14) also represents a time-relative change of segmented (damaged) areas that are separated by edges.

V. Conclusions

The proposed technique is an excellent beginning to Characterize damage in composite structures through edge detection and region segmentation using gradient field as a function of time displacement. From the obtained images and the mathematical model, the behavior of the charged regions in relations to each other can be quantified in Table I. The initial fully charged sample is (120 by 120 by 3 =43200). The factor of 3 is due to the conversion from color to gray level. The discharging relationship is presented in Fig. 10.

TABLE I. Relationship between Thermally Charged Regions

Time (min)	**A_1**	**A_2**	**Ratio (A_2 : A_1**
t=t+	43200	43200	Initial Charge
8	28500	25600	0.9
13	21600	14400	0.7
18	15000	6000	0.4
23	8700	400	0.1
28	1800	0	0

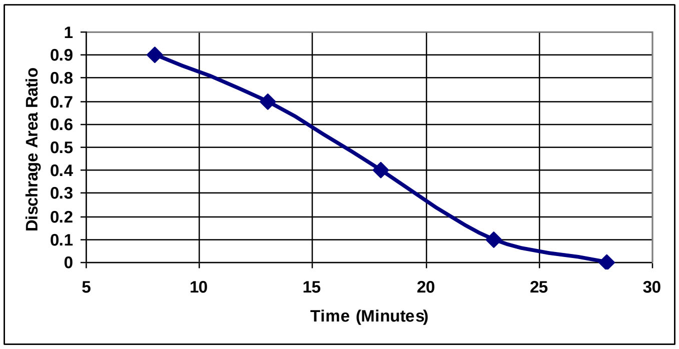


Fig. 10. Discharge relationship between two damaged areas in a composite structure.

It is evident the power law behavior of the discharging process of both damaged areas as a function of time and composite structure properties. The power law curve shows similar characteristics to Ostwald de Waele power law Model.

The presented technique is promising, and leads to potential development of detailed mathematical models covering the thermal behavior of composites by applying a more comprehensive version of the Gradient Field technique that takes into account the presented mathematical model, the optimum threshold level, type and structure of the tested composite component and various sources of thermal energy.

References

[1] P.Suganya, S.Gayathri and N.Mohanapriya," Survey on Image Enhancement Techniques", International Journal of Computer Applications Technology and Research vol. 2, no. 5,pp. 623 - 627, 2013.

[2] C. Jung, Q.Yang, T. Sun, and Q. Hyoseob Song," Low Light Image Enhancement with Dual-Tree Complex Wavelet Transform", J. Visual Communication and Image Representation, vol.42, pp. 28-36, 2017.

[3] P. Geng, X. Su, T. Xu , and J. Liu " Multi-modal Medical Image Fusion Based on the Multiwavelet and Non sub sampled Direction Filter Bank" International Journal of Signal Processing, Image Processing and Pattern Recognition vol.8, No.11 pp.75-84, 2015.

[4] K.Santhi, R.WahidaBanu, "Adaptive contrast enhancement using modified histogram equalization", Optik - International Journal for Light and Electron Optics, Elsevier, vol. 126, pp.1809 – 1814, 2015.

[5] R. Atta, R. Abdel-Kader, "Brightness preserving based on singular value decomposition for image contrast enhancement", Optik - International Journal for Light and Electron Optics, Elsevier, vol. 126, pp.799 – 803, 2015.

[6] H. Xu, Q. Chen, C. Zuo, C. Yang, and N. Liu, "Range limited double-thresholds multi-histogram equalization for image contrast enhancement", Optical Review, Springer, vol. 22, pp.246 – 255, 2015.

[7] K. Singh, R. Kapoorand, and S. Sinha "Enhancement of low Exposure Images via Recursive Histogram Equalization Algorithms", Optik - International Journal for Light and Electron Optics, Elsevier, vol. 125, pp.1385 – 1389, 2015.

[8] Y. Guo, R. Wei, and Y. Liu, "Weighted Gradient Feature Extraction Based on Multiscale Sub-Blocks for 3D Facial Recognition in Bimodal Images," Informatics, vol. 9, no.48, pp. 529–551, 2018.

[9] H. Lee J. Jeon, J. Kim, and S. Lee, "Structure - Texture Decomposition of Images with Interval Gradient", Computer Graphics Forum, vol.36, no.6, pp. 262-274, 2017.

[10] A. Furnari, G. Farinella, A. Bruna, and S. Battiato a, "Distortion adaptive Sobel filters for the gradient estimation of wide angle images," J. Vis. Commun. Image R, vol. 46, pp. 165-175, 2017.

[11] L. Huang, W. Zhao, and Z. Jun Wang, "An advanced gradient histogram and its application for contrast and gradient enhancement", Optik - International Journal for Light and Electron Optics, Elsevier, vol. 31, pp.86 – 100, 2015.

[12] Q. Songa, Y. Wanga, K. Baia, "High dynamic range infrared images detail enhancement based on local edge preserving filter", Infrared Physics & Technology, vol.77, pp. 464-473, 2016.

[13] M. Iskandarani, "Correlating and Modeling of Extracted Features from PVT Images of Composites using Optical Flow Technique and Weight Elimination Algorithm Optimization [OFT-WEA]", Journal of Computer Science, vol.13, no.9, 371-379, 2017.

[14] R.Usmantiagar, P. Venegas, J. Guerediaga, L. Vega, and I. Lopez,"Feature extraction and analysis for automatic characterization of impact damage in carbon fiber composites using active thermography", NDT & E International, vol. 54, pp. 123–132, 2013.

[15] L. Ma, D. Liu, "Delamination and fiber-bridging damage analysis of angle-ply laminates subjected to transverse loading", International Journal of Imaging Systems and Technology, vol. 50, pp. 3063–3075, 2016.

[16] F.Wang, J. Liu, Y. Liu, and Y. Wang, "Research on the fiber lay-up orientation detection of unidirectional CFRP laminates composite using thermal-wave radar imaging", International Journal of Imaging Systems and Technology, vol. 84, pp. 54–66, 2016.

[17] T.Liang, W. Ren, G. Tian, M. Elradi, Y. Gao "Low energy impact damage detection in CFRP using eddy current pulsed thermography", Composite Structure, vol. 143, pp. 352–361, 2016.

[18] B.Ashwini, B. Yuvaraju, "Feature Extraction Techniques for Video Processing in MATLAB," International Journal of Innovative Research in Computer and Communication Engineering, vol. 4, pp. 5292–5296, 2016.

[19] Xiangzhi Bai, "Morphological infrared image enhancement based on multiscale sequential toggle operator using the opening and closing as primitives", Infrared Physics and Technology, Elsevier, vol.68, pp. 143 – 51, 2016.

[20] W. Zhao, Z. Xu, J. Zhao, F. Zhao, and X. Han, "Variational infrared image enhancement based on adaptive dual-threshold gradient field equalization", Infrared Physics & Technology, Elsevier, vol. 66, pp.152-159, 2014.

[21] F. Ciampa, P. Mahmoodi, F. Pinto, and M. Meo, "Recent Advances in Active Infrared Thermography for Non-Destructive Testing of Aerospace Components", Sensors, vol. 609, pp.1-37, 2018.

Evaluating M-Learning in Saudi Arabia Universities using Concerns-Based Adoption Model Level of use Framework

Mohammed Al Masarweh
Management Information System, College of Business in Rabigh,
King Abdulaziz University, Rabigh, Saudi Arabia

Abstract—**Numerous studies have evaluated aspects of m-learning use in Saudi Arabia, mostly focused on technology use and its impact on students, or technology challenges and promises. Few studies have explored features of m-learning use and engagement among university faculty members. This paper presents a new methodology for evaluating the status of m-learning from faculty members' perspectives in Saudi Arabia by investigating level of use using Concerns-Based Adoption Model framework. Concerns-Based Adoption Model is well established in the United States of America and in research investigating innovation adoption in education, including recent efforts in the Middle East (Jordan and Saudi Arabia). The outcome of such research, including this study, promotes better use and engagement with m-learning and provides a better understanding of advantages, disadvantages and barriers. The outcomes of this research study can reflect positively on universities' status in the future and help in reforming policies and practices for developing the use of m-learning in Saudi Arabia.**

Keywords—Concern Based Adoption Model (CBAM); evaluation; M-learning; Saudi Universities; level of use; mobile phone

I. Introduction

Interest in m-learning in the Kingdom of Saudi Arabia (KSA) has grown immensely in recent years as a result of the rapid increase in mobile technologies, wireless networks, and the capabilities of today's mobile devices, facilitated by massive infrastructure and educational investment. The Saudi government is highly concerned to diversify the national economy and reduce oil dependency, and establishing a knowledge-based service economy is a key to this. Thus, more interest has been targeted to investing in IT and m-learning projects for Saudi educational institutions [1]. These range from simple SMS services for individuals and groups to complex operations that related to managing students and distributing learning materials.

Studies have revealed that positive attitudes among students towards m-learning improve adoption of and engagement with educational technologies and m-learning can promote the learning process to shift from a teacher-centric model to a learner-centric approach. Moreover, different research studies have been oriented towards investigating m-learning in terms of adoption, benefits, obstacles and future development needs in KSA [2]. However, few studies have considered issues related to faculty members' use of educational technologies, and defining their effects, impacts and challenges for future enhancements in use and policies. The exact engagement with m-learning by faculty members in KSA and level of use is thus unknown in Saudi m-learning literature, although different tools and frameworks are present for outlining and defining engagement levels [3].

The Concern Based Adoption Model (CBAM) is a framework known for its uses in investigating instructors' practices and providing a formative and summative evaluation methodology. CBAM consists of three different tools used for assessment, each focusing on a specific dimension with specific characteristics and strengths as a tool. The first tool is known as "Stages of Concern" (SoC), which is a quantitative questionnaire measuring users' feelings towards and innovation. The second tool is known as "Levels of Use" (LoU) which is an open-ended questionnaire based on a predefined matrix for measuring instructors' actions in eight behavioral categories against the variety of use. The last tool in this framework is known as "Innovation Configuration" (IC), which is a map of verbal description of the components of an innovation that describes the actions related to each component with differences on evaluating the actions, from "poor" to "ideal" [4].

CBAM tools can be used separately if needed in order to investigate different dimensions of m-learning adoption, or they can be used together sequentially, starting from SoC and ending with IC, to have a wider scope of understanding through assessing, monitoring and understanding different aspects of the implementation process related to m-learning. The CBAM tools provide a consistent and coherent taxonomy to describe emotional and behavioral domains with respect to the innovation used. CBAM has been widely deployed to study m-learning contexts in the United Stat of America (USA), Canada, Australia and more recently in the Middle East, specifically Jordan [4], [5].

This research is the first study to use the CBAM Levels of Use (CBAM-LoU) model to assess the use of m-learning in KSA. It is believed that the use of CBAM-LoU will provide a better understanding for m-learning implementation in KSA, which will promote better use, engagement and policies in the future. The next section presents CBAM-LoU and explains the use of this tool.

The structure of this paper is as follows. Section II discusses the levels CBAM. Section III describes the research methodology. Section IV presents the data analysis and results. Section V gives further discussion of the findings and Section VI concludes the paper.

II. CBAM Levels of Use (CBAM-LoU)

CBAM-LoU focuses on behavior and does not pay any attention towards attitudes, emotions and feelings. Moreover, it does not concentrate or investigate the quality of the researched innovation. It focuses on the behavior of a user or group and classifies them into eight different levels relating to use of the innovation [6]. The first three levels (0-II) are for the non-use of innovation, and the remaining five levels (III-VI) are for users using the innovation. CBAM-LoU has decision points that are triggered when users shift from one level of use to another. Table I shows the level of use with respect for each decision points used [6], [7]. Each presented level has seven different categories of behavioral indicators [7], [8], as shown in Table II.

TABLE I. CBAM Level of Use Categories and Decision Points

Category level	Category Name	Discription
Level 0	Non-use	In this level, participants have little or no knowledge of using the innovation, and show no interest or action towards becoming involved.
Decision Point A		Participants should act towards learning more about the used innovation
Level I	Orientation	Users in this level have developed or are developing information about the innovation, and have discovered its value and demands.
Decision Point B		Participants present decision to use the innovation through starting a time to begin
Level II	Preparation	Participants are organizing for the first interaction with the innovation.
Decision Point C		Participants performs user oriented changes
Level III	Mechanical Use	Participants make short-term effort with daily engagement with the innovation, with little time for reflection. The uses in this level are mainly disjointed and shallow.
Decision Point D-1		Participants form a routine pattern of use.
Level IVA	Routine	Participants’ use of the innovation is stable and few changes are made over the period of use. Few efforts are made to develop or improve the use or its consequences.
Decision Point D-2		Participants adjust the use of the innovation according to formal or informal evaluation to improve expected benefits.
Level IVB	Refinement	Participants vary the use of the innovation to increase benefits within the direct scope of influence. Differences are based on understanding of short- and long-term consequences.
Decision Point E		Participants starts change in the usage of the innovation for the advantage of learners and in coordination with other colleagues to enhance the benefits
Level V	Integration	Participants bring their efforts with each other for related activities in order to reach collective effect within the scope of influence.
Decision Point F		Participants starts exploring different modifications to the innovation they are currently involved with.
Level VI	Renewal	Participants re-asses the quality of interaction and use of the innovation, searching for possibilities of performing main modifications or alternatives, in order to bring better impacts of the innovation to their practices, students and the system.

CBAM-LoU outlines each participant in only one level, unlike CBAM’s SoC that defines users in different stages. In addition, the results of CBAM-LoU evaluation must be perceived as development status and not as summative end state [8]. The previously presented categories for each LoU are all concerned with action performed by participants except for the knowledge category, which deals with observable behaviors, dealing with understanding about the innovation and its use and effects.

The used categories represent subparts for each LoU that define each LoU through presenting more in depth description of a part of LoU, thus it makes it possible to have different data points for determining LoU. Each presented category can be evaluated.

Distinctly and the groupings of ratings can be used to define the overall LoU [9].

TABLE II. CBAM Behavioural Categories

Category Name	Description
Knowledge	Participants know how to use the innovation and its characteristics and consequences of use. Knowledge is cognitive and related to the use of the innovation, not to feelings or attitudes.
Acquiring Information	Participants ask for information about the innovation in different ways, such as asking about resources related to the innovation, persons, agencies and materials.
Sharing	Participants discuss the innovation use with other colleagues, and share plans, thoughts, resources, outcomes and challenges related to the use of the innovation.
Assessing	Participants explore the possible or actual use of the innovation, through mental evaluation or actual collection and analysis of data.
Planning	Participants make short- and long-plans during innovation adoption, such as scheduling, aligning resources, activities and coordinating the use of the innovation.
Status Reporting	Participants report their personal stand at the current time in regard to the use of the innovation.
Performing	Participants perform actions and activities related to the innovation.

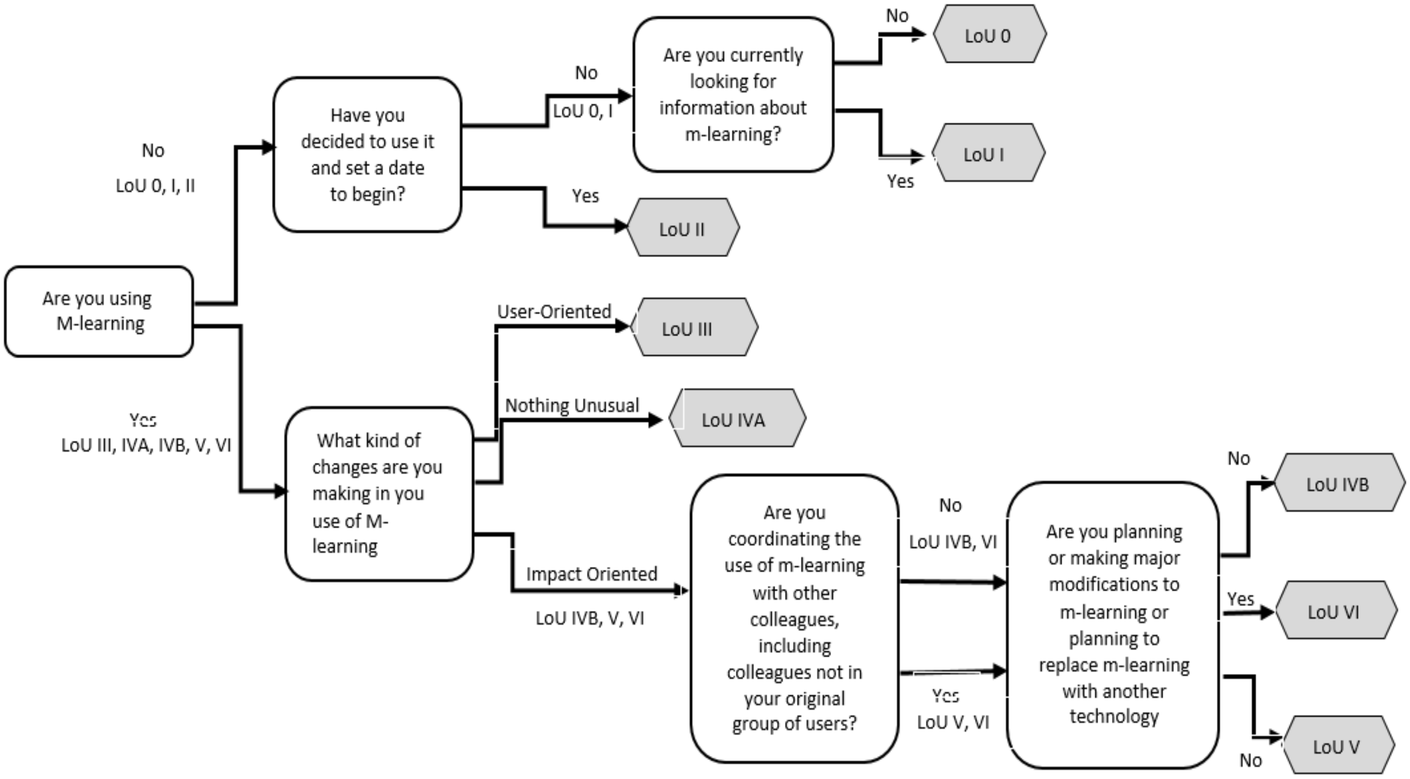

Are you using M-learning
No
LoU 0, I, II
Have you decided to use it and set a date to begin?
No
LoU 0, I
Are you currently looking for information about m-learning?
No
LoU 0
Yes
LoU I
Yes
LoU II
Yes
LoU III, IVA, IVB, V, VI
What kind of changes are you making in you use of M-learning
User-Oriented
LoU III
Nothing Unusual
LoU IVA
Impact Oriented
LoU IVB, V, VI
Are you coordinating the use of m-learning with other colleagues, including colleagues not in your original group of users?
No
LoU IVB, VI
Yes
LoU V, VI
Are you planning or making major modifications to m-learning or planning to replace m-learning with another technology
No
LoU IVB
Yes
LoU VI
No
LoU V

Fig. 1. Branching technique used to define level of use category.

III. Research Methodology

This research methodology is based on the specifications of CBAM-LoU tool applied to faculty members in Saudi universities concerning the use of m-learning.

The evaluation of participants was based on focused interviews that used a branching technique based on the output obtained from each participant. The used branching technique is presented in Fig 1.

CBAM-LoU reliability, validity and internal consistency have been demonstrated by different research studies for use with more than eleven educational tools [9]-[11]. The research sample consisted of 347 faculty members who agreed to participate in this research study, recruited by email from six different universities. A total of 119 faculty members were defined as users for m-learning, and 228 were classified as non-users. All participation was voluntary and fully informed consent was obtained. Interviews were performed in different settings and locations chosen by the participants and with respect for each university, as shown in Table III.

Classification distributed among different stages based on the derived results. The following table shows the categories and stages that are defined by CBAM-LoU tool [1], [13].

In terms of the gathered data, descriptive analysis was undertaken using SPSS V22 in order to define descriptive numerical data (frequencies) related to the LoU for all participants. The following section presents the output of this research study.

TABLE III. Participant Groups and Universities

University	M-learning groups		Total number	
	Users	*Non-users*	*Users*	*Non-users*
King Abdulaziz	4	4	42	30
Umm Al Qura	2	3	33	50
Taibah	3	5	19	32
Majmaah	2	6	12	38
Islamic (Madina)	3	4	8	33
King Khalid	1	6	5	45
Total	15	26	119	228

IV. CBAM-LoU Results

A. Participants using M-learning

Table IV shows the results of 119 faculty members using m-learning from six different universities in KSA.

The results indicate that participants are classified as mechanical use of m-learning according to the behavioral indications of CBAM-LoU. As noted previously, LoU is different from CBAM's Stages of Concern in that it defines only one level with the highest percent shown from the results [7], [8], [12].

B. Participants Not using M-learning

Table V shows the results for 228 participants not using m-learning in Saudi universities according to the CBAM-LoU matrix [7], [9].

Based on CBAM-LoU, the results for non-users of m-learning are classified as orientation based on the behavioral indications provided by LoU methodology.

V. CBAM-LOU RESULT'S DISCUSSION

This section discusses the results related to the status of using m-learning in Saudi universities based on the results shown in Tables IV and V.

A. *CBAM's LoU- Results Discussion (Using M-Learning)*

As stated before that LoU is based on semi-structured interview and it has 8 questions for participants that were used to outline their level of uses based on the predefined categories. The following discussion will be related to the results shown in (Table IV) with respect for the CBAM's LoU categories.

1) Knowledge category

In this category the following question was used: (What are the strengths and challenges of m-learning according to you, and have you tried to overcome challenges?). The results had two answers that identified participants as mechanical and routine use (86% and 14%, respectively). This category had different answers, and the majority of participants agreed that the main strength of m-learning was its availability and easy access anytime, anywhere.

TABLE IV. CBAM LOU RESULTS FOR (USING M-LEARNING)

Categories	**Level of Use**	**Frequency**	**Percent**
Knowledge	Decision point C mechanical use Decision point D-1 routine	102 7	86% 14%
Acquiring Information	Decision point C mechanical use Decision point D-1 routine	112 7	96% 6%
Sharing	Non-Use Decision point C mechanical use Decision point D-1 routine Decision point D-2 Refinement	79 30 7 3	66% 25% 6% 3%
Assessing	Non-Use Decision point C mechanical use Decision point D-1 routine Decision Point E- Integration	96 18 3 2	81% 15% 3% 2%
Planning	Decision point C mechanical use Non-Use Decision point D-1 routine Decision point D-2 Refinement	59 31 21 8	50% 26% 18% 7%
Status Reporting	Non-Use Decision point D-1 routine Decision point C mechanical use	108 6 5	91% 5% 4%
Performing	Non-Use Decision point C mechanical use Decision point D-1 routine	58 40 21	49% 34% 18%

TABLE V. CBAM LOU RESULTS FOR (NOT USING M-LEARNING)

Categories	**Level of Use**	**Frequency**	**Percent**
Knowledge	None Use Decision Point A Orientation	53 175	23% 77%
Acquiring Information	None Use Decision Point A Orientation	74 154	32% 68%
Sharing	None Use Decision Point A Orientation	191 37	84% 16%
Assessing	None Use Decision Point A Orientation	177 51	87% 22%
Planning	None Use Decision Point A Orientation	195 33	86% 14%
Status Reporting	None Use Decision Point A Orientation	189 39	83% 17%

Moreover, they agreed that many used applications facilitated collaborative learning and enhanced learners' engagement with the materials. In addition, the majority agreed that m-learning encourages self-paced learning and it addresses different learning styles (e.g. reading text, watching videos, listening to podcasts, working with interactive material and researching on the internet). On the other hand, the majority of participants agreed on some challenges related to m-learning, such as the connectivity in many cases can be a major challenge for m-learning and the challenge of screen size limitation.

Furthermore, the mentioned device compatibility issues with some technologies and finally they agreed that using mobile phones for learning can cause distractions due to phone calls, instant messaging and other applications. In terms of overcoming challenges, the majority of participants agreed that faculty members must be trained to create content suitable for mobile devices and to be educated on using m-learning strategies.

2) Acquiring information

The following question was used for this category: (At the present time are you looking for information related to m-learning, and if so, for what kind and for what purpose?). The answers identified mechanical use and routine use (96% and 6%, respectively). Mechanical use participants agreed that they are not currently seeking additional information related to m-learning, as they are overwhelmed by the currently available applications and tools. On the other hand, routine use participants agreed that they are seeking different information related to m-learning, especially those used for creating micro-learning objects to be used and distributed as learning content.

3) Sharing

In this category the following question was used: (Do you discuss m-learning with colleagues, and what do you tell them?). The results included non-use (66%), mechanical use (25%), routine use (6%) and refinement use (3%).

The non-use group agreed that they do not share any information related to m-learning, while the group defined as mechanical use agreed that they share some basic information related to some common tools and applications that are widely used in their institutions. The routine use group agreed that they share information about uses of some applications and its common uses among their close colleagues, while the refinement group agreed that they share the uses of m-learning tools and applications among different friends and on social media networks related to education and technological use, and they try to educate others on the benefits of using those applications.

4) Assessing

This category had two broad questions: (Based on your experience what is the effect of m-learning, how do you determine this, are you performing any formal or informal evaluation for using m-learning, and have you received feedback from students or colleagues, and what have you done with the information you got from the assessment?); and (Have you recently changed your use of m-learning, what, why and how recently are you considering making any changes?). The results from this category and the used questions came into four different results: non-use (96%), mechanical use (15%), routine use (3%) and integration use (2%).

The results show that the majority of participants in Saudi universities are not concerned with assessing their use of m-learning. This practice gives an indication for the need to educate faculty members in KSA about the importance of assessing their use of m-learning and its effect on future implementations and pedagogy in the educational sector. The mechanical use group agreed that they do some basic non-formal assessment for some features and uses of mobile applications and tools. The routine group agreed that they are performing informal assessment for their use of m-learning. The informal assessment is based on their daily use and discussion with other colleagues.

The integration group participants agreed that they are performing formal and informal assessment for their use of m-learning. The informal assessment is based on their discussion with other colleagues about the features and their impact on learning. On the other hand, the formal assessment is based on students' performance and output from participating and engaging with m-learning tools and applications.

5) Planning

In this category the following question was used: (What plans do you have for m-learning use?). This category had four different results: mechanical use (50%), non-use (26%), routine use (18%) and refinement (7%). The results show that half of participants are mechanical users with no real plans for future use of m-learning. All participants in this group showed no real plans for future use and enhancements. The non-use showed no interest in making any plans for future use of m-learning and mentioned that no plans are being made. The routine use group agreed that they are making small plans for enhancing their use and engagement with tools and applications used for educational purposes. The refinement group agreed that they are having serious plans for exploring different tools that are being used within the educational context, in KSA and abroad.

6) Status reporting

This category had the following question: (Are you coordinating your work with other colleagues out of your institution on the use of m-learning, and have you made any changes on m-learning based on your coordination?). The majority of participants agreed on not coordinating or working with others, thus they are classified as non-use (91%), with 5% of routine use participants showing some coordination with their colleagues outside their institutions using social media channels and educational groups. Mechanical use (4%) participants showed little and shallow coordination with some colleagues in different universities. The routine use group mentioned that they have made some changes related to m-learning use of tools and applications in that they have discovered better tools through reporting their use of existing tools and applications to others.

7) Performing

In this category the following question was used: (Are you planning to make major changes or to replace m-learning use at this time?). Three different groups emerged from answers: non-use (49%), mechanical use (34%) and routine use (18%). The first groups' answers stated that no planning had been made to change or replace m-learning. The second group (mechanical use) agreed that they have plans to change and enhance their use of m-learning through adopting different applications and tools and that they need to consider the use of m-learning with different subjects being taught and to encourage students on more interaction.

The routine use group agreed that they are seriously considering making major changes in their use and interaction with m-learning through exploring different tools used in education by different foreign institutions worldwide. Moreover, they agreed that they need to enhance the collaboration level among different faculty members in different universities to share experiences and knowledge. In terms of replacing m-learning, all participants agreed that m-learning is the latest technology being used by them and their institutions and no other technology is present that can act as substitute to m-learning.

The previous results for all categories were shown that the interaction level for participants with m-learning ranges from basic to intermediate. Moreover, it shows that the interaction level with m-learning is defined as mechanical use, as participants' efforts are based on short-range, day-to-day use of m-learning with little time for reflection. As stated before, mechanical use is related with participants directing their efforts to become proficient at tasks required for using m-learning, and their use is generally considered fragmented and shallow.

On the other hand, a smaller group of participants showed more interest and engagement with m-learning activities, and this group was defined as routine use. The routine use group showed different and simple changes in their ongoing use, and they provided some efforts and views to improve their use of m-learning. For some categories, participants' attitude was classified as non-use, in the categories sharing, assessing, planning, status reporting and performing.

The results of CBAM-LoU for participants in this study were identified as mechanical, as CBAM identifies one level of use only. The next section explains the results related to participants who did not use m-learning and discusses their status based on CBAM methodology.

B. CBAM's LoU- Results Discussion (Not using M-Learning)

The second group of 228 participants found not using m-learning has been interviewed using CBAM-LoU matrix, and this section will provide the discussion for their results with respect for LoU categories.

1) Knowledge category

For this category the following two questions were used: (Have you decided to use m-learning in the future, if so, when?); and (How do you describe m-learning as you see it?). The results indicated non-use (23%) and orientation (77%). The non-use groups were found to be uninterested in adopting m-learning as part of their educational activities and routine. They did not show any positivity for adopting m-learning in the future, nor did they provide satisfactory description of m-learning as they focused on the hardware aspects and some of its challenges rather than tools and services.

The orientation group showed positive attitudes towards adopting m-learning within their educational activities, and they agreed on using m-learning in the near future. They managed to describe m-learning in satisfactory manner, as they focused on the mobility, tools and applications that can be used in education. Moreover, it was found that they have not been using m-learning because of high responsibilities and being busy with many educational tasks and activities. In addition, they believed that m-learning will add more responsibilities that they cannot adopt; especially as such practices are not being recognized or encouraged by their institutions or educational policies.

2) Acquiring information category

In this category the following question was used: (Are you currently searching for any information related to m-learning, and if so, for what kind and for what purpose?). The answers for this category indicated non-use (32%) and orientation (68%). The non-use participants showed no interest in knowing more about m-learning and the wide selection of services it can offer. They stated that they are satisfied with the traditional approach or the current knowledge they have about m-learning.

On the other hand, the orientation group showed a positive attitude for learning more about m-learning and the services it offers related to content creation, sharing resources and providing collaboration among users. Moreover, they were interested in knowing about tools and services that can support their teaching, with the focus on minimizing efforts and fatigue inherent in traditional approaches of teaching.

3) Sharing category

In this category the following question was used: (Do you share information about m-learning with others, and what do you share?). The answers for this category indicated non-use (84%) and orientation (16%). The non-use group agreed that they do not discuss or share any information related to m-learning use, tools, services and applications. On the other hand, a small percent of the orientation group showed positive attitudes as they discussed and shared information with other users. The shared information was related to the perceived benefits of m-learning, tools and services that are used in education and the possibility of saving efforts using m-learning technology.

4) Assessing category

For this category two questions were used: (What are the assets and flaws of m-learning for your situation?); and (What questions do you ask related to m-learning use, giving examples if possible?). The results indicated non-use (87%) and orientation (22%). The non-use group did not specify or identify any assets related to m-learning, but they focused on the flaws they mentioned, such as that m-learning is time-consuming and not beneficial compared with the traditional approach and activities, or even if compared with internet use through personal computers.

In terms of the questions asked related to m-learning use, the most frequent questions were related to the benefits of m-learning on educational activities, such as how m-learning supports different pedagogies, assessments and user engagements. On the other hand, the orientation group recognized some benefits for using m-learning as they mentioned it might help in learning more about different technologies and bridging the gap between Saudi and foreign universities. Moreover, they mentioned that using m-learning can provide the same benefits of e-learning. In terms of flaws, they mentioned the lack of Arabic digitized resources, tool and services, and the absence of consensus of specific tools and services in education are another main flow.

In terms of questions and concerns related to m-learning, the main concerns were related to possibilities of supporting m-learning with policies and recognition for the efforts. The possibilities of providing tools in Arabic language were the main concern for many faculty members KSA.

5) Planning category

The following question was used with this category: (Do you have any plans or preparations for m-learning adoption?). The results indicated non-use (86%) and orientation (14%). The majority of participants mentioned that they do not have plans or preparations for m-learning use, especially as most of the tools are presented in non-Arabic languages and there is no agreement on using specific tools in educational settings in KSA or having a policy for such use. On the other hand, the orientation group mentioned that they need to have training on the use and benefits of m-learning, and they need better management skills for their traditional educational activities. Moreover, they focused on the need for support by their institutions to be able to have serious plans of use and adoption.

6) Status reporting category

For this category the following question was used: (Where do you see yourself in relation to the use of m-learning?). The answers indicated non-use (83%) and orientation (17%). The non-use group agreed that they are currently not using m-learning as they are satisfied with traditional learning and with the use of e-learning through Personal Computers (PCs). On the other hand, the orientation group agreed that they are interested in m-learning but there are some serious obstacles for adopting such services and practice. The challenges are related to policies, time management, language and content creation.

Investigating participants not using m-learning gave better understanding for the challenges facing m-learning adoption in Saudi universities. Different challenges that act as barriers for m-learning use were identified, and it was found that eliminating those challenges will enhance the adoption by the groups defined as orientation by CBAM-LoU. However, there will be a group that resists the change and favors traditional approaches only, seeing the obstacles and neglecting the opportunities for enhancing educational outcomes.

VI. Conclusion

The importance of evaluating the use and interaction with educational tools, applications or systems is essential to provide better quality of education and to form and reform appropriate policies. Different evaluation tools are used to evaluate the interaction with m-learning. This research deployed CBAM-LoU tool, as CBAM is a well-used framework for evaluating educational tools within educational settings. Moreover, CBAM-LoU managed to provide different views on the evaluation results through the matrix it provides. This research study managed to evaluate the use of m-learning in six different Saudi universities with 347 participants, using an evaluation methodology based on using semi-structured interviews.

The evaluation sample was divided into two groups, users and non-users of m-learning. The first group consisted of 119 participants, and the results from CBAM-LoU identified them as mechanical use, which defines participants as having their focus on short-term daily engagement with m-learning with little time for reproduction. Their efforts are largely focused on learning important issues related to the use of the educational tool, thus their use is fragmented and shallow. In terms of challenges related to that group, it was found that using m-learning adds more responsibilities for managing their tasks and activities. Moreover, it was found that their practice of m-learning use is based on self-efforts and less is related to university's support and policies.

The categories that had challenges and which need more attention from Saudi management and officials in universities are sharing, accessing, planning, status reporting and performing. On the other hand, the second group consisted of 223 participants who are not using m-learning. The CBAM-LoU matrix with semi-structured interviews was performed and the results identified some participants as the non-use category. This category classifies users as having little or no knowledge of using m-learning, and participants are not showing interest or action towards becoming involved.

In addition, it was found that participants are favoring the traditional approach of teaching and the use of e-learning using personal computers. According to CBAM-LoU, participants showed the lowest impact in the categories sharing, assessing, planning and status reporting.

These results give a good indication about the level of use and the status of m-learning in Saudi universities. The information from this research study can be used to promote better practice and engagement with m-learning through the defined strengths and challenges, and through planning for better training and policy settings by university officials and decision makers.

References

[1] A. Khan, H. Al-Shihi, Z. A. Al-Khanjari, & M. Sarrab, "Mobile Learning (M-Learning) adoption in the Middle East: Lessons learned from the educationally advanced countries. Telematics and Informatics", 32(4), 909-920, 2015..

[2] O. Al-Hujran, E. Al-Lozi, & M. Al-Debei, "Get Ready to Mobile Learning": Examining Factors Affecting College Students' Behavioral Intentions to Use M-Learning in Saudi Arabia". Jordan Journal of Business Administration, 10(1), 111-128, 2014.

[3] M. Sarrab, M. Elbasir, & S. Alnaeli, "Towards a quality model of technical aspects for mobile learning services: An empirical investigation". Computers in Human Behavior, 55, 100-112, 2016.

[4] N. Matar, "Evaluating E-Learning System Use by CBAM-Stages of Concern Methodology in Jordanian Universities". World of Computer Science & Information Technology Journal, 5(5), 2016.

[5] N. Matar, "Presenting Structured Evaluation Framework Towards E-Learning Adaption In Jordanian Universities-The Use Of Cbam-Soc Framework". Journal Of Theoretical And Applied Information Technology, 95(5), 1008, 2017.

[6] D. Niederhauser, & D. Lindstrom "Instructional Technology Integration Models and Frameworks: Diffusion, Competencies, Attitudes, and Dispositions. Handbook of Information Technology in Primary and Secondary Education, 1-21, 2018.

[7] N. Matar, "Defining E-Learning Level of Use in Jordanian Universities Using CBAM Framework". International Journal of Emerging Technologies in Learning (iJET), 12(03), 142-153, 2017.

[8] A. Galloway, & D. Gutmore, "What concerns faculty about teaching online: The effect of organizational structuration to teaching online". In E-Learn: World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education (pp. 661-672). Association for the Advancement of Computing in Education (AACE). 2016.

[9] S. Kim, & S. Paik, "An Analysis of Science Teachers' Stages of Concern and Levels of Use on Descriptive Assessment". Journal of the Korean Chemical Society, 60(5), 2016.

[10] J. Wyant, E. Jones, & S. Bulger, "A mixed methods analysis of a single-course strategy to integrate technology into PETE". Journal of Teaching in Physical Education, 34(1), 131-151, 2015.

[11] J. Tondeur, J. van Braak, P. Ertmer, & A. Ottenbreit-Leftwich, "Understanding the relationship between teachers' pedagogical beliefs and technology use in education: a systematic review of qualitative evidence". Educational Technology Research and Development, 65(3), 555-575, 2017.

[12] B. Barrio, & B. Combes, "General education pre-service teachers' levels of concern on Response to Intervention (RTI) implementation". Teacher Education and Special Education, 38(2), 121-137, 2015.

[13] J. Golden, & V. Brown, "A Holistic Professional Development Model: A Case Study to Support Faculty". Handbook of Research on Teacher Education and Professional Development, 259.Chicago, 2016.

Automatic Arabic Image Captioning using RNN-LSTM-Based Language Model and CNN

Huda A. Al-muzaini, Tasniem N. Al-yahya, Hafida Benhidour
Dept. of Computer Science
College of Computer and Information Sciences
King Saud University
Riyadh, Saudi Arabia

***Abstract*—The automatic generation of correct syntaxial and semantical image captions is an essential problem in Artificial Intelligence. The existence of large image caption copra such as Flickr and MS COCO have contributed to the advance of image captioning in English. However, it is still behind for Arabic given the scarcity of image caption corpus for the Arabic language. In this work, an Arabic version that is a part of the Flickr and MS COCO caption dataset is built. Moreover, a generative merge model for Arabic image captioning based on a deep RNN-LSTM and CNN model is developed. The results of the experiments are promising and suggest that the merge model can achieve excellent results for Arabic image captioning if a larger corpus is used.**

Keywords*—AI; image caption; natural language processing; neural network; deep learning convolutional neural network; recurrent neural network; long short-term memory

I. Introduction

Automatic generation of captions for images by describing the content of an image using natural language sentences has become a fundamental task in Artificial Intelligence and has recently attracted the attention of the research community [1]. Given the huge number of images that are available online, image captioning has become nowadays central to image retrieval tasks such as the one carried by search engines or newspaper companies. More specific applications, like describing images for blind persons or teaching children concepts, can also be given as examples on the importance of captioning images.

Image captioning has been identified as a cross-modal task which grounds and relates the visual and the natural language model. Despite the challenging nature of this task, several image caption generation models, one can cite [2]–[6] as examples, have achieved promising results due to the advances in training neural networks [7] and the large image datasets that are now available [8].

The sparsity of annotated resources other than English is an issue in morphological complex language such as Arabic. Thus, there is a need for corpora sufficiently large for image captioning in other languages.

The aim of this work is to take a step towards the goal of developing an image caption generation model for describing images in Arabic language (see Fig. 1). The model is inspired by the merge model proposed in [10] and [11]. It consists of two sub-networks: a deep recurrent neural network (RNN) for sentences and a deep convolutional neural network (CNN) for images. These two sub-networks interact with each other in a merge layer to predicate and generate the caption. Moreover, the first public Arabic image caption corpus is presented. This Arabic version is a subset of the Flickr [11] and MS COCO [12] caption data sets. The remainder of the paper is organized as follows. Influential work as well as state-of-the-art models for image caption generation for English as well as other languages are presented in Section II. Detailed description of the image caption generation model for Arabic language is given in Section III. A description of the image dataset with Arabic captions is presented in Section IV. The process of building the Arabic image caption corpus through crowdsourcing is presented in Section V. The experiment evaluation and results are described in Section VI. Finally, the conclusion with some directions for future work is given in Section VII.

II. Related Work

This section covers recent advances in the development of image caption generation models for different languages including: English, Arabic, Chinese, Japanese, and German.

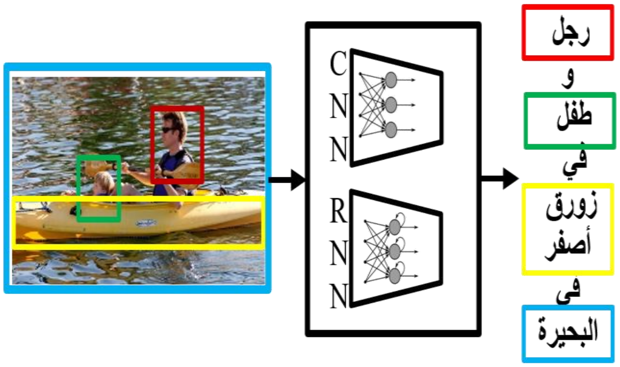


Fig. 1. Example of an Arabic caption generated for an image (caption translation in English: a man and a child in a yellow canoe in the lake).

A. *Image Captioning for English Language*

The different approaches for image caption generation can be either based on retrievable or constructive approaches as pointed out in [9], [10], [13], [14]. This taxonomy is clearly depicted in Fig. 2. An image caption generator based on a retrievable approach models the problem as a retrieval task. A database based on image features and captions is constructed. Given an image, the most suitable annotation is retrieved. This approach however lacks the ability of generating novel sentences, does not scale to describe raw images, and the caption generation is limited to the features and the size of the database. Thus, this approach is not suitable for today's demand. Example of work based on this approach includes [1], [15]–[17].

Recent progress in automatic image captioning is based on a constructive approach. A constructive approach gradually constructs a novel caption for each image. This can be further divided into computer vision and natural language generation methods (CV/NLG) or CNN/RNN methods. For the first category, image attributes are extracted from images using computer vision techniques which are fed to natural language generation models to generate a syntactically correct caption. This approach is the base of the work in [18]. The CNN/RNN approaches have proven to be the most successful ones. They model the caption generation process in two phases, the first phase is image features learning phase and the second is the sentence generation phase. Depending on whether the image is injected to the language model or left out and then later merged with the output of the language model using a feedforward layer, one can distinguish two models, the inject and the merge model. A complete empirical study of these two models can be found in [10] and [11]. In the inject class (see Fig. 3), the language model, such as the RNN, is the primary generation component where an image is directly injected to the model during training time. The output of the RNN is a mixed vector that is handled in a subsequent feedforward layer to predict the next word in the caption. Works under this class includes [2] and [19]. In [2], the Neural Image Caption (NIC) model is presented. This model is based on an end-to-end neural network that works by first pre-training it for an image classification task using a CNN and then using the last hidden layer as an input to the RNN that generates sentences. Experiments on several datasets including Flickr [11], MS COCO [12], Pascal VOC [17], and SBU [16] using different metrics: BLEU-{1,2,3,4} [20], CIDER [21], and METEOR [22] reported an accuracy comparabale to state-of-the-art approaches; for instance, on the Pascal dataset, NIC yielded a BLEU score of 59, to be compared to the current state-of-the-art of 25, while human performance reaches 69. On Flickr30k, an improvement was achived from 56 to 66, and on SBU, from 19 to 28. In [19], the process starts by decomposing the input image by detecting objects and other regions of interest to produce a vector representation richly expressing the image semantics. This feature vector is taken as input by a hierarchical RNN. The hierarchical RNN is composed of two levels: a sentence RNN and a word RNN. The sentence RNN receives image features, decides how many sentences to generate in the resulting paragraph, and produces an input topic vector for each sentence. Given this topic vector, the word RNN generates the words of a single sentence. The model was experimented on a novel dataset of paragraph annotations, comprising of 19,551 MS COCO [12] and Visual Genome [23] images, and evaluated across six language metrics: BLEU-{1,2,3,4} [20], CIDER [21], and METEOR [22]. The scores show the superior advantages of this method over traditional image captioning methods and was close to human performance.

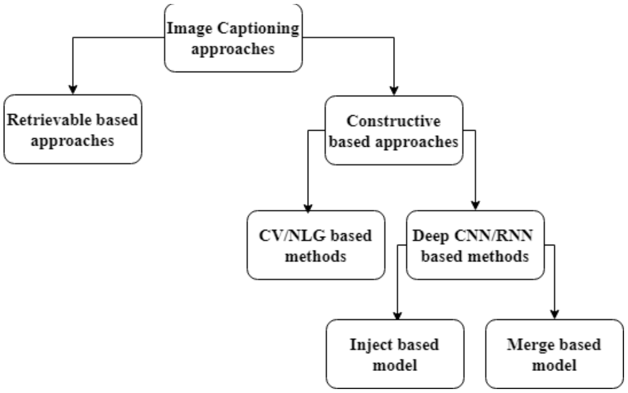


Image Captioning approaches
Retrievable based approaches
Constructive based approaches
CV/NLG based methods
Deep CNN/RNN based methods
Inject based model
Merge based model

Fig. 2. Taxonomy for English image captioning approaches.

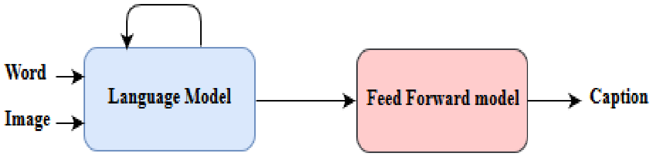


Fig. 3. Inject model.

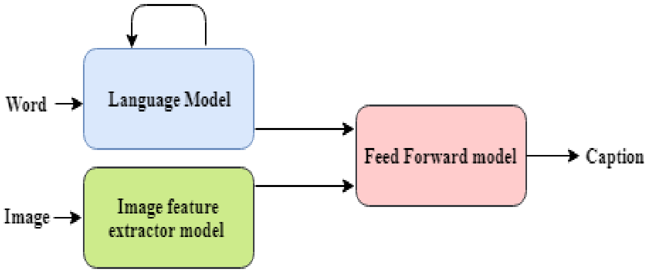


Fig. 4. Merge model.

The second class is the merge model in which the image features and linguistic models are learned independently and then merged in a feed forward model in which the prediction takes place (see Fig. 4). The work of [24] was the first to prpose a merge model for image captioning and shortly after their work was published, several papers appeared with promising results including [4], [25], [26]. This demonstrates the effectiveness of this model. In [24], Mao et al. proposed the merge model then refined it in [4] and [26]. Their image representation is learned independently by a CNN model then inputted to the RNN-LSTM model along with every word in the sentence description. The approach uses the capacity of the RNN-LSTM more efficiently. The RNN-LSTM model incorporates a two-layer word embedding system which learns the word representation more efficiently than the single-layer word embedding. These two models interact with each other in a multimodal layer. The effectiveness of their model was validated on four benchmark datasets IAPR TC-12, Flickr 8K, Flickr30K, and MS COCO. Experimental results based on BLEU-{1,2,3,4} [20], CIDER [21], METEOR [22], and ROUGE [27] showed the outstanding performance of their model for almost all evaluation metrics. In [25], the Deep Compositional Captioner (DCC) is introduced. DCC builds on recent deep captioning models which combines a CNN and RNN networks for visual and language generation, respectively. Then, both models are combined into a deep caption model which is trained jointly on paired image-sentence data. However, unlike previous models which can only describe objects that are present in paired image-sentence data, DCC is able to generate sentences that describe objects presented in unpaired image/data but not present in paired image/sentence data. To accomplish this task, the training is preformed into three stages: 1) CNN and RNN are trained with unpaired data, then 2) both models are combined into a caption model which is trained on paired image-sentence data, and finally, 3) the knowledge is transferred from words that appear in paired image-sentence data to words that do not appear in paired image-sentence data. DCC performance was empirically evaluated by studying results on a training split of the MS COCO [12] dataset by deliberately excluding certain objects. Moreover, DCC performance to describe objects in the ImageNet7k dataset which are not present in the caption datasets was assisted. DCC scored 69.36 and 23.98 on the BLEU and METEOR metrics respectively. In addition, the F1-score was reported, which indicates that DCC can integrate new vocabulary in captions.

The literature on English caption generation although new, is rich of models that have proven their efficiency. However, few explicit comparison between the performance of the inject and merge architectures has been investigated. In [26], the authors compared the inject and merge architectures based only on the BLEU metric and concluded that merge is superior. The first work that studies extensively and systematically the difference between the inject and merge architecture is presented in [10] and [11]. Experimental evaluation concluded the following: 1) inject architectures tends to be slightly better on standard corpus-based metrics such as CIDER [21], 2) merge architectures produce sentences that are rich in vocabulary; that is inject models tends to re-generate captions wholesale from the training data, 3) inject models tend towards more generic and less image specific captions, especially for longer captions; a problem that merge models is not susceptible of, and 4) from an engineering perspective, merge architectures make better use of their RNN memory and avoids overfitting.

B. Image Caption for Arabic language

Automatic image captioning in Arabic was addressed only by the work of [28] by using root-word based RNN and Deep Belief Network (DBN). The approach adopted can be summarized in three stages. In the first stage, a Region CNN (RCNN) [29] is used to map image objects to Arabic root words by the aid of a transducer based algorithm for Arabic root extraction [30]. After that, stage two uses a word based RNN with LSTM memory cell to generate the most appropriate words for an image in Modern Standard Arabic (MSA). Finally, the caption sentences are generated by using dependency tree relations; specifically the Prague Arabic Dependency Treebank (PADT) [31]. For evaluation, two datasets were created. The first consists of annotating 10,000 images from the ImageNet dataset with Arabic captions and the second 100,000 images from Al-Jazeera news website. Experiments show a promising result considering BLEU-1 score with value 34.8 for Arabic caption generation.

C. Image Caption for Other Languages

The limitation of image description corpora in languages other than English is an issue, particularly for morphologically rich languages such Arabic and Japanese. In [32] a Japanese version of MS COCO caption dataset has been created using Yahoo! Crowdsourcing. The authors developed a model for image caption generation for Japanese language using deep learning. They pre-trained the model with the English portion of the corpus to improve the performance then trained it using Japanese captions. The resulting bilingual model has better performance comparing to the monolingual model that uses only the Japanese caption corpus. Cross-lingual image captioning for Chinese language has been developed by applying machine translation [33]. The experiment has been done on Flickr8k-cn and Flickr30-cn datasets. To improve the translated English-Chinese sentences, a fluency-guided learning framework has been proposed using LSTM neural network. The proposed approach improves both the fluency and the relevance without using any manually written caption in Chinese. In [34], an RNN model for generating Chinese captions has been presented. The authors developed two methods, one that takes the list of words from a Chinese sentence as input, and the second takes the list of characters and feed them to the same RNN model. The Chinese caption is obtained by translating Flickr30 dataset from English to Chinese using Google Translation API. They observed that the character level method outperform the word level in this task.

Multi30K, a Germen version of Flickr30K dataset, has been presented in [35]. Each image has a German translation of the English description obtained from Flickr30K dataset and five independent German captions obtained using Crowd flower platform. The translated sentences were collected by professional English-German translators without seeing the image.

III. Methodology

The Arabic image captioning model proposed in this work follows the merge architecture that was previously described in [10], [11]. This architecture is a simplified version of the architecture in [2]. It was chosen for its simplicity whilst still being the best performing system in the 2015 MS COCO [12] image captioning challenge.

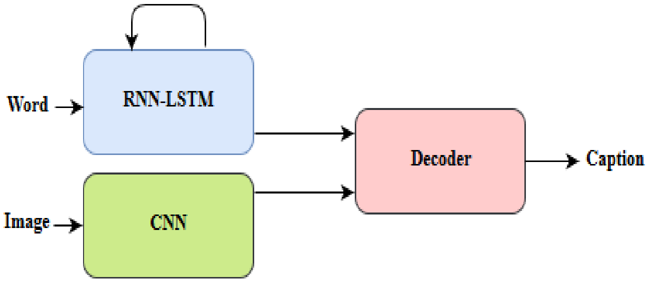


Fig. 5. The proposed model.

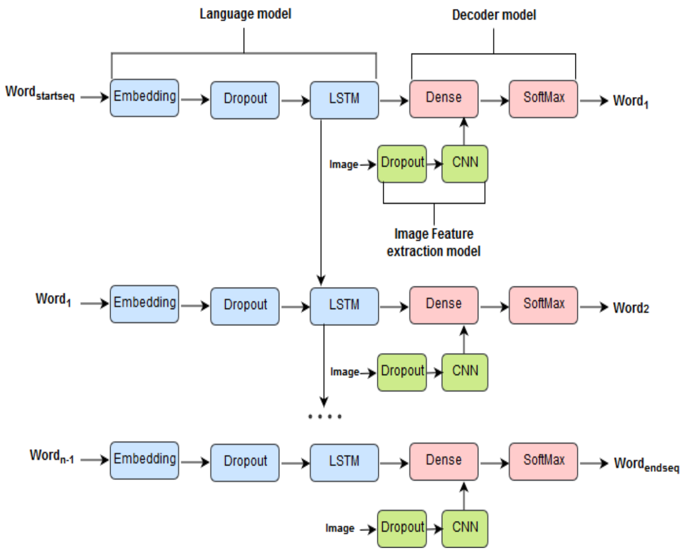


Fig. 6. Image captioning system for Arabic based on merge model.

The model is composed of three parts as shown in Fig. 5: 1) a language model based on RNN-LSTM [36] to encode linguistic sequences of varying length, 2) an image feature extractor model based on CNN [7] to extract image features in the form of a fixed-length vector, and 3) a decoder model that takes as input the outputted fixed vectors from the previous models and makes a final prediction.

A detailed illustration of the three parts of the proposed system is shown in Fig. 6. First, the language model inputs sequences with a pre-defined max length -the maximum words in the longest caption- which are fed then into an embedding layer that uses a mask to ignore padded values. Further, a 50% dropout is performed in a form of a regularization then the output is forwarded to the LSTM layer with 256 memory units. Independently, the second stage is the image feature extractor model that expects an input photo features to be a vector of 4,096 elements. A 50% dropout is also done before the image being processed by a CNN layer to produce a 256-element representation of the image. The final stage is the Decoder model that merges the 256-output of both models to an output Softmax layer that makes the final prediction over the entire output vocabulary for the next word in the caption.

IV. Dataset

The model has been trained and tested with images, from MS COCO dataset [37] and Flickr8K datasets [38]. The MS COCO images contain multiple objects in the scene collected by searching for pairs of 80 object categories. This dataset contains 2.5 million captions labelling over 330,000 images. To gather Arabic captions for the images, Crowd-Flower Crowdsourcing service [39] was used. Given 1166 images taken from the training set of the MS COCO dataset, a total of 5358 captions were collected. The images have on average 4.6 captions; the maximum number was 6 and the minimum was 4. Flickr8K dataset contains 8000 images and each comes with 5 English sentences. The images were selected with different locations and scenes from 6 Flickr groups. The first 2261 images from the training set were selected. A professional English-Arabic translator translated the captions of 150 images from Flicker, a total of 750 Arabic captions. The rest of the images (2111) were translated to Arabic using Google translator and then checked by Arabic native speakers. The total of images from both datasets (COCO and Flicker) is 3427, with a vocabulary size of 9854 and the longest caption consisting of 27 words. Since the dataset consists of some images from MS COCO and some from Flicker training sets, all images were divided for the experiments to 2400 for training, 411 for the development, and 616 for testing with a percentage of 70:12:18 respectively (see Table I).

V. Crowdsourcing Procedure

All captions used to build the dataset were human generated using Crowd-Flower Crowdsourcing [39]. A job was posted that asked the contributors to describe an image. In the job page, a user interface was provided with instructions in Arabic and one example. Each task includes only 5 images in each page to prevent contributor's exhaustion. Some of the instructions were translated directly from English instructions that were used in the MS COCO captions [37] and instructions specific to Arabic language were added. The job has the following instructions:

1) Please adhere to the standard Arabic language.

2) Write a useful sentence that ends with a period (.). Do not just type multiple words or phrases.

3) The sentence must contain at least 20 Arabic letters.

4) Use a polite style of speech and correct punctuation marks.

5) Please comment on the image by giving only factual data:

a) Do not write about things that may happen in the future.

b) Do not write about sounds, such as, the child heard the sound of the horn.

c) Do not speculate or imagine. Do not write about something that makes you feel uncertain.

d) Do not write about your feelings regarding the scene in the picture.

e) Do not use excessive poetic style.

1) Do not use demonstrative pronouns such as 'this' or the adverb of place 'here'.

2) Please do not write the names of the persons, places or nationalities; e.g. Washington City, American Flag.

3) Please describe all important parts of the scene; do not describe unimportant details.

TABLE I. Train/Dev/Test Split

Train	Train	2400
	Dev	411
Evaluate	Test	616
Total		3427

The job only appears for Arabic contributors to be sure that non-Arabic workers do not participate in this job. Six captions per image were collected. To guarantee the quality of the captions and that they are well written in Arabic and not using an Arabic dialect, a data-cleaning task was assigned to a professional Arabic language specialist. For some images, he selected the best 4 captions, for others he kept all 6 captions with small modifications.

VI. Experimental Evaluation

In this section the relative importance of different components of the proposed model is assessed, the implementation environment is defined, and finally the obtained results are presented and analyzed.

For the image encoder, a fully convolutional network based on Visual Geometry Group (VGG) OxfordNet 16-layer CNN [40] is adopted. Prior to training, all images were vectorized using the activation values which is trained to perform object recognition on a 4096-element vector and returns a 256 vector. For the language model, a single hidden RNN-LSTM layer with 256 memory units is defined. This layer is supported in the Keras [41] API library. The network uses a dropout of 50% on both models.

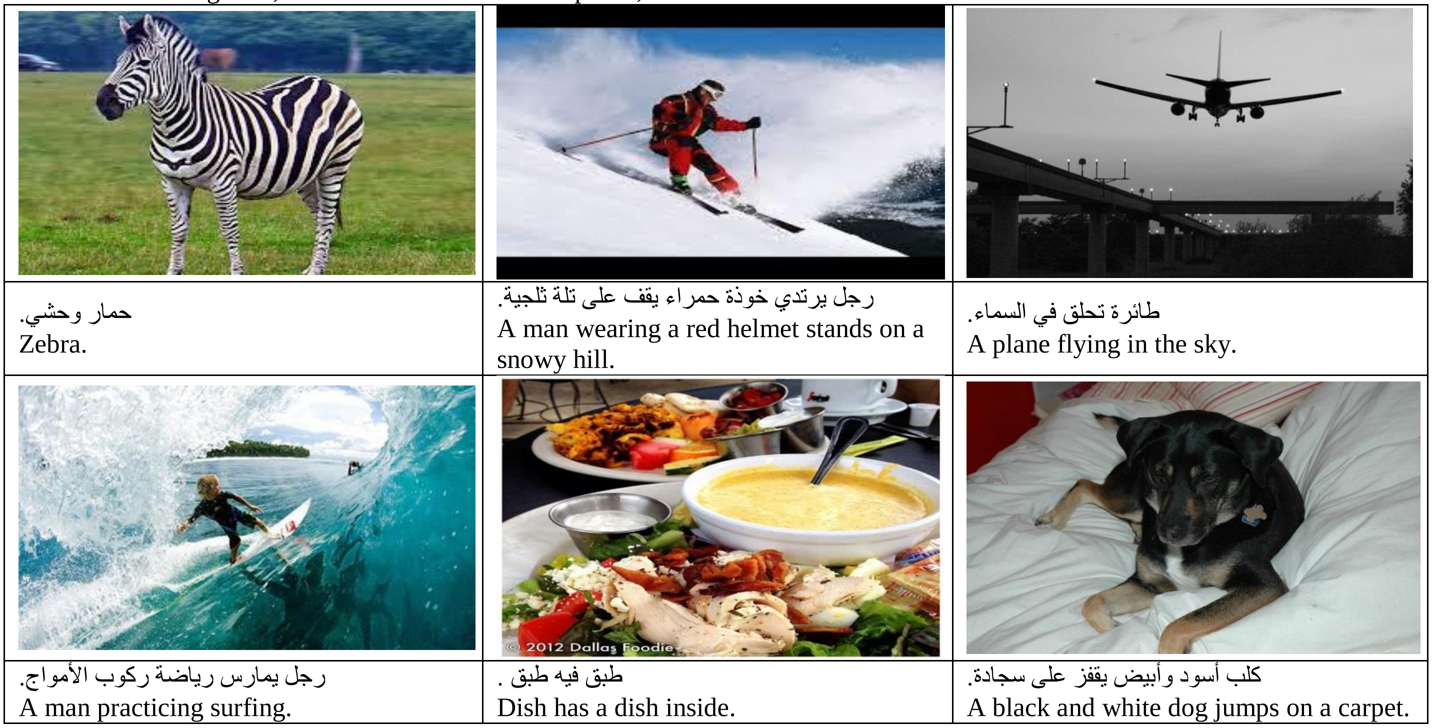
The complete model was implemented in python using latest version 2.1.6 of Keras [41]. Eexperiments were conducted on the commercial cloud server FloydHub [42]. FloydHub servr uses Nvidia Tesla K80 GPUs (12GB vRAM) and 61GB RAM and supports Keras [41] API.

In the experiment, the maximum length of a description within the data set is 27 words. This value is essential because it defines the input length to RNN-LSTM model. Given the amount of training data, the model was fit for 10 epochs, and the model stabilized after the 6th epoch. At the end of the 6th model, the loss computed was 4.278 on the training dataset and a loss of 4.859 was on the development dataset. The model generates correct descriptions of images (see Fig. 7), the syntax and the semantic of the sentences is accurate. For the middle image in the second row, the obtained caption “Dish has a dish inside”, even though correct, it fails to describe the important part of the image which is the food.

Following previous works, the model was evaluated on the BLEU-{1,2,3,4} [20], which evaluates a candidate sentence by measuring the fraction of n-grams that appear in a set of references. The model scored a value of 46 which is considered excellent in the BLUE scale. All BLEU scores obtained by the proposed model are given in Table II. Moreover, Table II gives a comparison of the proposed model with the BLEU scores of the Arabic captions obtained by translating the English captions derived from the NIC model [2] using Google Translate [28]. This translated model is evaluated on the Flickr8K dataset. As seen from Table II, the Arabic caption based merge model is comparable on the BLEU-{1} score. Also comparing the proposed model with [28], the proposed model results a 10% higher BLEU-{1} score. The obtained results are promising and can be improved with the availability of more data.

TABLE II. BLEU-{1,2,3,4} METRICS FOR THE ARABIC MODEL & NIC [2]/GOOGLE TRANSLATION

Test Dataset	**Model**	**BLEU-**			
		{1}	**{2}**	**{3}**	**{4}**
Flickr616	**Arabic caption based on merge model**	46	26	19	8
Flickr8K	**NIC [2]/Google translate**	52	46	34	18



حمار وحشي.
Zebra.

رجل يرتدي خوذة حمراء يقف على تلة ثلجية.
A man wearing a red helmet stands on a snowy hill.

طائرة تحلق في السماء.
A plane flying in the sky.

رجل يمارس رياضة ركوب الأمواج.
A man practicing surfing.

طبق فيه طبق .
Dish has a dish inside.

كلب أسود وأبيض يقفز على سجادة.
A black and white dog jumps on a carpet.

Fig. 7. Examples of image captions generated using the proposed model.

VII. Conclusion

A novel corpus of image captions in Arabic is built by collecting 5358 captions for 1176 images using a Crowd-Flower Crowdsourcing service, and 750 captions for 150 images were obtained from a human translator. The rest of image captions were translated from English to Arabic using Google translator. The RNN model trained by Arabic captions works well for image caption generation even with the small dataset that has been used for training and validating the model. Till now, no other RNN models were proposed for image caption generation for Arabic language except the paper of Jindal [28] that used a different methodology based on Deep Belief Network. The performance of the proposed model on the test set gave a promising result of 46.2 for the BLEU-1 score, which is 10% higher than the Jindal result.

The proposed model can give better performance with larger dataset. Therefore, for future research the image dataset with Arabic captions will be expanded and made publicly available. Further experiments will be conducted with the expanded corpus.

Acknowledgment

The authors would like to express their deep gratitude and sincere appreciation to their families that collaborated and were one hand to build the Arabic image captioning pilot corpus. Especial thanks to Mrs. Hend Al-yahya, -an English linguistic researcher- for her ongoing support and skills in auditing part of this dataset.

References

[1] J.-Y. Pan, H.-J. Yang, P. Duygulu, and C. Faloutsos, “Automatic image captioning,” in IEEE International Conference on Multimedia and Expo (ICME), 2004, vol. 3, no. 22, pp. 0–3.

[2] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, “Show and tell: A neural image caption generator,” in IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2015, vol. 07–12–June, pp. 3156–3164.

[3] J. Johnson, A. Karpathy, and L. Fei-Fei, “DenseCap: Fully Convolutional Localization Networks for Dense Captioning,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[4] J. Mao, X. Wei, Y. Yang, J. Wang, Z. Huang, and A. L. Yuille, “Learning like a child: Fast novel visual concept learning from sentence descriptions of images,” in IEEE International Conference on Computer Vision (ICCV), 2015, vol. 2015 Inter, pp. 2533–2541.

[5] J. Donahue et al., “Long-term recurrent convolutional networks for visual recognition and description,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 4, pp. 2625–2634, 2017.

[6] A. Karpathy and F. F. Li, “Deep visual-semantic alignments for generating image descriptions,” IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., pp. 3128–3137, 2015.

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” in 25th International Conference on Neural Information Processing Systems, 2012, pp. 1097–1105.

[8] O. Russakovsky et al., “Imagenet large scale visual recognition challenge,” arXiv Prepr. arXiv1409.0575, 2014.

[9] M. Tanti, A. Gatt, and K. P. Camilleri, “What is the Role of Recurrent Neural Networks (RNNs) in an Image Caption Generator?,” Aug. 2017.

[10] M. Tanti, A. Gatt, and K. P. Camilleri, “Where to put the Image in an Image Caption Generator,” Mar. 2017.

[11] C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier, “Collecting Image Annotations Using Amazon’s Mechanical Turk,” in NAACL HLT Workshop on Creating Speech and Language Data with Amazon’s Mechanical Turk, 2010, no. June, pp. 139–147.

[12] T.-Y. Lin et al., “Microsoft COCO: Common Objects in Context,” 2014, pp. 740–755.

[13] J. Devlin et al., “Language Models for Image Captioning: The Quirks and What Works,” arXiv Prepr. arXiv1505.01809, 2015.

[14] S. Pratap, S. Gurjar, S. Gupta, and R. Srivastava, “Automatic image annotation model using LSTM approach,” Signal Image Process., vol. 8, no. 4, 2017.

[15] J. Devlin, S. Gupta, R. Girshick, M. Mitchell, and C. L. Zitnick, “Exploring Nearest Neighbor Approaches for Image Captioning,” arXiv:1505.04467, 2015.

[16] V. Ordonez, G. Kulkarni, and T. L. Berg, “Im2Text: Describing Images Using 1 Million Captioned Photographs,” NIPS, pp. 1143–1151, 2011.

[17] A. Farhadi et al., “Every Picture Tells a Story: Generating Sentences from Images,” in European Conference on Computer Vision (ECCV), 2010, pp. 15–29.

[18] D. Elliott and F. Keller, “Image Description using Visual Dependency Representations,” Emnlp, no. October, pp. 1292–1302, 2013.

[19] J. Krause, J. Johnson, R. Krishna, and L. Fei-Fei, “A Hierarchical Approach for Generating Descriptive Image Paragraphs,” in Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3337–3345.

[20] K. Papineni, S. Roukos, T. Ward, and W. Z. Ibm, “BLEU: a Method for Automatic Evaluation of Machine Translation,” in 40th Annual Meeting on Association for Computational Linguistics (ACL), 2002, no. July, pp. 311–318.

[21] V. Ramakrishna, L. Zitnick, and P. Devi, “Cider: Consensus-based image description evaluation,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.

[22] M. Denkowski and A. Lavie, “Meteor Universal: Language Specific Translation Evaluation for Any Target Language,” in 9th Workshop on Statistical Machine Translation, 2014, pp. 376–380.

[23] R. Krishna et al., “Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations,” Int. J. Comput. Vis., vol. 123, no. 1, pp. 32–73, May 2017.

[24] J. Mao, W. Xu, Y. Yang, J. Wang, and A. L. Yuille, “Explain Images with Multimodal Recurrent Neural Networks,” arXiv:1410.1090, Oct. 2014.

[25] L. A. Hendricks, S. Venugopalan, M. Rohrbach, R. Mooney, K. Saenko, and T. Darrell, “Deep Compositional Captioning: Describing Novel Object Categories without Paired Training Data,” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1–10.

[26] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille, “Deep captioning with multimodal recurrent neural networks (m-rnn),” in nternational Conference on Learning Representations (ICLR), 2015.

[27] C. Y. Lin, “Rouge: A package for automatic evaluation of summaries,” in Workshop on text summarization branches out (WAS 2004), 2004, no. 1, pp. 25–26.

[28] V. Jindal, “A Deep Learning Approach for Arabic Caption Generation Using Roots-Words,” AAAI, pp. 4941–4942, 2017.

[29] Y. Jia et al., “Caffe: Convolutional Architecture for Fast Feature Embedding,” arXiv:1408.5093, 2014.

[30] Q. Yaseen and I. Hmeidi, “Extracting the roots of Arabic words without removing affixes,” J. Inf. Sci., vol. 40, no. 3, pp. 376–385, Jun. 2014.

[31] J. Hajič, O. Smrz, P. Zemánek, J. Šnaidauf, and E. Beška, “Prague Arabic dependency treebank: Development in data and tools,” in International Conference on Arabic Language Resources and Tools (NEMLAR), 2004, pp. 110–117.

[32] T. Miyazaki and N. Shimizu, “Cross-Lingual Image Caption Generation,” in The Association for Computational Linguistics (ACL), 2016, pp. 1780–1790.

[33] W. Lan, X. Li, and J. Dong, “Fluency-Guided Cross-Lingual Image Captioning,” in ACM on Multimedia Conference (ACMM), 2017, pp. 1549–1557.

[34] H. Peng and N. Li, “Generating Chinese Captions for Flickr30K Images,” 2016.

[35] D. Elliott, S. Frank, K. Sima'an, and L. Specia, "Multi30K: Multilingual English-German Image Descriptions," in 5th Workshop on Vision and Language, hosted by the 54th Annual Meeting of the Association for Computational Linguistics, 2016, pp. 70–74.

[36] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," Neural Comput., vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[37] X. Chen et al., "Microsoft COCO Captions: Data Collection and Evaluation Server," arXiv:1405.0312, 2015.

[38] M. Hodosh, P. Young, and J. Hockenmaier, "Framing image description as a ranking task: Data, models and evaluation metrics," in 24th International Joint Conference on Artificial Intelligence (IJCAI), 2015.

[39] CrowdFlower, "Machine Learning, Training Data, and Artificial Intelligence Solutions: Figure Eight." [Online]. Available: https://www.figure-eight.com/.

[40] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv:1409.1556, 2014.

[41] F. Chollet, "Keras Documentation," Keras.Io, 2015. [Online]. Available: https://keras.io/. [Accessed: 28-Apr-2018].

[42] "FloydHub Documentation." [Online]. Available: https://docs.floydhub.com/. [Accessed: 28-Apr-2018].

Use of Technology and Financial Literacy on SMEs Practices and Performance in Developing Economies

Juma Buhimila Mabula
Department of Business Administration,
Harbin Institute of Technology
Harbin, China

Han Dong Ping
Department of Accounting
Harbin Institute of Technology
Harbin, China

***Abstract*—Micro, Small and Medium Enterprises (SMEs) practices in developing economies experience a unique set of challenges to attain their success. With a view of analyzing double impact of SME financial literacy and use of technology on practice of record keeping and risk management as echoed on firm performance, the partial least square structural equation modelling was used to configure the perceived impact of these variables. The results posit a significant relationship between the firm use of technology to its practice of record keeping and performance, a significant positive association of financial literacy and firm risk management practices. Nevertheless the study found insignificant association of financial literacy and firm book keeping practice; it offers unleashed dual practical role of financial literacy and use of technology for improving SMEs financial practices in developing economies.**

Keywords—Technology use; financial literacy; book keeping; risk management; developing economies

I. Introduction

As SMEs in developing countries being considered as important pillars of their economies they also serve as a mean for people’s life sustenance and survival [1]. The fact that SMEs are managed by single or few owner/managers and their smallness nature implicate high cost of having relevant information as a base for rational decision making. Therefore possessing a well-placed technology utilization and financial literacy niceties can greatly enhance the firm decision making process.

The inception of new digital information and technology has revolutionized the world of business practices [2]-[4]. Besides of the challenges it has posed, the potential ample benefits justify its adoption by individuals and firms. The use of technology has an anchoring utility especially when it comes to learning process. Adopting new technology opens up multiple avenues for acquiring other skills and knowledge. In this quest the yielded results transmuted into business practices encompassing e-commerce, e-business and e-marketing. These practices has demonstrated a positive impact to all business of any size globally [5].

The analysis of the adoption and applicability of Information Communication and technology can only be meaningful if studied by satisfying the conditions for a company success [6]. Financial literacy as one of the basic gasp that is inevitable for an individual or firm has implacable anecdotes with the use of technology. Financial literacy and use of technology at individual and firm level has a tendency of bolstering each other. In [7], [8] embedding use of technology as a construct in financial literacy studies verify its noteworthy impact on firm financial practices and performance. At firm level the concept of financial literacy has been shown to have an influence on some basic firm practices [9]-[11] such that the attainment of improved financial literacy by the firm managers will ultimately improve the performance of the firm.

Coupled with the general low level of technological advancement, developing economies has shown a deviant picture in ICT adoption, while some countries have been able to yield the ICT potentials, other countries has lagged behind [12]. The choice of technologies is deemed to be the most important decision for developing economies, therefore understanding the intricacies for making such a choice is imperative [13]. On the other hand the general financial literacy in developing economies is also considered to be relatively low [14]. Therefore studying the double array of financial literacy and technology utilization impact on SMEs practices and performance forms an interesting phenomenon.

The use of technological at firm level reckoned to have positive impact on firm practices, and a financial literacy well placed firm managers provides an array for SMEs better financial management and improved performance. The analysis of concurrence and collectiveness of SMEs technology use with financial literacy impact on firm practice and performance is the focus of this study. Specifically this study improvise a model that works on assessing the impact of financial literacy and use of technology into firm record keeping practices and risk management replicating their consequential results on firm performance in developing economies.

II. Hypothesis Development

This study is based on an “evolutionary theory of economic change” by [15]. The theory contends that a firm searching for innovative (imitative) solutions boost their profits, in such a way prosperous firms grows at the expense of the outperformed firms. This process dynamically has an interaction of many variables in business competitive environment. Therefore the firms may not be able to grasp the finest technological bases, ultimately fail to heighten their performance perfectly.

The theory clearly shows the need for the firm to be innovative and the interactiveness of variables to boost the performance of the firm. The theory is applied to amplify the importance for right technology adoption as basis for improving firm performances. Technology enhances innovations and widen the sphere of knowledge and skills acquisition, the applicability of it is reflected in the firm interactive practices for enhancing its performance. With this base the general model is drawn and presented in Fig. 1.

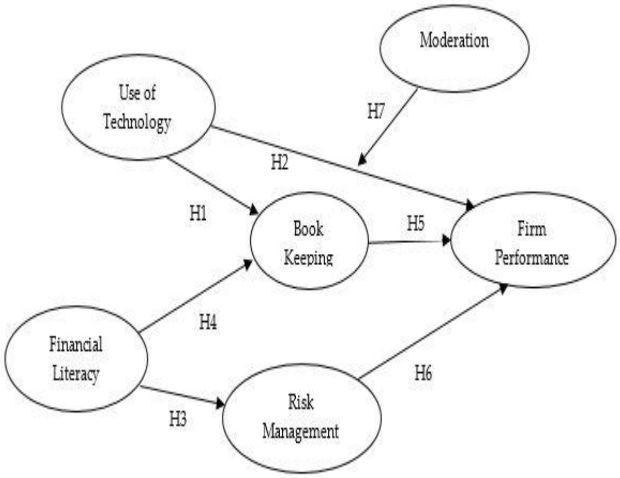


Fig. 1. The conceptual model of the study.

The application of information technology has had an obvious positive results into firm practices. [16] assert that firm with strong IT resources are able to assimilate the technological business planning cases more effectively, apprehend and make a dependable cost effective use that support the firm more fast than competitors, have an effective inter-business unity communications and able to foresee the future business needs before their competitors. The investment in information technology is a vital organization resource that can be used to advance communications, enhance product design worth, melt down the product development costs and lessen the design cycle [17]. Previous studies gives us a clue on the positive impact of information technology application on firm performance [16], [18], [19]. With this base the study therefore propose that:

Hypothesis $H_{1:}$ The use of technology has a positive impact on firm practice of book keeping in developing economies.

Hypothesis H_2: The use of technology positively influence firm performance in developing economies.

Financial literacy at individual and firm level is a key to sound financial management. At a firm level, considering the nature of smallness of organizations, managers are easily capable of disseminating their financial knowledge and skills into firm practices. Financially literate SME manager is capable of using financial resources optimally, knows and understand the sources of reliable financing, capable of implementing sound business practices of budgeting, planning and control, put in place a sound system of acquiring, processing, storing and dissemination of financial information. The literature gives us ample overt evidences of financial literacy to financial practices [20]-[23] at individual level and [24]-[26] at firm level. Therefore researchers propose that:

Hypothesis H_3: Financial literacy positively influence the practice of risk management practices by SMEs in developing economies.

Hypothesis H_4: Financial literacy has a positive impact on firm book keeping practice in developing economies.

The SMEs functional practices are be a base of it success or failure. Of particular interest record management has been identified to be one of the principle success factor for entrepreneurs [27]. Therefore there is a need to have a comprehensive system prioritizing its authenticity. And enterprise risk management currently is viewed as one of the key conditions for successful firms because it enable them to put a glance at all risks facing a company through systematic plan [28]. Therefore SMEs are well placed to attain better performance if they have sound systems of record keeping and risk management. With these arguments it is proposed that:

Hypothesis H_5: The practice of book keeping positively influences the firm performance in developing economies.

Hypothesis H_6: The risk management practices in developing economies has a significant impact on firm performance in developing economies.

Author in [6] gives us clues that the information and technology is very crucial to SMEs because it enable them to implement their business competitive advantage. SMEs benefits by enhancing their visibility, easy access to information, overcome traditional trade areas and facilitate financial transactions. SMEs record keeping, processing and disseminations overtly enhanced by use of technology in terms of hard and software. The use of computers software enhances the preparation of instant periodical financial statements. Therefore SME use of technology in record management considered to increase the impact of firm use of technology on its performance. Therefore it is argued that:

Hypothesis H_7: Firm book keeping practices moderates the use of technology-firm performance relationship.

III. LITERATURES

A. Financial Literacy

Many individual has attempted to define financial literacy at personal level. The main focus has been simply summed up by Jump$tart definitions who simply define financial literacy as the ability to use knowledge and skills to manage one's financial resources effectively. At firm level financial literacy stresses the ability of the owner/manager to be able to propagate their financial skills and knowledge into firm operations. The concept of financial literacy for SMEs spring from the nature of smallness and few owners of SMEs creating an avenue of easily transferring their expertise into firm endeavor. The Bank association of South Africa has attempted to define SME financial literate managers as entailing the following qualifications: having adequate level of entrepreneurial competencies, personal skills, and business management skills, having appropriate level of understanding of functional financial management systems, has appropriate level of understanding of SMEs life cycle funding and other financial requirements, understand legal, regulatory and tax issues as they relate to financial matters and understand the range of legal resources it can opt to when necessary, in case of bankruptcy or other financial distress situations [29].

B. Use of Technology

In this context [30] gives us a definition by stating that 'Information technology is the technology that is used to hoard, manipulate, distribute and create information'. [31] provides four dimensions of the use of technology readiness which are optimism, innovativeness, discomfort and insecurity. The optimism dimension presents a positive view of technology adoption and perception of the benefits to be accrued from its use for enhancing work efficiency and improving performance. Innovativeness refers to the extent to which individuals prefer to experiment with the new technology to come out with the best solution towards solving their problems. The dimension of discomfort displays a sense of lack of technological mastery and lack of courage to use the latest technology. And the insecurity dimension presents a state of complete mistrust of technology-based transactions and complete doubt of its capabilities. The first two dimensions are considered to be contributors to the process of adopting new technology while the last two represent inhibitors of the technological adoption. Author in [32] shows that optimist and innovators are prone to adopting to new technology easily. In terms of financial supply technology advances is the ability to effortlessly create a free online connection structure, the secure online money transfer, correct credit scores to be utilized by a bunch of financiers, and the free media tools to market their products therefore engage large group of geographically dispersed people [33].

C. SMEs Definition

Defining SMEs has been a contextual endeavor due to lack of general widely accepted definition. The study define SMEs by Tanzania context the setting of where the study was undertaken. The Tanzania SME policy defines SME to mean micro, small and medium enterprises. These organizations comprises of non-farm activities mainly manufacturing, mining, commerce and services. The micro organizations are those employing up to 4 people with capital threshold of Tanzanian shillings 5.0million. The small enterprises are considered to employ ranging from 5 to 49 human resources with capital investment ranging from Tshs. 5 million to Tshs. 200 million. The medium enterprise they are the one which employ between 50 to 99 people within a range of capital of Tsh.200 to 800 million [34].

D. Use of Technology and Financial Literacy

The adoption of technology by individual and firm now has become an inevitable matter boosted by expanded coverage and presence of wireless cellphones, computers, internet and other gadgets. In developing economies where there is prevalence of low level of technology, the use of technology has become a pillar for individual and firm success. The connection of technology and financial products offers crucial tie on enhancing financial literacy. For instance the electronic banking which has proliferated recently has increased financial products consumers' accessibility. These technological products may include direct deposit, computer banking, store value cards and debit cards [35]. A study of financial literacy by [8] on financial literacy of micro entrepreneurs in South Africa found that there is a limited use of technology by most firms. In this study they discovered that most of the firms had no email address and inaccessibility to internet at their work station, none of their respondents had a webpage. Financially excluded individuals and firms suffer from lack of financial literacy and basic education resulting into limited access into financial services. Therefore concentrate much on the use of informal financial system of which they are more familiar with. The barrier can be bypassed by employing information technology, which gives a feasible option to reduce transaction costs at large extent and form user friendly platforms convenient to the users [36].

E. Use of Technology SME Practices

The general extent of utilization of technology by SMEs is at low level [37]. Nonetheless literature offer ample evidence that utilization of technology is an important determinant for SME success [38]-[40]. Considering the scale of many small businesses such that they are more prone to cash flow problems, and therefore have less resources to devote to the sophisticated management of financial instruments and, since they don't partake into economies of scale and market power enjoyed by large organizations [41], [42]. The use of information technology by SMEs can benefit them by developing competences for managing, information intensive resources, reduce the transaction costs and develop capability to gather information locally and internationally resulting into rapid flow of information [30]. For instance crowd funding via internet based forums can improve SMEs access to finance and potentially remedy certain market failure, such as sufficiency of funding for start-ups it is a flexible, cost-effective and quickly raise finances, and it could be a marketing tool to entrepreneurs [43].

F. SMEs use of Technology and Record Keeping

Firm records are business information that contains the transactions between parties in the form of electronic and paper documents kept for present and future references [44]. The electronic business records, refers to the records in a form of digital records processed, manipulated and/or transmitted using a computer system. The practice of keeping information electronically by SMEs poses a new set of challenges for the SMEs to retain the authenticity of such information because they can easily be altered.

G. Financial Literacy and SMEs Risk Management

Authors in [45] stresses the importance of insurance as the main driving risk tool for SMEs because the other sophisticated means of risk management might prove to be cost ineffective. Financial literacy enables individuals and businesses to make better financial decisions and therefore undertake firm risk management more effectively [8].

Author in [46] measuring financial literacy discovered one's protection from unexpected happenings using either insurance or risk management techniques is one of the important constructs in financial literacy studies. Risk management is an interdisciplinary theme which may call integration of knowledge of fields of economic, technological, social and political knowledge. SME managers function in a macro, and micro market environment that is affected by numerous internal and external influences which continuously prone to change. The situation confronts entrepreneurs to be vigilant therefore able to identify opportunities and threats. Having more knowledge of financial literacy specifically in strategic risk management enables SMEs owner/managers to objectively evaluate their actions. The challenge with risk management is that every risk is linked to a different discipline which may not be necessarily in conversancy of firm manager proficiency. The concept of managing risks is an essential part of borrowing and investing. Therefore risk management is considered as part and parcel of individual and firm financial literacy.

IV. Methodology

Data was collected through a survey of SMEs in region of Morogoro and Dar es salaam region Tanzania. The survey was directed to SMEs owners/managers' at the top level who could provide the required information. The data collection was done by researchers' personal visit at the firm location and the researcher supervised the process of filling the questionnaire for prompt response. The introduction was preceded by the introduction letter assuring the participants that the information needed was solely academical, confidential treatment of the results and researchers' readiness to avail the results to them and the public at large. The choice of firm to be included in this study was based on convenient sampling due to limiting factors including firm proximity, time, financial and respondents readiness. The survey was done by a team of 4 researchers with an average of 8 SMEs reached per day per each researcher. The researcher visited 520 firms and the complete usable results obtained from 311 firms. A response rate of about 60 percent. This study is part of the comprehensive questionnaire designed capture main aspect of financial literacy at personal and firm level. The survey questionnaire was used in pilot study for instruments refinements and then evaluated by experts. Cronbach's alpha test was used to verify its quality and consistency. Basing on the PLS sampling rule that the minimum sample size has to be ten times of the portion the model that require for multiple regression. This model has 5 independent variables tolerating the use a minimum sample of 50, therefore 311 is more than six times enough.

A. Measured Variables

The indicators making up the constructs for financial literacy were gathered using instruments validated by [14] where by financial literacy (FL), was categorized into general financial knowledge, financial knowledge, financial attitudes, and financial behavior. Financial knowledge was captured using multiple choice questions and the remaining questions of financial attitudes and behavior were stylized into a likert scale, the choice of questions was made in such a way they are relevant to firm level business operation. The use of technology (UT), book keeping (BK) and risk management (RMGT) were are all adopted from [8], thereafter made into likert scale questions of 5 points. Some items were added from the original format questions for questionnaire comprehensiveness. The firm performance (FP) indicators was drawn from the enterprise survey questionnaire, formed into a five point likert scale, [47] uses the same approach.

B. Demographic Information

The study captured demographic information at individual and firm level. At individual level the survey collected information about gender and education level and at firm level information collected was about business type, location, firm age and number of employees. Table I shows the detailed results. The demographic information were used as control variables. These control variables were linked to the independent variables and assessed whether there was a significant increase in R square, the results with all control was insignificant, therefore the influence of these control variables was excluded in further analysis.

TABLE I. Summary of Respondents Demographic Information

		Number	Percentage
Gender	Male	164	52.9
	Female	146	47.1
Education Level	Primary	11	3.5
	Secondary	87	28.1
	College	134	43.2
	University	78	25.2
Business type	Wholesale and Retail	101	32.2
	Agricultural	56	18
	Construction	31	10.0
	Food and Accommodation	31	10.0
	Manufacturing	30	9.7
	Others	61	19.1
Location	Rural	127	41
	Urban	183	59
Firm Age	Below 1 Year	39	12.6
	1-5 Years	132	42.6
	6-10 Years	100	32.3
	More than 10 Years	39	12.6
Number of employees	1-5 Micro	144	46.5
	6-49 Small	147	47.4
	50-99 Medium	19	6.1

C. Examination of Model Reliability and Validity

By examining the parameters of the PLS structural model the hypotheses are assessed. The standardized path coefficients shows how strong the relationship between the independent and dependent variable. In a partial least square- structural equation modelling composite reliability is an alternative preference to Cronbach's alpha to test the convergent validity in a reflective model, this is because Cronbach's alpha may under/overestimate the scale of reliability [48].

The first step was to examine the outer model acceptability in the research algorithm. The outer loading acceptable range be ≥ 0.5, [49]-[52]. According to the result presented in a Table II, all of the construct meet this criterion with exception of UT3 and therefore considered further analysis.

Average variance extracted (AVE) can be used to assess both convergent and divergent validity. In actual fact AVE mirrors the average communality for every latent factor in a reflective model. The AVE can be calculated by having the sum squared component loading for each construct divided by sum of component loading plus the sum of the error variance. The adequacy of a model is reflected by AVE being greater than 0.5 [49], [53]. The meaning of this condition is that factors in the model at least explain half of the variance of the corresponding indicators. Once more Table II presents acceptable results for the model.

TABLE II. RELIABILITY AND CONVERGENT VALIDITY

	Mean		Standard Deviation	Loading	t-static
	Book keeping (composite reliability=0.765; AVE=0.5)				
BK1-BK	0.649		0.225	0.706	3.257*
BK2-BK	0.715		0.225	0.771	3.639**
BK3-BK	0.612		0.225	0.635	2.812*
BK4-BK	0.543		0.225	0.561	3.142**
	Use of technology (composite reliability=0.762; AVE=0.5)				
UT1-UT	0.794		0.149	0.829	6.222**
UT2-UT	0.788		0.165	0.826	5.926**
UT3-UT	0.443		0.271	0.469	1.783
	Financial Literacy (composite reliability=0.794; AVE=0.5)				
GFL-FL	0.476		0.201	0.537	5.599**
FA-FL	0.772		0.151	0.795	4.627**
FK-FL	0.772		0.156	0.814	4.860**
FB-FL	0.620		0.201	0.637	3.061*
	Firm Performance (composite reliability=0.803; AVE=0.6)				
PF1-PF	0.916		0.135	0.947	10.770**
PF2-PF	0.647		0.227	0.703	3.472*
PF3-PF	0.602		0.179	0.601	3.205**
	Risk Management(composite reliability=0.817; AVE=0.6)				
RMGT1-RMGT	0.832		0.169	0.891	4.860**
RMGT2-RMGT	0.642		0.256	0.692	2.510*
RMGT3-RMGT	0.674		0.258	0.728	2.925*
	*p≤0.05 and ** p≤0.001				

1) *Discriminant validity*

The Fornell-Larcker criterion was used to establish the discriminant validity. The criterion is based on calculation of AVE. which states that "for any latent variable, the square root of the AVE should be higher than its correlation with any other latent variable" [48]. This imply that for any latent variable the shared variance with its cluster of indicators is higher than the variance it shares with any other latent variable. In result Table III verify that there is discriminant validity in the model since all the numbers in diagonal cells (square root of AVE) in absolute terms are greater than the corresponding correlation of the variables.

TABLE III. FORNEL-LARCKER CRITERION AND HETEROTRAIT-MONOTRAIT RATIO (HTMT) IN PARENTHESIS

	BK	FL	FP	RMGT	UT
BK	0.673				
FL	-0.164 (0.3)	0.705			
FP	0.35 (0.6)	0.014 (0.2)	0.764		
RMGT	-0.020 (0.3)	-0.269 (0.4)	0.112(0.3)	0.775	
UT	-0.243(0.5)	-0.110 (0.3)	-0.269 (0.5)	0.007 (0.1)	0.728

2) *Cross loading and heterotrait-monotrait ratio*

An alternative to assessing discriminant validity beside of AVE is the examination of the cross-loading. Accordingly a good model has to have indicators which load well on their intended factors. The cross loading with other factors they are not intended to measure should be given attention. The rule of thumb is that cross loading should be under 0.3 and some use 0.4 [48]. Evaluating the variables in Table IV, all indicators have less or equal to 0.3 cross loading weight signifying that the model has discriminant validity.

Superior to the use of cross-loading examination and use of Fornell-Larcker criterion the heterotrait monotrait ratio (HTMT) criterion was established. The HTMT is the geometric mean of the heterotrait-heteromethod correlation divided by the average of the monotrait-heteromethod correlations [54]. The condition for this criterion for a well-fitting model, heterotrait correlations should be smaller than the monotrait correlations, implying that the HTMT ratio should be below 1. A cut of 0.9 is established in literature [55], [56]. The results from model in Table IV shows all ratios displayed are fine below 0.9, again further confirming that the model has discriminant validity.

TABLE IV. VARIANCE INFLATION FACTOR (VIF) AND CROSS LOADING

	VIF	Cross Loading				
		BK	***FL***	***FP***	***RMGT***	***UT***
BK1	2.196	0.7	-0.1	0.2	0.0	-0.2
BK2	2.234	0.8	0.0	0.3	-0.2	-0.3
BK3	1.264	0.6	-0.3	0.3	0.1	-0.0
BK4	1.264	0.6	-0.1	0.2	-0.0	0.0
FA	1.629	-0.2	0.8	-0.0	-0.1	-0.1
FB	1.617	-0.1	0.6	-0.1	-0.1	-0.1
FK	1.010	-0.0	0.8	0.2	-0.3	-0.1
GFL	1.061	-0.2	0.5	-0.0	-0.2	-0.0
PF1	2.089	0.3	0.0	0.9	0.1	-0.3
PF2	1.669	0.2	-0.0	0.7	0.2	-0.2
PF3	1.572	0.3	-0.2	0.6	-0.0	-0.1
RMGT1	2.649	-0.1	-0.3	0.1	0.9	-0.0
RMGT2	1.870	-0.0	-0.2	-0.0	0.7	0.0
RMGT3	1.846	0.0	0.1	0.2	0.7	-0.0
UT1	2.286	-0.3	-0.1	0.1	0.0	0.8
UT2	1.331	-0.2	-0.1	-0.2	0.0	0.8
UT3	1.874	-0.2	-0.1	-0.2	0.0	0.5

Source: Authors' analysis

3) *Multicollineariry*

The possibility of multicollinearity existent in the reflective model was also tested. Multicollinearilty occurs when two or more variables are highly intercorrelated. This may deter a researcher from assessing the relative importance of the independent variable. To examine the presence of multicollinearity a variance inflation factor (VIF) is used. The common rule of thumb provides a cut off of 4 or higher VIF for multicollinearity to exist [48]. In this model the results in Table IV offers all indicators VIF below 4 signifying that multicollinearity hasn't affected this scrutiny.

4) *Common method bias*

In a survey model of self-reported data there is a possibility of common method bias to exist. [57] posit that common method bias can be a result of a combination of many factors that may entail consistency and social interest. Specifically common method bias is a caused by the measurement approach used in a structural equation modelling study and the implied social desirability associated with the questionnaire responses eliciting the indicators to share a certain amount of conjoint variation [58]. To assess whether the data are affected with common method bias two approaches were employed, first is the psychological separation [59] done during data collection the questionnaire was structured in such a way the dependent and independent variables question was not asked on chronological order. Secondly was to further asses the data collected in the model using the approach by [58]. The approach emphasize on assessing the variance inflation factor (VIF) acceptability say if they are ≥ 3.3 provided that the data passed the convergent and discriminant validity then confirm the absence of common method bias problem in the model. Table IV verify VIF results in the model are below 3.3, therefore it validate that the data are not affected by common method bias.

D. *Evaluation of Study Structural Model*

With satisfactory measurement model and no worry for multicollinearity, common method bias and the non-response biases the analysis proceeded to test the study structural model. The test comprises of estimates of path coefficients signifying the strength of the relationship between dependent and independent variable and the evaluation of the significance of such path coefficients.

TABLE V. PATH COEFFICIENT

	Sample	Mean	Standard deviation	T statistics
BK->FP	0.372	0.371	0.116	3.134**
FL->BK	-0.193	-0.208	0.153	1.227
FL->RMGT	0.269	0.287	0.125	2.227*
ModE->FP	0.195	0.174	0.093	2.075*
RMGT->FP	0.113	0.117	0.144	0.790
UT->BK	0.264	0.285	0.121	2.273*
UT->FP	0.197	0.204	0.095	2.065*
*p≤0.05 and ** p≤0.001				

Source: Authors

Examination of the specific hypotheses exemplified in the study model was assessed by two-tailed t-statistics for the standardized path coefficients running basic bootstrap with no sign changes option at 1000 samples. The results are shown in Table V and Fig. 2.

H_1 Predicted the presence of positive direct relationship between the firm use of technology and the practice of firm book keeping in developing economies. The PLS structural path coefficient was significant at ($p< 0.05$) in hypothesized direction. This proposes that firm which has the high rate of use of technology are the one with improved of record keeping practices. Therefore H_1 is supported.

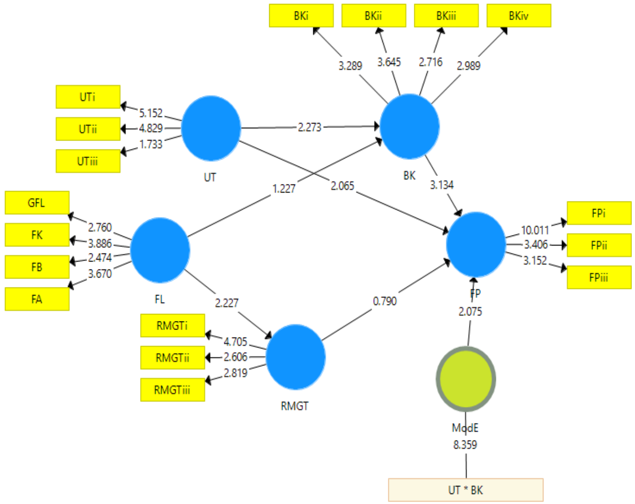


Fig. 2. Results after bootstrap.

H_2 Also predicted a positive relationship of the firm use of technology with its performance. According to the results the PLS structural path coefficient on this relationship was significant at ($p<0.05$) on the positive hypothesized direction. The results suggests that the more the firm employ the use of technology the more it will improve its performance over a steady period of time. Therefore H_2 is supported by the study algorithm.

H_3 States that financial literacy has a positive relationship with firm risk management practices. From the results the PLS structural path coefficient confirm that this path is significant at ($p< 0.05$) again the result suggest that, the higher the financial literate managers the more likely to engage into risk management of the firm. Therefore financial literacy of those who manage the SMEs in developing economies has an important impact on firm risk management. Again H_3 is supported by the algorithm.

H_4 suggested that there is positive relationship between financial literacy and firm practice of record keeping. The PLS structural path coefficient in this it is not significant. Suggesting that there is no direct relationship between financial literacy and the practice of book keeping. Though it is hard to draw a conclusion on insignificant results speculation can be on factors like the managers' level of education and factors for financial information flow constructs may mediate or moderate the relationship between financial literacy and firm book keeping practice.

H_5 predicted that there is positive relationship between firm practice of book keeping and firm performance. The partial least square structural path coefficient is significant at ($p< 0.01$) in a hypothesized direction. The results indicate that, the more the firm put more emphasis on practice of record keeping the more will enhance its performance over time. This higher significance emphasize on how important record keeping can be for firm sustainability. Keeping records is the base of all analysis aiding financing, investment, marketing and pricing decisions. Therefore H_5 is supported by the analysis.

H_6 predicts that there is a direct positive relationship between firm risk management practices and firm performance. The PLS structural path coefficient in this equation is not significant though it display a positive relationship. With this results speculation can be that there are other factors that mediate/moderate the risk management-firm performance relationship. The risk management factor determinants also would be not likely to be captured in a short while because the happening of uncertain event is not certain. Therefore H_6 is not supported.

H_7 Predicted the moderating role of firm book keeping practice on use of technology-firm performance relationship. The PLS structural path coefficient is significant on this equation at ($p< 0.05$). The result suggests that firm practice of record keeping enhance the relationship of firm use of technology and its performance in developing economies. In other words emphasizing on plausible firm record keeping practice implicate more use of technology which eventually improve firm performance. Therefore H_7 is supported by the analysis.

V. Discussion of Results

This study links variables at individual and firm level, examining how financial literacy and use of technology affect the business practices of book keeping and risk management translating the ultimate effect on firm performance in developing economies. The results presents significant results on the impact of use of technology and firm practices and performance. This result is consistent [38]-[40], [60]. Adoption and actual use of more sophisticated information technology system technology of more sophisticated in developing countries would have a cunningly better way of SMEs practices and performance.

Enhancing financial literacy of SMEs managers in developing economies still proves its vitality. Even though the model offer a slight insignificant results of ($p< 0.2$) suggesting an indirect relationship of financial literacy on the firm practice of book keeping, the usefulness of SMEs managers understanding of financial information collection, processing, storage and dissemination is of utmost importance. Firm managers well placed on the knowledge and skills of process financial data electronically will have timely information as a base for their present and future decision.

The model also offers significant results of the relationship of financial literacy and firm risk management practices. As it is demonstrated in [8], [45], [46], risk management is an important firm determinant of financial literacy. Therefore firm managers high level of financial literacy will embark on more effective business insurance and internal risk management practices and strategies. The lack of insurance cover by micro entrepreneurs demonstrated in [8], verify the failure of these SME managers to recognize the its importance.

Well placed firm practice of data management has been shown to have a high significance on the performance of the firm; the results are elucidated in H_5. Proper book keeping is one of the firm sound accounting practices. Consistent with the result [61] offers more evidence that proper book keeping practices as the one of the key factor for SME success or failure in United Kingdom and Nigeria comparative study. And also [62] contended that majority of SMEs lacked accounting knowledge as a result they never kept complete accounting information therefore fail to effectively utilize financial data as basis for their decision and performance measurement. The use of technology in SMEs information processing has had a strengthening tie to firm performance as it shown in H_7. Authors in [63] further stress on the importance of IT support and book keeping for entrepreneurs effort to maintain an efficient firm operation, if taken for granted could be a great hindrance for business success.

VI. Conclusion

Coupled with the general impediment of information technology and financial system in developing economies, SMEs faces relatively copious challenges in terms of technology availability, awareness, and technical knowhow. Most of SMEs finds it is cost ineffective to monitor the present and updating to upcoming newer information technology versions which are emerging now and then. In addressing these challenges it is important to understand the relative impact of financial literacy and information technology on firm practices as it has been exemplified in this study. Situating SMEs owner/managers' financial knowledge and skills entwined with right technology adoption has an array of operational reinforcement of which if well understood and utilized will definitely enhance firm effective practices ultimately improve performance. Investment on SMEs risk management good practices and a good system of financial information collection, processing, keeping and dissemination has valuable tie with firm performance.

Limited by the nature of the study in terms of data and analytical approach, nevertheless the study gives glimpses of unharnessed benefits of integrating financial management and information technology in developing economy firms.

VII. Further Study Recommendation

There is much untapped potential of the applicability of technology adoption and its implications on awareness, access and actual use of financial products by individual and firms in developing economies. This study recommends more rigorous scientific research to be undertaken on this phenomenon including more variables of firm practices.

References

[1] M. A. Aremu and S. L. Adeyemi, "Small and medium scale enterprises as a survival strategy for employment generation in Nigeria," Journal of sustainable development, vol. 4, p. 200, 2011.

[2] T. Forester, High-tech society: the story of the information technology revolution: Mit Press, 1987.

[3] B. Hobijn and B. Jovanovic, "The information-technology revolution and the stock market: Evidence," American Economic Review, vol. 91, pp. 1203-1220, 2001.

[4] D. Byrne, S. Oliner, and D. Sichel, "Is the information technology revolution over?," 2013.

[5] M. Iansiti and K. R. Lakhani, "Digital ubiquity:: How connections, sensors, and data are revolutionizing business," Harvard Business Review, vol. 92, p. 19, 2014.

[6] A. Tarutė and R. Gatautis, "ICT impact on SMEs performance," Procedia-Social and Behavioral Sciences, vol. 110, pp. 1218-1225, 2014.

[7] M. A. A. Kumar, "Financial Literacy awareness among SME's in Western Division of Fiji," EPH-International Journal of Educational Research (ISSN: 2208-2204), vol. 1, pp. 12-27, 2017.

[8] F. Olawale, "The financial literacy of Micro Enterpreneurs in South Africa," Journal of social sciences, vol. 42, p. 7, 2014.

[9] S. Wise, "The impact of financial literacy on new venture survival," International Journal of Business and Management, vol. 8, p. 30, 2013.

[10] A. Drexler, G. Fischer, and A. Schoar, "Keeping it simple: financial literacy and rules of thumb," American Economic Journal: Applied Economics, vol. 6, pp. 1-31, 2014.

[11] H. K. Mutegi, P. W. Njeru, and N. T. Ongesa, "Financial literacy and its impact on loan repayment by small and medium entrepreneurs," 2015.

[12] P. Datta, "A preliminary study of ecommerce adoption in developing countries," Information Systems Journal, vol. 21, pp. 3-32, 2011.

[13] E. F. Schumacher and S. I. Beautiful, economics as if people mattered: thesis, 1977.

[14] A. Atkinson and F.-A. Messy, "Measuring financial literacy," 2012.

[15] R. R. Nelson, An evolutionary theory of economic change: harvard university press, 2009.

[16] A. S. Bharadwaj, "A resource-based perspective on information technology capability and firm performance: an empirical investigation," MIS quarterly, pp. 169-196, 2000.

[17] T.-P. Liang, J.-J. You, and C.-C. Liu, "A resource-based perspective on information technology and firm performance: a meta analysis," Industrial Management & Data Systems, vol. 110, pp. 1138-1158, 2010.

[18] R. Santhanam and E. Hartono, "Issues in linking information technology capability to firm performance," MIS quarterly, pp. 125-153, 2003.

[19] N. M. Menon, B. Lee, and L. Eldenburg, "Productivity of information systems in the healthcare industry," Information Systems Research, vol. 11, pp. 83-92, 2000.

[20] M. Abreu and V. Mendes, "Financial literacy and portfolio diversification," Quantitative finance, vol. 10, pp. 515-528, 2010.

[21] S. Agarwal, G. Amromin, I. Ben-David, S. Chomsisengphet, and D. D. Evanoff, "Financial literacy and financial planning: Evidence from India," Journal of Housing Economics, vol. 27, pp. 4-21, 2015.

[22] P. Babiarz and C. A. Robb, "Financial literacy and emergency saving," Journal of Family and Economic Issues, vol. 35, pp. 40-50, 2014.

[23] S. A. Cole, T. A. Sampson, and B. H. Zia, Financial literacy, financial decisions, and the demand for financial services: evidence from India and Indonesia: Harvard Business School Cambridge, MA, 2009.

[24] D. A. Adomako S, "Financial Literacy and firm performance: the moderating role of financial capital availability and resource flexibility," International Journal of Management and Organizational studies, vol. 3, 2014.

[25] P. Dahmen and E. Rodríguez, "Financial literacy and the success of small businesses: An observation from a small business development center," Numeracy, vol. 7, p. 3, 2014.

[26] P. Chepngetich, "Effect of Financial Literacy and Performance SMEs. Evidence from Kenya," 2016.

[27] A. H. Seyal, M. Noah Abd Rahman, and H. Awg Yussof Hj Awg Mohammad, "A quantitative analysis of factors contributing electronic data interchange adoption among Bruneian SMEs: A pilot study," Business Process Management Journal, vol. 13, pp. 728-746, 2007.

[28] G. Lukianchuk, "The impact of enterprise risk management on firm performance of small and medium enterprises," European Scientific Journal, ESJ, vol. 11, 2015.

[29] F.-A. Messy and C. Monticone, "The status of financial education in Africa," 2012.

[30] M. Berisha-Namani, "The role of information technology in small and medium sized enterprises in Kosova," in Fulbright Academy Conference, 2009, pp. 1-8.

[31] A. Parasuraman, "Technology Readiness Index (TRI) a multiple-item scale to measure readiness to embrace new technologies," Journal of service research, vol. 2, pp. 307-320, 2000.

[32] N. Tsikriktsis, "A technology readiness-based taxonomy of customers: A replication and extension," Journal of Service Research, vol. 7, pp. 42-52, 2004.

[33] Y. Pierrakis and L. Collins, "Crowdfunding: A new innovative model of providing funding to projects and businesses," 2013.

[34] URT, "Small and Medium Enterprises Development Policy," M. o. I. a. Trade, Ed., ed. Dar es Salaam, 2003.

[35] L. J. Servon and R. Kaestner, "Consumer financial literacy and the impact of online banking on the financial behavior of lower-income bank customers," Journal of Consumer Affairs, vol. 42, pp. 271-305, 2008.

[36] S. Bansal, "Perspective of technology in achieving financial inclusion in rural India," Procedia Economics and Finance, vol. 11, pp. 472-480, 2014.

[37] N. A. Rahman, Z. Yaacob, and R. M. Radzi, "An overview of technological innovation on SME survival: a conceptual paper," Procedia-Social and Behavioral Sciences, vol. 224, pp. 508-515, 2016.

[38] C. Chittithaworn, M. A. Islam, T. Keawchana, and D. H. M. Yusuf, "Factors affecting business success of small & medium enterprises (SMEs) in Thailand," Asian Social Science, vol. 7, p. 180, 2011.

[39] Z. M. Makhbul and F. M. Hasun, "Entrepreneurial success: An exploratory study among entrepreneurs," International Journal of Business and Management, vol. 6, p. 116, 2010.

[40] O. M. Oyeku, O. Oduyoye, O. Asikhia, M. Kabuoh, and G. Elemo, "On entrepreneurial success of small and medium enterprises (SMEs): A conceptual and theoretical framework," Journal of Economics and Sustainable Development, vol. 5, pp. 14-23, 2014.

[41] P. Fariselli, C. Oughton, C. Picory, and R. Sugden, "Electronic commerce and the future for SMEs in a global market-place: Networking and public policies," Small Business Economics, vol. 12, pp. 261-275, 1999.

[42] G. Premkumar and M. Roberts, "Adoption of new information technologies in rural small businesses," Omega, vol. 27, pp. 467-484, 1999.

[43] A. A. Eniola and H. Entebang, "SME firm performance-financial innovation and challenges," Procedia-Social and Behavioral Sciences, vol. 195, pp. 334-342, 2015.

[44] S. N.-I. M. Kamal, N. H. M. Yatim, S. Osman, M. N. Ali, N. M. Ali, and J. M. Jali, "Developing the Authenticity Framework of Electronic Business Records in SMEs Companies," Procedia Economics and Finance, vol. 31, pp. 834-838, 2015.

[45] K. W. Hollman and S. Mohammad-Zadeh, "Risk management in small business," Journal of Small Business Management, vol. 22, pp. 47-55, 1984.

[46] S. J. Huston, "Measuring financial literacy," Journal of Consumer Affairs, vol. 44, pp. 296-316, 2010.

[47] C. M. Lau and K. Roopnarain, "The effects of nonfinancial and financial measures on employee motivation to participate in target setting," The British accounting review, vol. 46, pp. 228-247, 2014.

[48] D. Garson, Partial Least Square: Regression and Structural Equation Models. North Carolina State University: Statistical Associates 2016.

[49] W. W. Chin, "The partial least squares approach to structural equation modeling," Modern methods for business research, vol. 295, pp. 295-336, 1998.

[50] O. Götz, K. Liehr-Gobbers, and M. Krafft, "Evaluation of structural equation models using the partial least squares (PLS) approach," in Handbook of partial least squares, ed: Springer, 2010, pp. 691-711.

[51] J. F. Hair, M. Sarstedt, C. M. Ringle, and J. A. Mena, "An assessment of the use of partial least squares structural equation modeling in marketing research," Journal of the academy of marketing science, vol. 40, pp. 414-433, 2012.

[52] J. Hulland, "Use of partial least squares (PLS) in strategic management research: A review of four recent studies," Strategic management journal, pp. 195-204, 1999.

[53] M. Höck and C. M. Ringle, "Strategic networks in the software industry: An empirical analysis of the value continuum," in IFSAM VIIIth World Congress, 2006, p. 2010.

[54] J. Henseler, C. M. Ringle, and M. Sarstedt, "A new criterion for assessing discriminant validity in variance-based structural equation modeling," Journal of the Academy of Marketing Science, vol. 43, pp. 115-135, 2015.

[55] T. S. Teo, S. C. Srivastava, and L. Jiang, "Trust and electronic government success: An empirical study," Journal of management information systems, vol. 25, pp. 99-132, 2008.

[56] A. H. Gold, A. Malhotra, and A. H. Segars, "Knowledge management: An organizational capabilities perspective," Journal of management information systems, vol. 18, pp. 185-214, 2001.

[57] P. M. Podsakoff, S. B. MacKenzie, J.-Y. Lee, and N. P. Podsakoff, "Common method biases in behavioral research: a critical review of the literature and recommended remedies," Journal of applied psychology, vol. 88, p. 879, 2003.

[58] N. Kock, "Common method bias in PLS-SEM: A full collinearity assessment approach," International Journal of e-Collaboration (IJeC), vol. 11, pp. 1-10, 2015.

[59] N. Podsakoff, "Common method biases in behavioral research: A critical review of the literature and recommended remedies," Journal of Applied Psychology, vol. 88, pp. 879-903, 2003.

[60] K. Saira, M. A. Zariyawati, and M. N. Annuar, "Information system and firms' performance: the case of Malaysian small medium enterprises," International business research, vol. 3, p. 28, 2010.

[61] U. B. Ihua, "SMEs key failure-factors: a comparison between the United Kingdom and Nigeria," Journal of Social Sciences, vol. 18, pp. 199-207, 2009.

[62] N. Maseko and O. Manyani, "Accounting practices of SMEs in Zimbabwe: An investigative study of record keeping for performance measurement (A case study of Bindura)," Journal of accounting and taxation, vol. 3, p. 158, 2011.

[63] E. G. Carayannis and M. Von Zedtwitz, "Architecting gloCal (global–local), real-virtual incubator networks (G-RVINs) as catalysts and accelerators of entrepreneurship in transitioning and developing economies: lessons learned and best practices from current development and business incubation practices," Technovation, vol. 25, pp. 95-110, 2005.

Experimental Study of Spatial Cognition Capability Enhancement with Building Block Learning Contents for Disabled Children

Kohei Arai, Taiki Ishigaki
Graduate School of Science and Engineering
Saga University, Saga City, Japan

Mariko Oda
Hagoromo University of International Studies
Osaka, Japan

***Abstract*—In this research, we develop learning teaching materials using building blocks for children with disabilities, and verify learning effect. It is important to prepare input equipment according to children with disabilities and to prepare learning materials according to the ability you have learned. Therefore, this time we developed a teaching material using building blocks to improve spatial recognition capability using touch pad and tablet as input device. It is decided to measure the effect by comparing the scores learned by actually combining the input device and the learning material. Through experiments with participants of disabled children, it is found that the learning contents are effective and appropriate for improvement of their spatial recognition capability.**

***Keywords*—Experiment; slope surfaces; interaction between two surfaces**

I. Introduction

Children have acquired the ability to become the basis of learning from activities such as living activities and playing. However, some children have restricted activities due to obstacles, and they miss opportunities to learn. For example, in the study of building blocks for young children In the research of Ito - Takahashi, the child A performs the actions such as "loading", "breaking", "hitting", "arranging", which are basic motions for playing using building blocks Experienced. After winning acts to pile up, they repeatedly challenged play and confirmed play and experienced "interval", "width/depth/height", "center of gravity", "balance", etc.

It can be pointed out that child A is acquiring the concept of "plane" "solid" through building blocks [1]. However, some children who have disabilities use the real building blocks. It is difficult to learn. Children with severely overlapping obstacles have difficulty in displaying intention based on voice and behavior, and there are many passive learning that only learning is heard, learning is listening. In special support schools, various education is conducted according to the progress of each individual development and disability based on the guidelines for teaching special support schools [2]. The faculty prepares for the learning environment tailored to the student. ICT can be used effectively in order to teach instruction according to students. Therefore, in this research, we aim to improve the spatial recognition capability by using a learning environment that combines input equipment according to student's disability and learning materials using building blocks. Learning materials, by changing input devices, consider the teaching materials that can be used regardless of the degree of disability.

Development of learning support software with CG animations for intellectually disabled children is conducted [3]. Spatial comprehension exercise system with 3D CG of toy model for disable children is proposed and is well reported [4]. Also, learning content for improvement of spatial comprehension capability with 3D CG toy model for disable children is proposed [5].

In this research, we develop learning teaching materials using building blocks for children with disabilities, and verify learning effect. It is important to prepare input equipment according to children with disabilities and to prepare learning materials according to the ability you have learned. Therefore, this time we developed a teaching material using building blocks to improve spatial recognition capability using touch pad and tablet as input device. It is decided to measure the effect by comparing the scores learned by actually combining the input device and the learning material.

Research background is described in the next section followed by learning content creation. Then experiments with specific disabled children are described followed by conclusion with some discussions.

II. Research Background

A. Disabled Children

A child with physical disability is a child with disabilities in motor function such as hands, feet, spine, etc. due to injury at birth or childbirth, or illness or accident at young age. The condition varies according to inconvenient parts and degrees. In the case where the exercise / motion of the right hand or right body only, or both feet, and even the whole body is inconvenient.

There is a match. In addition, the extent is also diverse, such as those who do not feel difficulty in everyday life, those who need prosthetic gear such as pine needle stick and wheel chair, and even those who need assistance for many activities [6].

Involuntary movement is an exercise that cannot normally be suppressed or can only be partially restrained [7]. Student becomes difficult to bend arbitrary movements arising temporarily in arms and hands. In that case, they may move their fingers to attract objects.

Intellectual disability is the state where the development of cognitive abilities has remained at a level where the cognitive ability development is generally delayed due to intellectual functional disorder that occurred until the developmental stage. The degree of intellectual impairment is shown below [8].

- Mild intellectual disability refers to intellectual disability with IQ generally 50-70. Meals, clothes removal, drainage etc. There is no hindrance to the everyday life skills of. But the development of language is slow, even 18 years old and above elementary school. It is often stayed at the raw level of academic ability.
- Medium intellectual disorder generally refers to intellectual disability with IQ of 35-50. Language development and delay in athletic ability. They can do yourself partly, but it is difficult to do everything.
- Severe intellectual disability refers to a mental disorder whose IQ is generally 20-35. Development of language and motor functions is slow. On the learning side, it is limited to reading and writing hiragana. The development of emotions is immature and one thing around us. Because it is difficult to do with people, protection and assistance may be necessary for food, clothing and shelter.
- Most severe intellectual disability refers to a mental disorder with an IQ of 20 or less. Words do not develop. It's almost time to stay screaming. They can handle all aspects of their life.

Mental rotation is a cognitive activity that rotates the visual stimulus presented in a rotated state mind and recognizes figures in an upright state without rotation [9].

Spatial recognition capability is the ability to quickly and accurately grasp and recognize the state and relationship of objects in 3D space such as position, orientation, attitude, size, shape and spacing of objects.

B. Children in Concern

Student A (Male) has a high intellectual level but it is difficult to express by speech language due to paralysis of upper and lower limbs. At school, he is sitting in a wheelchair. The physical condition of Student A can move his hands and arms with his own will and grips and opens his hands. When grasping objects, there is an influence of involuntary movements from one minute to five minutes. Legs are difficult to move by paralysis. Although there are physical restrictions due to paralysis of the upper and lower limbs in response to the instruction of the teacher, it is possible to speak "ah" "uh" or waving his own intention.

The range that Student A moves mainly by hand is shown in Fig. 1. Student A moves his face to see for those who are interested by judging who the ears are very good at the footsteps heard from the corridor. As an educational support for one year from 2015, we have been developing input devices to support Student A's intention display. Educational support from 2016 aims at improving the spatial recognition ability by learning to show intention by using Student A's own hand rather than listening, passive learning which is only viewing. Therefore, prepare two kinds of problems using building blocks and learn mental rotation.

Student B (Female) is difficult to express by speech language with the most severe intellectual disorder. Student B understands the daily movements such as correcting posture and instructions on learning using things, and acts as instructed. As a result of educational support so far, she can recognize plan figures and pictures. We tried to grasp the touch pen and select the same shape and picture as Fig. 2 using a tablet [10]. When we started educational support, we dropped or thrown tablet. However, with repeated support, we understand the content of the learning, and now we can correctly answer the problem more than a healthy person. Therefore, in this educational support, we will see what kind of response to the learning of the three-dimensional figure. We think that the ability to capture things in three dimensions in the future life is necessary and aim to improve mental rotation as well as Student A to improve space recognition ability.

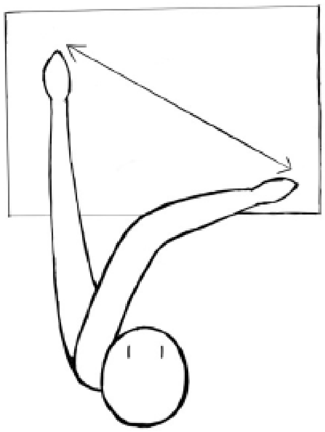


Fig. 1. Movable range of his / her hand.

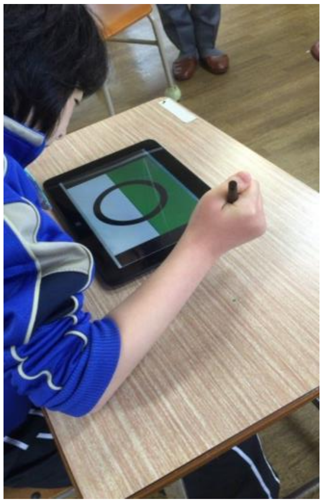


Fig. 2. Photo which shows character learning process.

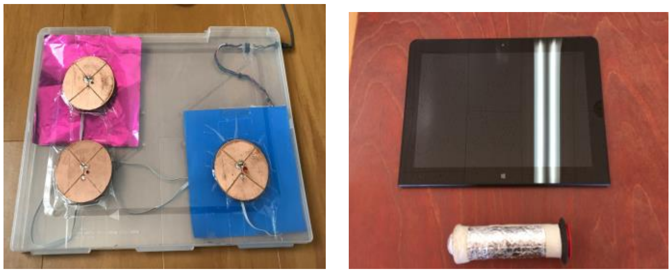
III. Experiments

A. Experimental Configuration

In this section, we will describe the input device according to the target state described in Section 2. Touch pad that touches the copper plate to output signals and touch table to learn by touching the pen were prepared as input equipment considering the physical restriction due to the target obstruction.

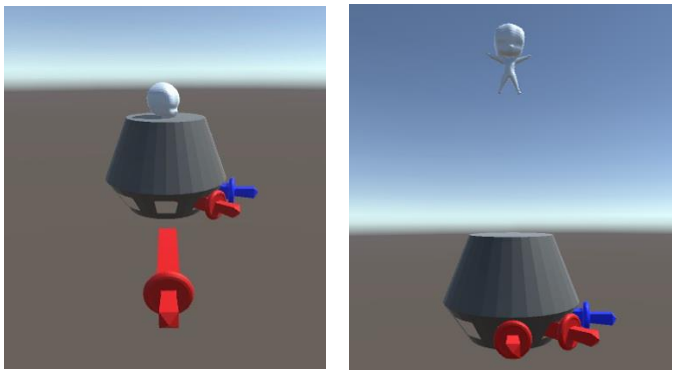
Student A can wield his hand at his own will as described in 2.2. Therefore, we prepared a touch pad that outputs signals only by touching the hands. The touchpad is a mechanism that outputs signals (keys 1, 2, 3) by touching three copper plates. For this time, we are trying to learn by touching the upper left and lower right copper plates. Student A easily moves the hand to the upper left and the lower right, so the buttons are arranged accordingly. Understand the copper plate to understand that it corresponds to the learning material choice by sticking colored paper.

Student B used the tablet and the stylus used for educational support from three years ago. Student B sometimes drops things or throws things. Therefore, we use a tablet resistant to impact. In addition, because Student B is difficult to grab a thin pen, it made thick the touch pen so that it grips easily. The operation of the learning material can be learned by touching the option with the touch pen. Touchpad and tablet and touch pen are shown in Fig. 3 as building block learning tools.



(a) Touchpad (b) Tablet and touch pen

Fig. 3. Building block learning tools.



(a) Alternately stab a sword (b) If you hit the hit area you fly

Fig. 4. Toy model.

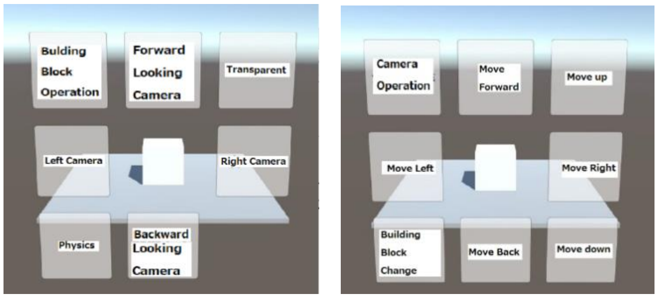
We describe the environment which we can play with 3DCG. For this time, learning is done with the teaching materials using 3DCG for the first time. Therefore, in order to give interest to 3DCG teaching materials which we have not seen so far, we have reproduced the familiar toys. By playing with the toy with that toy, we have not been able to play by myself, but by using this input device we can experience what we can operate and draw interest. After that, let's learn when the operation of the input device gets used.

Toy model reproduced the toy playing with teacher usually with 3DCG. Until now it was difficult to play with toys in hand by himself, but by using the input device according to the subject, they can operate their own toys and play. In the developed toy, they can "turn" the red of the touchpad by touching and touch the blue to "stab the sword" (Fig. 4). By touching the button on the tablet by displaying the button, the same action can be done.

We asked the special support school teacher to evaluate the developed toy model. Children who have difficulty playing with objects due to the obstacles this time are highly valued, such as wanting to try also by other students because they play very happily.

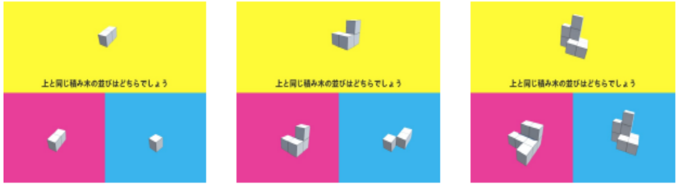
Children who have difficulty playing toys due to obstacles tend to give up a lot of things, becoming passive even if they are playful if their own thoughts do not pass. Therefore, it was found that the environment where the input device and the toy model can be combined and played as desired can be effective for children's active activities.

The building block play was reproduced on the PC (Fig. 5). The subjects A and B of this time let the students play with complicated operations but were not able to play, but had another obstacle. The student has never played with a building block because it is difficult to move down from the neck due to obstacles. Therefore, we prepared an environment where you can play building blocks on a PC. The student can move the mouse pointer using the line of sight input device. The operation method of this building block is to operate by performing the functions by hovering over the button displayed around the building block.



(a) Camera switch mode (b) Building block operation mode

Fig. 5. Building block toy model.

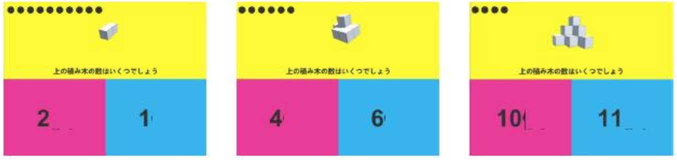

上と同じ積み木の並びはどちらでしょう
上と同じ積み木の並びはどちらでしょう
上と同じ積み木の並びはどちらでしょう

(a) Level 1 (b) Level 2 (c) Level 3

Fig. 6. Three difficulty level of building block toy model.

The alignment teaching material of building blocks is a question concerning the arrangement of building blocks. Since the object understands the planar figure, it is a teaching material to check whether you understand the three-dimensional shape before doing the building block problem. By rotating the arranged building blocks so that the whole is visible, we made it possible to see the viewpoint seen from the front and the viewpoint seen from behind. The question presentation (yellow frame) is placed on the screen and the options are arranged on the left and right (red frame, blue frame) under the screen. When presenting the question, rotate 360° so that the whole can be seen and pause at every 180° rotation. Three difficulty levels are set to change difficulty level, question number, number, and number of questions (Fig. 6).

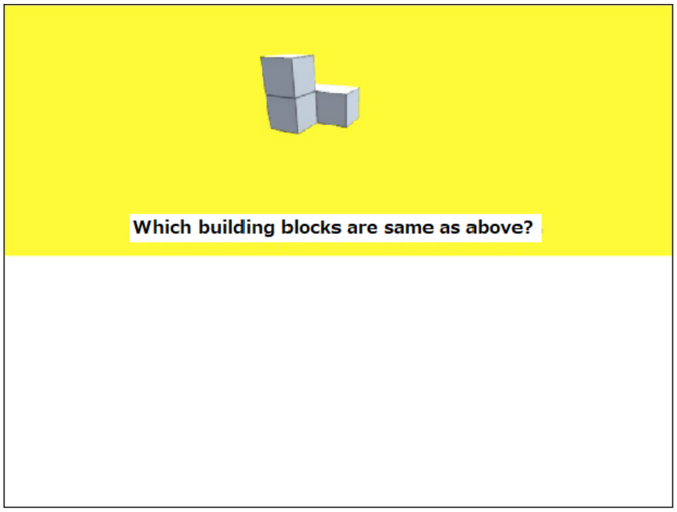
This teaching material has become a teaching material to check whether the number of building blocks necessary to comprehend the shape of the arranged blocks and to compose the arranged shape can be counted. This teaching material also rotates so that the whole of the building blocks arranged in the same way as previously described model which is shown in Fig. 6. The number of remaining questions from this teaching material is displayed on the upper left of the screen as requested by the teacher (Fig. 7).


●●●●●●●●●●
上の積み木の数はいくつでしょう
2
1
●●●●●●
上の積み木の数はいくつでしょう
4
6
●●●●
上の積み木の数はいくつでしょう
10
11

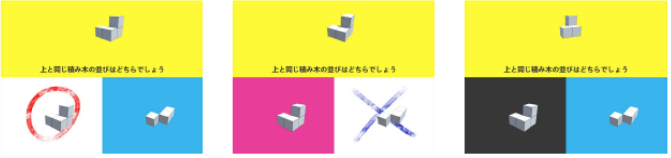
(a) Level 1 (b) Level 2 (c) Level 3

Fig. 7. Three difficulty level of block counting toy model.

In learning process flow, questions are presented when selecting the question number first (Fig. 8(a)). As soon as it is presented, options are displayed after the question sentence is read aloud. When answering, the question presentation section is inclined at the same angle as the selected building block, so that the difference between correct answer and incorrect answer can be understood (Fig. 8(b)).


Which building blocks are same as above?

(a) Question


上と同じ積み木の並びはどちらでしょう
上と同じ積み木の並びはどちらでしょう
上と同じ積み木の並びはどちらでしょう

(b) Answer (Correct (left), Incorrect (middle), Mouse miss-operation

Fig. 8. Building block toy model.

The explanatory learning material is a teaching material for explaining the wrong answer after confirming whether the arrangement and number of building blocks are understandable. Choose the wrong answer with the learning materials, and the teacher draws a commentary with the viewpoint of the building blocks, the movement of the building blocks, and the building blocks drawing attention (Fig. 9). The operation procedure is shown below:

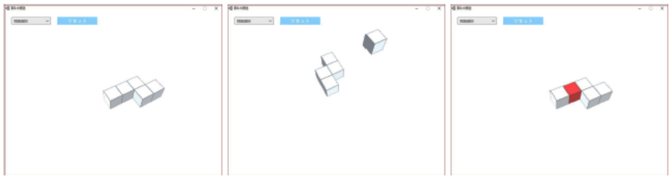
a) Move drag block while right clicking the block to move block.

b) Move drag view while left clicking on places other than building block to move view point.

c) Right click the block to change the color of the block.

d) Mouse wheel scaling to enlarge or shrink the block.

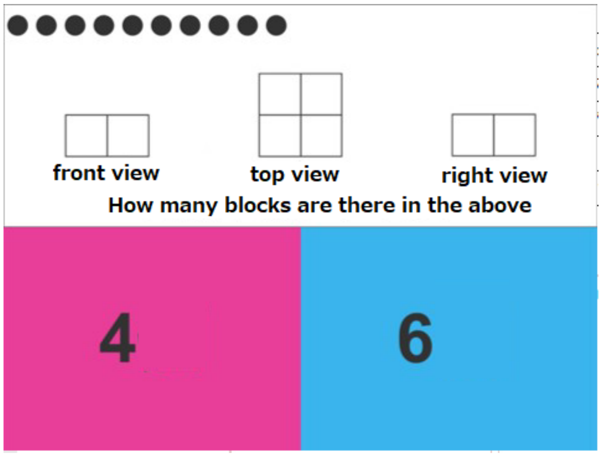
e) Click reset button to reset the position of the building block.



(a) Change the view point (b) Move the block (c) Look at the block

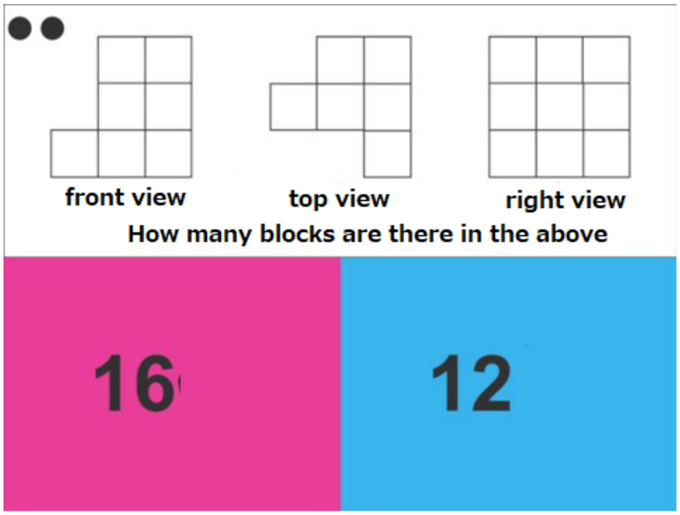
Fig. 9. Displaying the instructions.

The building block problem is a teaching material to answer the number of building blocks by performing mental rotation while imagining the arrangement of building blocks in the head from the viewpoint of the front, the upper, and the right from the viewpoint of the front side, the upper side, and the right side. This learns after learning the understanding of the shape of the building blocks and the number of building blocks with the learning materials of Fig. 10 (a) and 10 (b).



How many blocks are there in the above

(a) Question No.1



(b) Question No.2

Fig. 10. Building block questions.

In this question, let them answer the maximum number of blocks of blocks from the figure. By imagining from three viewpoints and grasping the position of the building blocks, it improves the space recognition ability. It is very good to be able to know the students' comprehension level as an evaluation by a special support school teacher. In school life, students have a lot of passive learning and students' intention is reflected in teaching materials, so that mistakes can be supplemented with commentary materials.

Even for Student A is difficult to touch the real building blocks, they can learn a lot through building teaching materials and it is very effective for learning. We would like to see students' reactions by letting other students learn using building blocks teaching materials. Very good evaluation was obtained from the target teacher this time.

We experimented to see if the shape of the building block was recognizable in three dimensions. Tables I and II as well as Fig. 11 show the results of making the blocking problem of the blocks.

A cross mark in the answer result is a case of an unintended answer or incorrect answer due to involuntary movements. When the score of Quiz 3 of the first learning and the score of Quiz 1 of the second learning were high, after seeing the screen for a while, they moved hands by hand and answered (average response time is 5 minutes).

TABLE I. SCORE AT THE FIRST TIME (STUDENT A)

Quiz1	Ans.	Quiz 3.	Ans.
1	○	1	○
2	○	2	x
3	○	3	x
4	x	4	○
5	○	5	○
6	○	6	○
7	x	7	○
8	x	8	○
9	x	9	○
10	x	10	○

TABLE II. SCORE AT THE SECOND TIME (STUDENT A)

Quiz1	Ans.	Quiz3	Ans.
1	○	1	○
2	○	2	x
3	○	3	○
4	○	4	○
5	x	5	○
6	○	6	x
7	○	7	○
8	○	8	○
9	○	9	○
10	○	10	x

When Quiz 1's first learning score was low, they answered without looking at the screen (average response time is 24 seconds). For the wrong problem, the choice of correct answer and the choice of incorrect answer were relatively similar by one block difference. Since the score of learning when student A himself was concentrating was high, it was confirmed that he was firmly recognizing the shape of the building block.

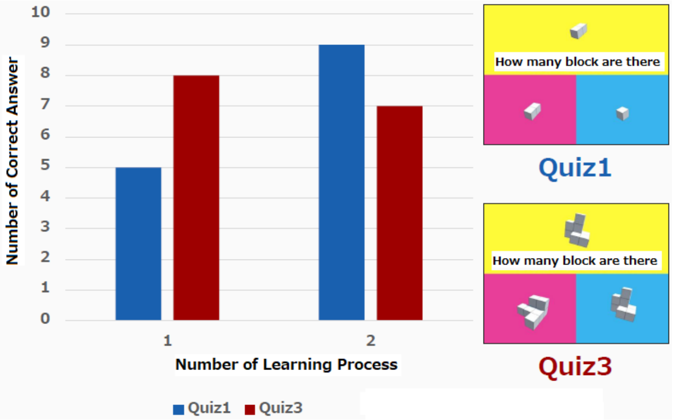


Fig. 11. Results of making the blocking problem of the blocks.

We experimented to see if we can count the number of building blocks that are lined up. Tables III to V as well as Fig. 12 show the results of counting the number of building blocks.

The score of Quiz 1 is high for the 1st and 2nd learning, but it is wrong with Question 8 (Fig. 13). What is considered as a cause is that since the problem is being rotated, it seems that the numbers are wrong because the front and rear shapes are similar.

TABLE III. SCORE AT THE FIRST TIME (STUDENT A)

Quiz1	Ans.	Quiz2	Ans.
1	○	1	x
2	○	2	x
3	○	3	x
4	○	4	x
5	○	5	x
6	○	6	x
7	○	7	○
8	x	8	○
9	○	9	x
10	○	10	x

TABLE IV. SCORE AT THE SECOND TIME (STUDENT A)

Quiz1	Ans.	Quiz2	Ans.
1	○	1	○
2	○	2	x
3	○	3	x
4	○	4	x
5	○	5	○
6	○	6	x
7	○	7	○
8	x	8	○
9	○	9	○
10	○	10	○

TABLE V. SCORE AT THE THIRD TIME (STUDENT A)

Quiz1	Ans.	Quiz2	Ans.
1	x	1	○
2	x	2	x
3	○	3	○
4	x	4	○
5	○	5	x
6	x	6	○
7	x	7	○
8	○	8	x
9	x	9	x
10	x	10	○

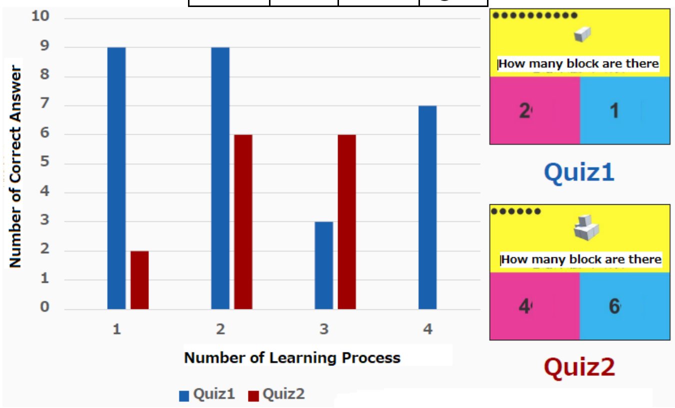


Fig. 12. Results of making the blocking problem of the blocks for Question 8.

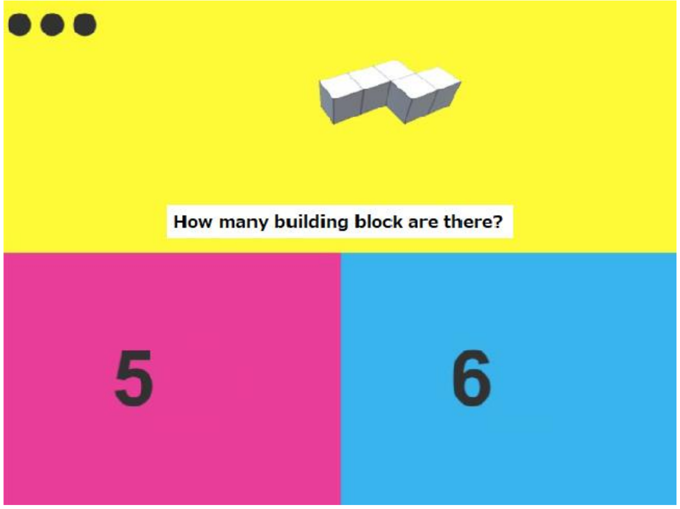


Fig. 13. Question No.8.

It seems to be because it judged that the number of building blocks before and after the problem presentation part and options are repeated while looking it repeatedly. Regarding Quiz 2, when three times of learning is repeated for the correct part, a high score is obtained. It was confirmed that the number of building blocks can be counted from the score.

Student A reacts sensitively to the sound, it is a level to judge who is just a footstep. He moves his face every time. He heard footsteps and tries to see it. Therefore, student A was enclosed in a partition and concentrated on the screen (Fig. 14). Until now, the teacher was supplemented (such as reading the position of the face and the problem) next to Student A, but Student A alone saw the screen and answered more. Furthermore, in order to draw motivation for learning, we changed the scenery at the time of learning for each question (Fig. 13). In the change of the landscape, although there was no influence on the score, time to see the screen and the input device increased, and it was possible to change the willingness to learn.

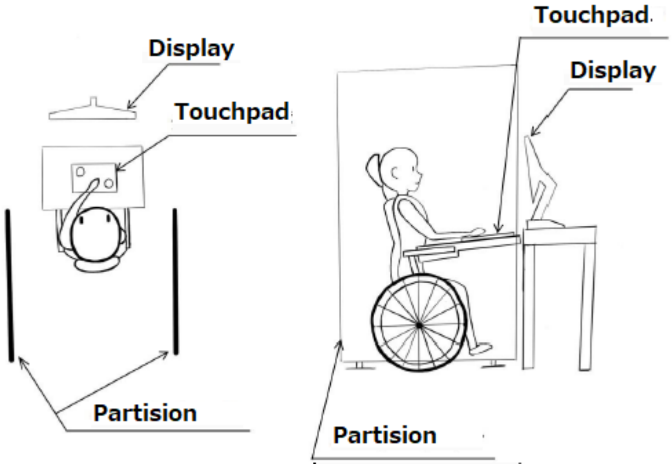


Fig. 14. Learning environment.

This section describes experiments on reactions of Student B to learning three-dimensional drawings. Student B knows that long-term learning is necessary from educational support so far to learning ability through learning. Therefore, we looked at what kind of response it would be to learning of a three-dimensional figure to work on for the first time. We got motivation for learning and examined the iterative learning till learning can be done just like educational support so far.

In addition to the educational support of plan figures and pictures so far, learning of building skill teaching materials of a three dimensional figure was done. Student B was unable to bring motivation for learning by dropping or throwing the tablet on the first teaching material. However, we were interested or had time to look at the rotating problem. From the teacher who saw this reaction, there was an opinion that the shape and the size did not change very much even if the plan view rotated, but it is difficult to rotate the three-dimensional figure because the size and shape change.

IV. CONCLUSION

In this research, improvement of spatial recognition ability and learning materials were effective for education for children with disabilities using educational materials using input devices and building blocks suitable for the target and educational support. It became possible for Student A to display their own intention rather than past passive learning. As a result, it was confirmed that Student A's positional relationship and spatial grasping ability are provided. Since the period of educational support was short, the measurement of space recognition ability due to the block trees problem was insufficient and we could not confirm the improvement of capacity.

Student B dropped or threw the teaching material for the first time for the teaching material to see, so learning effect was not seen. However, since we have to make repeated learning for a long period of time from educational support so far, we will measure the learning effect by looking at the future response. Overall, we could not measure the learning effect from the learning and the score using the teaching materials, but in terms of providing a learning environment tailored to each individual disability, we highly evaluated from the faculty and tried it for other students, I was able to know that it was effective as well.

As a future task, it is difficult to learn using only input devices and learning materials. Differences appear in the ability to concentrate due to the sound and scenery at learning, and it is influenced by the learning score. In the future, it is a task to measure the learning effect by long-term educational support while considering not only the object but also surrounding environment.

ACKNOWLEDGMENTS

Author would like to thank the participants (Student A and B) of the experiment.

REFERENCES

[1] Ito Tomisato, Takahashi Toshiyuki: Various developmental features seen in building blocks of an infant, Journal of Art Education Society, 32 (0), pp 41-53, 2011.

[2] Special education school guidelines for teaching, etc., high 1,2009.

[3] Kotaro Taguchi, Mariko Oda, Hiroshi Kouno Seio Oda, Kohei Arai, Development of learning support software with CG animations for intellectually disabled children, Journal of Education System and Information Society of Japan, 30,1, 48-56, 2014

[4] Kohei Arai, Taiki Ishigaki, Mariko Oda, Spatial comprehension exercise system with 3D CG of toy model for disable children, IJACSA, 8, 4, 189-194, 2017.

[5] Taiki Ishigaki, Mariko Oda, Kohei Arai, Learning content for improvement of spatila comprehension capability with 3D CG toy model for disable children, Proceedings of the 3rd ADADA Conference of the Asian Society of Digital Art and Design, at the Kurume Institute Technology, 2017.

[6] Social Welfare Corporation Japan Association for Handicapped Children http://nishikyo.or.jp/about_us/index.html (Accessed on June 21 2018)

[7] Kaori Dragon: Examination of involuntary movements, clinical neurophysiology, vol. 43, No. 4, pp. 122-141, 2015.

[8] Developmental navigation https://h-navi.jp/column/article/73 (Accessed on June 21 2018)

[9] Information dissemination for therapists http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0180985; http://www.comp.tmu.ac.jp/locomotion-lab/higuchi/higu-therapist-old17.html (Accessed on June 21 2018)

[10] Moriko Oda, Seio Oda, Hiroshi Kono, Hideto Sazuka, Masahito Takahashi: Engineering and Educational Support for Regional Special Support Schools by Service Learning, Research Report on Educational System Information Society, vol. 29, No. 6, pp 115- 120, 2015.

AUTHOR'S 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 counselor 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 Science Commission "A" of ICSU/COSPAR since 2008 then he is now award committee member of ICSU/COSPAR. He wrote 37 books and published 570 journal papers. He received 30 of awards including ICSU/COSPAR Vikram Sarabhai Medal in 2016, and Science award of Ministry of Mister of Education of Japan in 2015. He is now Editor-in-Chief of IJACSA and IJISA. http://teagis.ip.is.saga-u.ac.jp/index.ht

Improved Langley and Ratio Langley Methods for Improving Sky-Radiometer Accuracy

Kohei Arai
Dept. of Information Science
Saga University
Saga City, Japan

***Abstract*—Improved Langley Method (ILM) is proposed to improve the calibration accuracy of the sky-radiometer. The ILM uses that the calibration coefficients of other arbitrary wavelengths can be presumed from the calibration coefficients in a certain reference wave length, and improves the calibration accuracy of a full wave length region by Ratio Langley Method (RLM) in long wavelength paying attention to calibration accuracy being good comparatively was proposed. Specifically, the calibration coefficient of other wavelengths was presumed by the RLM from the calibration factor by ILM in 0.87 micrometer. The numerical simulation based on measured data of solar direct and aureole when the calibration error of the proposed method was evaluated about the case where ±3% and ±5% of measurement error is superimposed on the measurement data solar direct and aureole, the maximum with error was 0.0014 and 0.0428, and they of ILM were 0.011 and 0.0489. Therefore, the proposed calibration method is robust for a measurement error compared with ILM, and was understood that highly precise calibration is possible over full wavelength. When the standard deviation of a calibration coefficients estimated the accuracy of the proposed calibration method based on the measured data of the sky-radiometer for 15 days which fits calibration among the measured data for four years or more, it was 0.02016, and since it was smaller than the standard deviation 0.03858 of the calibration coefficients by ILM, the predominance of the proposed calibration method has been confirmed.**

***Keywords*—Calibration; Langley plot; improved Langley method; ratio Langley method; aerosol optical depth; volume spectrum**

I. Introduction

Attempts have been made to estimate the aerosol complex refractive index and particle size distribution using the downward radiance on the ground obtained by a radiometer (sky-radiometer) that measures the direct sunlight, scattering, and marginal light [1]. There is the aerosol parameter distribution on a global scale like AERONET: Aerosol Robotic Network and SKYNET: Sky-radiometer Network by developing the sky-radiometer all around the world, and aerosol which is a contributing factor of earth radiation budget, global warming etc. Attempts have been made to elucidate the wide area distribution of. In that case, the calibration accuracy of the sky-radiometer greatly affects the aerosol complex refractive index and the estimation accuracy of the particle size distribution, which is extremely important.

Solar irradiance measurement on the ground is a very effective method for knowing the optical and physical properties of the sun and the Earth's atmosphere [2]-[4]. The first of this method is the Langley plot method based on the Beer Bouquet Lambert's rule [5]: LPM [6]. Measure the direct sunlight from sunrise to southern middle time to estimate the atmospheric mass to 0, that is, to estimate radiance outside the solar atmosphere. A calibration method has been proposed to determine the calibration coefficient of the sky-radiometer so that the difference between this estimated value and the stable solar atmospheric radiance and irradiance model [5] called so-called solar constant is 0, which is widely used [7], [8]. In addition, a method of reducing the influence of noise in observation has been proposed [9].

The calibration of the sky-radiometer has dealt with the Rati Langley method [10] which aimed at improving the calibration accuracy by obtaining the calibration coefficient according to the ratio between the LPM and the measurement wavelength, and the fluctuation of the atmospheric condition in the Langley plot This is done by the modified Langley method (ILM [11]). LPM is effective when the atmospheric condition does not change, but such a state is difficult to obtain, generally it is not possible to expect high calibration accuracy. ILM measures not only direct sunlight but also scattering and marginal light on the ground surface, estimates the optical thickness of air molecules, ozone, aerosol, aerosol complex index of refraction, and particle size distribution, and obtains the scattering phase function from these The scattering contribution component is estimated and the calibration coefficient of the sky-radiometer is determined with the downward radiance on the ground where the value obtained by multiplying the aerosol optical thickness by the atmospheric mass becomes zero. Therefore, since ILM does not assume stability of atmospheric condition for a long time compared with LPM, relatively high calibration accuracy can be expected.

In addition, the RLM utilizes the fact that the logarithm of the optical thickness and the logarithm of the wavelength are in a linear relationship, and by taking into account the aerosol particle size distribution by the aerosol optical thickness at a certain reference wavelength, an aerosol of any arbitrary wavelength and estimates the optical thickness. The calibration coefficient of the sky-radiometer is determined by the downward radiance (irradiance) on the ground where the proportional coefficient of the linear relation is multiplied by the aerosol optical thickness at the reference wavelength and the atmospheric mass to zero.

This research aimed at newly devising a calibration method that combines the merits of ILM and RLM and to further improve the calibration accuracy. In other words, I propose a method that achieves high calibration accuracy over all wavelengths by exploiting the features of ILM with high calibration accuracy in the relatively long wavelength region, and using this as the reference wavelength of RLM to obtain calibration coefficients of other arbitrary wavelengths

At this time, consider the aerosol particle size distribution. That is, the volume spectrum [11] is reanalyzed using sunlight and marginal light radiance. I evaluated the proposed method based on measured numerical simulation and measured values themselves, and ascertained the superiority over the existing method, I report here. Section 2 introduces LPM, ILM and RLM as a theoretical background on calibration of sky-radiometer, and explains the proposed method in Section 3. In addition, the numerical simulation method and results in Section 4 and the advantage of the proposed method based on measured data are shown in Section 5.

II. Proposed Method

A. Theoretical Background

LPM observes the sun's direct light *F* from the sunrise to the south middle and estimates the radiance F_0 of the sun's atmosphere outside from (1).

$$\ln F = \ln F_0 - m\tau \tag{1}$$

where *F* and F_0 are the direct solar radiation radiance on the earth's surface and the sun's atmospheric radiance outside the atmosphere, τ is the optical thickness of the whole atmosphere and *m* is the atmospheric mass. It is a precondition that the optical thickness of the whole atmosphere does not change during observation time for that reason,

As a result, LPM is effective only at high-altitude sites where the atmosphere is clear and stable with consistent conditions. In addition, the calibration accuracy of measuring equipment for solar direct light is also important. This is evident from the fact that it appears as an estimation error of the calibration coefficient of 2.6 to 10% [12] as it is clear from the past study evaluating the influence of the calibration precision in the plural aerosol models. ILM based on simultaneous observation of sunlight and marginal light was proposed to cope with the atmospheric condition variation in the Langley plot. The optical thickness of the aerosol is estimated by Volume Spectrum Analysis (VSA). At VSA, solar marginal light is replaced by the relative radiance of (2).

$$R(\theta) = \frac{F(\theta)}{Fm\Delta\Omega} = \omega\tau P(\theta) + q(\theta) \tag{2}$$

where, *R* (θ) at the angle θ is the relative radiance at the angle θ normalized by the solar direct ray *F* of the sun marginal light: *F*(θ), $\Delta\Omega$ is the small solid angle, ω is the single scattered albedo, and $q(\theta)$ is multiple scattering component. Also, $P(\theta)$ is the scattering phase function of the aerosol and the air molecule at the scattering angle θ, and is defined by (3).

$$P(\theta) = (\omega_a \tau_a P_a(\theta) + \omega_m \tau_m P_m(\theta)) / \omega\tau \tag{3}$$

where, ω_a , τ_a and $P_a(\theta)$ are the single scatter albedo, optical thickness and phase function of the aerosol, and ω_m, τ_m and $P_m(\theta)$ are those of the air molecule. If the aerosol particles are spherical homogeneous, from the Mie scattering theory $\omega_a \tau_a P_a(\theta)$ is defined by (4).

$$\omega_a \tau_a P_a(\theta) = \int_{r1}^{r2} K(\theta, kr, \tilde{m}) v(r) d\ln r \tag{4}$$

where, $v(r) = (4\pi/3) r^4 n(r)$, and $n(r)$ is the vertical particle size distribution of the aerosol. In addition, $k = 2\pi/\lambda$ and $\tilde{m} = n - i\xi$ is the aerosol complex index of refraction, $K_{ext}(kr, \tilde{m})$, $K(\theta, kr, \tilde{m})$ is the integral kernel function defined by Mie scattering theory. An approximate solution of the aerosol volume spectrum $v(r)$ is obtained by using a radiative transfer inversion that iteratively updates the multiple scattering contribution $q(\theta)$ according to the measurement value of the sun marginal light [11]. In this way the aerosol optical thickness can be estimated by (5).

$$\tau_a = \int_{r1}^{r2} K_{ext}(kr, \tilde{m}) v(r) d\ln r \tag{5}$$

By rewriting equation (1),

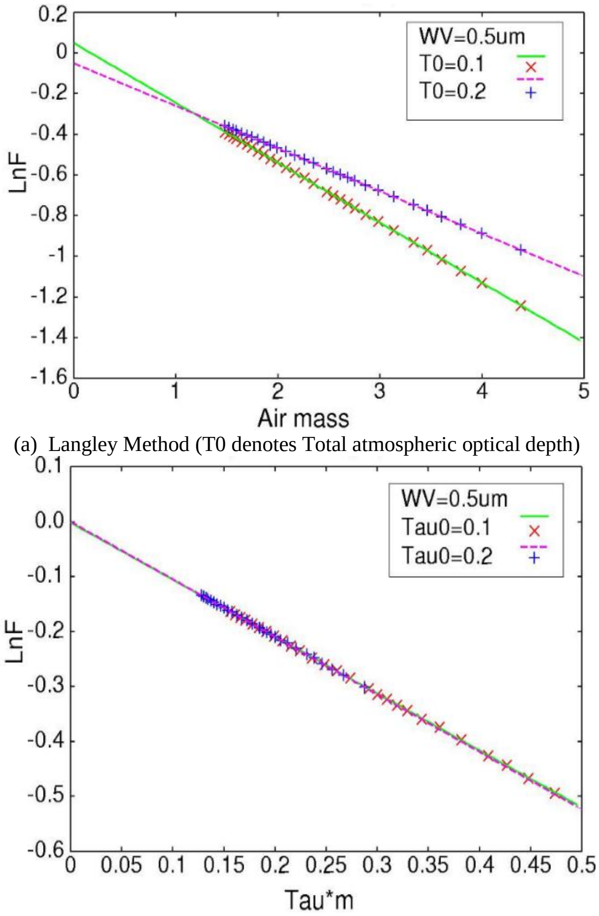
$$\ln F + m(\tau_m + \tau_o) = \ln F_0 - m\tau_a \tag{6}$$

where, τ_o is the optical thickness of ozone. Therefore, the calibration coefficient is obtained when $m\tau_a$ = 0. This method is called a modified Langley method (ILM). This method excludes most of the influence on atmospheric state change in aerosol optical thickness estimation by solar marginal light observation and is superior to LPM with deterioration of calibration coefficient estimation accuracy due to atmospheric state change. Fig. 1 shows comparison between LPM and ILM.

In this figure, despite the total atmospheric optical thicknesses being both 0.1 and 0.2, the estimated value of radiance outside the solar atmosphere is not matched because of the calibration problem in the case of LPM, but in the case of ILM, the radiance outside the solar atmosphere is matched when both the optical thicknesses are 0.1 and 0.2.

$$\ln F / m = \ln F_0 / m - \tau \tag{7}$$

The proposed method deforms (1) to reduce the influence due to temporal and spatial atmospheric conditions [13].



(a) Langley Method (T0 denotes Total atmospheric optical depth)

(b) Improved Langley Method (ILM), Tau0 denotes

Fig. 1. Illustrative comparison between Langley method and Improved Langley Method (ILM).

It is also applied to measurement data of Multi-Filter Rotating Shadow Band Radiometer (MFRSR) [14] and is trying to use it for atmospheric correction [15]. Furthermore, accuracy comparison with regular LPM has also been done, proving that the precision according to (7) is higher [9].

B. Proposed Method

In solving the radiative transfer equation by inversion due to influences such as the brightness of marginal light θ <30°, scattering by air molecules, error in peripheral light measurement, volume spectrum estimation error, other assumptions about atmospheric conditions, etc. The effect of multiple scattering tends to be estimated low. These errors appear when VSA estimates the aerosol optical thickness. An example is shown in Fig. 2.

Aerosol optical thickness was measured at Saga University under fine weather on November 26, 2003, December 3 and December 4, 2003 by the Sky-radiometer POM-1 manufactured by PREDE Co. Ltd. This figure compares the aerosol optical thickness estimated by VSA and the volume spectrum obtained by reanalysis using the measurement result by the Sky-radiometer. The difference between them is as much as 10%, which means that the calibration coefficient estimation accuracy obtained by ILM will be low. The figure also shows that the difference between the aerosol optical thicknesses of both is smaller in the long wavelength region than in the short wavelength, meaning that the accuracy is higher when the calibration coefficient is obtained in the long wavelength region.

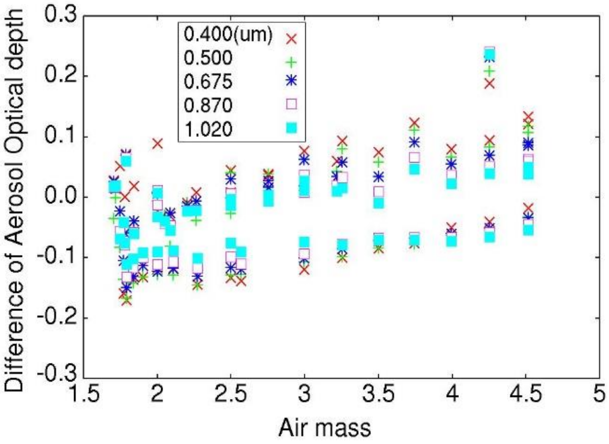


Fig. 2. The Differences of aerosol optical depth by means of VSA and reanalysis of volume spectrum from air-mass 1.5 to 4.5. Data are observed by POM-1 Sky-radiometer in 11/26/2003, 12/03/2003 and 12/04/2003 at Saga, Japan.

Fig. 3 shows the minimum of the estimated sunlight direct radiance when the measurement error is ± 10% for the optical thickness change range (0.08 to 0.28) under the typical atmospheric condition of Fig. 2 (L2) and the maximum (L1) logarithm. When compared with the case without error, it is found that there is a difference of -2% between the former and the + 3% from the latter. In other words, it means that the measurement accuracy of the optical thickness of ± 10% appears as the logarithm difference of the direct solar radiation radiance of 5%.

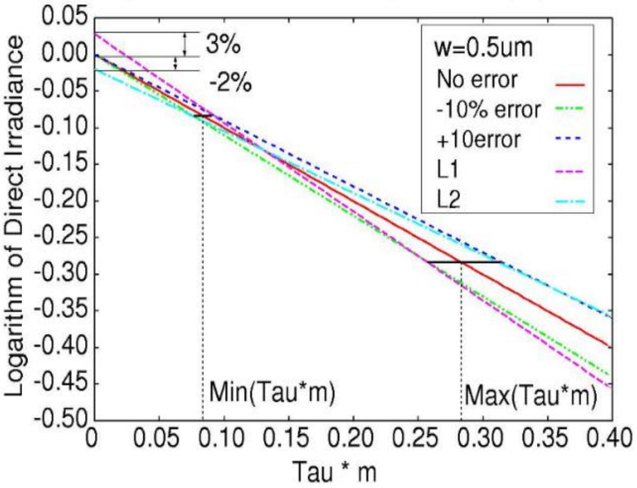


Fig. 3. Importance of optical depth measurement accuracy for Improved Langley Method: ILM.

From the aerosol optical thickness at a certain reference wavelength, there is a method of obtaining it from other wavelengths in consideration of the particle size distribution $f(r)$ [10]. This method is called RLM. At this time, the aerosol optical thickness can be obtained by multiplying the particle diameter distribution by the logarithm of the particle diameter *r* and integrating by multiplying by the kernel function.

$$\tau_a(\lambda,t)=\pi A(t)\int K_{ext}(r,\lambda)f(r)d\ln r \qquad (8)$$

Therefore, the ratio of the aerosol optical thickness at different wavelengths is a constant as in equation (9).

$$\tau_a(\lambda_1,t)/\tau_a(\lambda_2,t)=\tau_a(\lambda_1,t_0)/\tau_a(\lambda_2,t_0)=\psi \quad (9)$$

where, $A(t)$ is a constant for deriving the actual particle shape distribution from the particle size distribution shape at time t. From these, it is found from the calibration coefficient at the reference wavelength at other wavelengths.

$$\ln F(\lambda_1)+m(\tau_m(\lambda_1)+\tau_o(\lambda_1))=\ln F_0(\lambda_1)-\psi m\tau_a(\lambda_0) \qquad (10)$$

where, λ_0, λ_1 is the reference and calibration wavelength, and ψ is a constant. Since $m\tau_a(\lambda_0)$ is well calibrated, $\ln F_0(\lambda_1)$ can be obtained by regression analysis with the left side of (10) and $m\tau_a(\lambda_0)$.

The proposed calibration method calibrates based on the ILM in the long wavelength region where the calibration accuracy is relatively high and uses the result as the reference wavelength calibration coefficient to obtain calibration coefficients of other wavelengths. At this time, the Skyradpack software code [11] was used when ILM was applied. The flow of the proposed calibration method is shown in Fig. 4.

First, based on Skyradpack ver. 4.2, VSA is obtained by using measurement data by Sky-radiometer, and aerosol optical thickness and volume spectrum are calculated. At this time, the calibration of the sky-radiometer is performed by ILM. Next, the volume spectrum is reanalyzed using the direct sunlight and the peripheral light radiance to re-calculate the aerosol optical thickness, the volume spectrum, the complex refractive index and the like with higher accuracy. Calibration is performed at other wavelengths based on the calibration coefficients at various amounts and reference wavelengths obtained by these reanalyzes. These are performed in two stages of levels 0 and 1 shown below.

Level 0: Estimate volume spectrum based on VSA. The aerosol optical thickness is obtained using the sunlight directly and peripheral measurement data. Calibrate F_0 from the plot until *lnF-m* τ *a* becomes zero.

Level 1: Re-analyze the sunlight and marginal light, volume spectrum, update the VSA, and recalculate the phase function and volume spectrum.

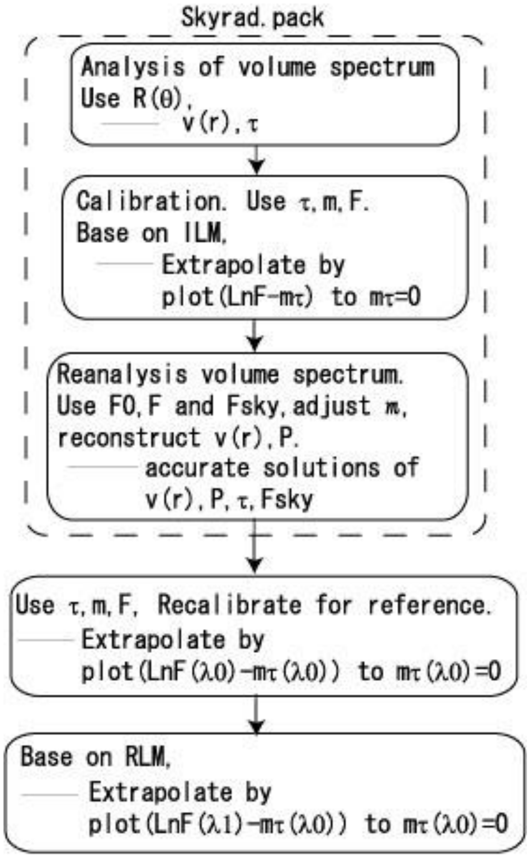

Skyrad.pack
Analysis of volume spectrum
Use R(θ),
v(r), τ
Calibration. Use τ, m, F.
Base on ILM,
Extrapolate by
plot(LnF-mτ) to mτ=0
Reanalysis volume spectrum.
Use F0, F and Fsky, adjust m,
reconstruct v(r), P.
accurate solutions of
v(r), P, τ, Fsky
Use τ, m, F, Recalibrate for reference.
Extrapolate by
plot(LnF(λ0)-mτ(λ0)) to mτ(λ0)=0
Base on RLM,
Extrapolate by
plot(LnF(λ1)-mτ(λ0)) to mτ(λ0)=0

Fig. 4. The algorithm of multi stage calibration method.

III. Simulation

Simulation data was generated based on Skyradpack ver.4.2 mentioned above. The wavelength to be used is the set wavelength of the Sky-radiometer POM-1 manufactured by PREDE Co. Ltd., 0.4, 0.5, 0.675, 0.87 and 1.02 µm. The reference wavelength was set to 0.87 µm which is a long wavelength. In addition, the lognormal distribution of (11) was assumed for the particle size distribution of aerosol.

$$n(\ln r)=\sum_{i=1}^{2}\frac{C_i}{\sqrt{2\pi}\log\sigma_i}\exp\left(-\frac{(\log r-\log\overline{r_i})^2}{2\log^2\sigma_i}\right) \qquad (11)$$

where *n(lnr)dlnr* is the number density of aerosol particles between particle size *r* and *r+dlnr*. Also C_i set to 1, the standard deviation and average σ_i, $\overline{r_i}$ of the particle size distribution was set to be the same as the aerosol type measured at Saga University in 2003. These amounts are shown in Table I.

TABLE I. THE PARAMETERS FOR LOG-NORMAL DISTRIBUTION

No. Mode	C_i	r_i(μm)	σ_i
1	1	0.37	1.95
2	1	3.06	2.36

That is, the particle size distribution is a bimodal characteristic (bimodal), with the first mode appearing at 0.37 μm and the second mode at 3.06 μm. Furthermore, the aerosol complex refractive index is set to m=1.50-0.01i, and the radiance of the sun outside the atmosphere is set to 1. The temporal variation of the aerosol optical thickness was assumed to conform to (12) [16].

$$\tau_a = \tau_{a0}(1+\alpha t^2) \qquad (12)$$

where, τ_{a0} is the aerosol optical thickness at the south-middle time, and they were set to 0.1, 0.2 and 0.3 by numerical simulation. Also, α is assumed to be 0.011. Thus, the aerosol optical thickness will vary from 0 to 20%, whereas the atmospheric mass varies between 1.5 and 4.5. An error (random noise) of ± 3% and ± 5% was intentionally superimposed on the measured value of the Sky-radiometer. Simulation results are shown in Fig. 5. In the figure, the wavelengths are limited to 0.4, 0.5 and 0.87 μm to avoid expression complexity. These are the ILM, VSA (L_0 in the figure) and the proposed method, i.e. the estimation error of the aerosol optical thickness (L_1 in the figure) obtained by reanalysis of the volume spectrum. At this time, the total atmospheric optical thickness was set to 0.1, 0.2, 0.3. Measurement errors of 0%, ± 3% and ± 5% were randomly superimposed on the radiance measured by the Sky-radiometer.

Table II shows the maximum of the estimation error of aerosol optical thickness by ILM, that is VSA. From this table the accuracy of estimating the aerosol optical thickness of ILM is more sensitive to the measurement error superimposed on the sky-radiometer than the optical thickness of the whole atmosphere.

TABLE II. MAXIMUM ERROR IN AEROSOL OPTICAL DEPTH ESTIMATION WITH THE METHODS BY (A) VSA AND BY (B) REANALYSIS OF VOLUME SPECTRUM AS THE OPTICAL DEPTH OF 0.1, 0.2 AND 0.3

		Optical Depth	0.1			0.2			0.3		
Error(%)	Method	Wavelength(nm)	400	500	870	400	500	870	400	500	870
0	LM		0.025	0.015	0.007	0.042	0.025	0.009	0.042	0.024	0.009
±3			0.09	0.05	0.028	0.063	0.036	0.022	0.057	0.033	0.026
±5			0.14	0.07	0.045	0.1	0.045	0.038	0.081	0.038	0.038
0	PM		-0.015	-0.011	-0.003	-0.013	-0.01	-0.003	0.042	0.033	0.013
±3			0.058	0.025	0.024	0.045	0.02	0.025	0.044	0.02	0.014
±5			0.058	-0.038	-0.02	-0.042	-0.03	-0.02	-0.038	-0.032	-0.021

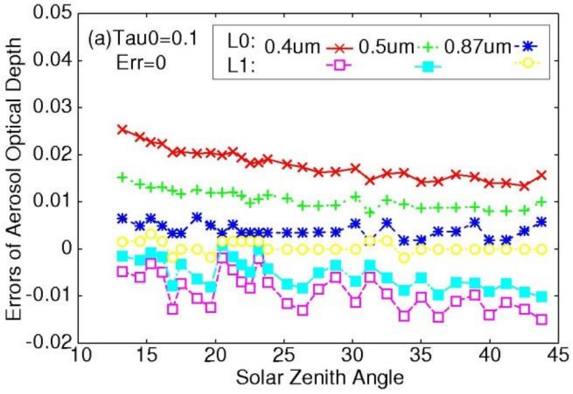
Table II and Fig. 5 also show that

a) estimation of aerosol optical thickness by reanalysis of volume spectrum is more accurate than by VSA,

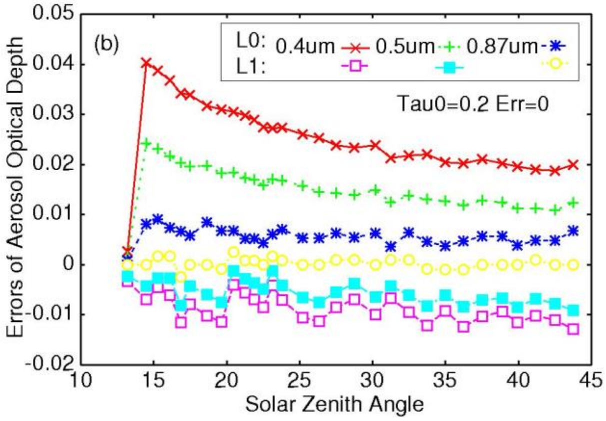
b) estimation of aerosol optical thickness at 0.87 μm

c) the influence of measurement error depends largely on the case of VSA and not sensitive for volume spectrum reanalysis,

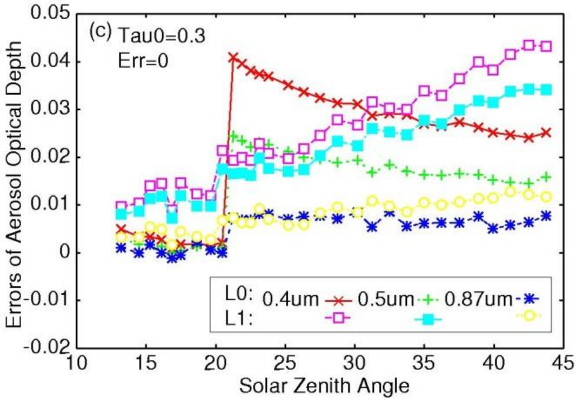
d) optical thickness: The same is true for the influence on VSA, in the case of VSA, the estimation error of the aerosol optical thickness depends on the optical thickness, but in the case of reanalysis of the volume spectrum the dependence is small. Aureole is discussed in the previous paper [17].



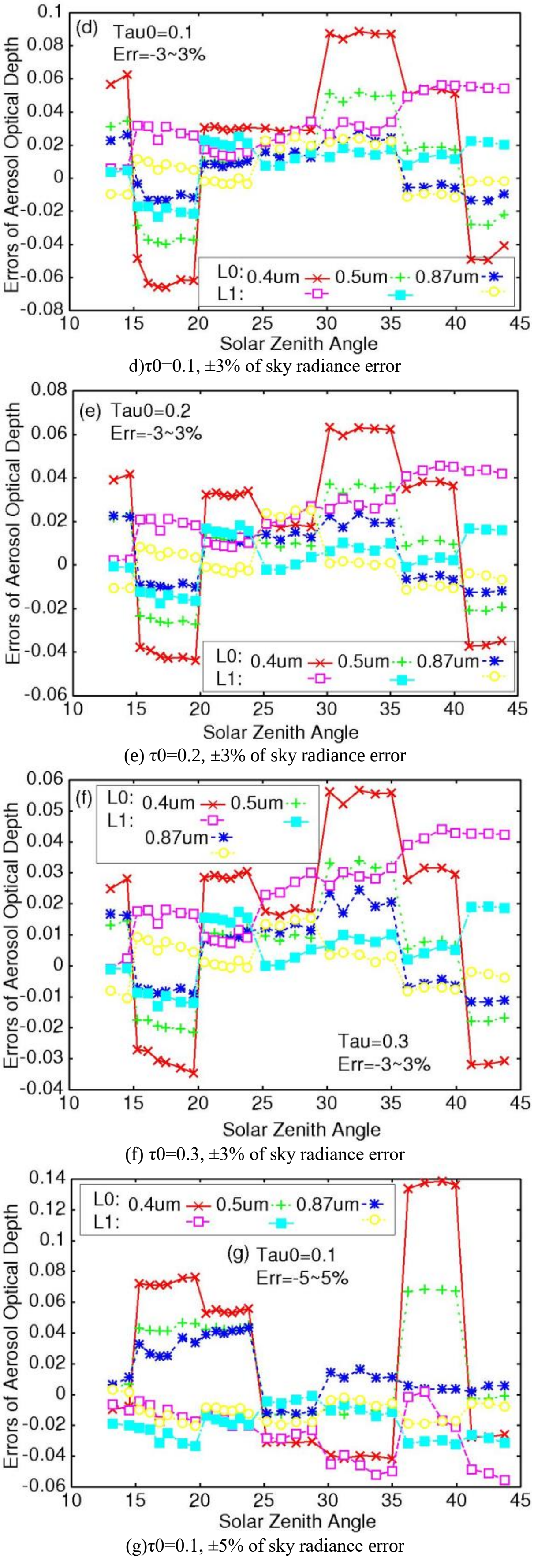
(a) τ0=0.1, no sky radiance error



(b) τ0=0.2, no sky radiance error



(c) τ0=0.3, no sky radiance error

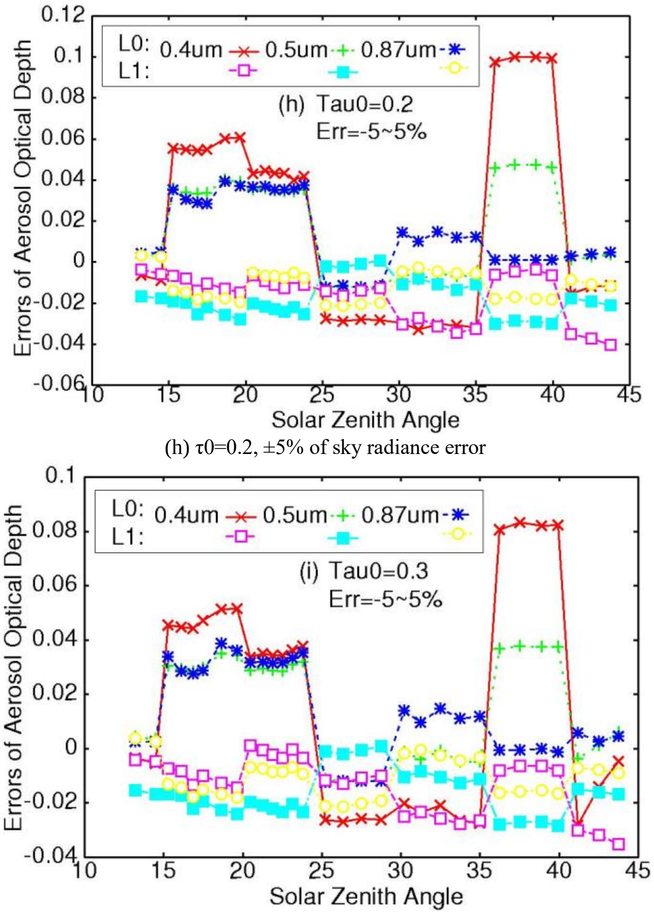


d)τ0=0.1, ±3% of sky radiance error

(e) τ0=0.2, ±3% of sky radiance error

(f) τ0=0.3, ±3% of sky radiance error

(g)τ0=0.1, ±5% of sky radiance error



(h) τ0=0.2, ±5% of sky radiance error

(i) τ0=0.3, ±5% of sky radiance error

Fig. 5. The estimation errors of aerosol optical depth with VSA and reanalysis of volume spectrum for the wavelengths 0.4, 0.5 and 0.87um with 0, ±3% and ±5% of random noise in sky radiance measurements.

Table III compares the calibration coefficient estimation accuracy by ILM and the proposed method for five wavelengths. The calibration accuracy of the proposed method exceeds ILM especially at short wavelength (0.4 μm). Also, as the optical thickness increases, the more the noise superimposed on the measurement of the sky-radiometer, the greater the difference in calibration coefficient accuracy between ILM and the proposed method.

TABLE III. COMPARISON OF ESTIMATION ERROR FOR CALIBRATION BETWEEN ILM (USE VSA) AND THE PROPOSED METHOD (PM: USE REANALYSIS OF VOLUME SPECTRUM) AS THE OPTICAL DEPTH ARE 0.1, 0.2 AND 0.3

(a)0% Error

Aerosol Optical Depth	0.1		0.2		0.3	
Wavelength(nm)	LM	PM	LM	PM	LM	PM
400	0.0008	0.0006	0.0029	0.0009	0.013	0.0014
500	0.0003	0.0006	0.0015	0.0006	0.01	0.0009
675	0.0012	0.0005	0.0006	0.0006	0.005	0.0005
870	0.0002	0.0002	0.0001	0.0001	0.003	0.0004
1020	0.0002	0.0002	0.0001	0.0001	0.002	0.0004

(b)±3% Error

Aerosol Optical Depth	0.1		0.2		0.3	
Wavelength(nm)	LM	PM	LM	PM	LM	PM
400	0.011	0.004	0.017	0.009	0.023	0.011
500	0.008	0.003	0.009	0.006	0.012	0.009
675	0.003	0.002	0.003	0.002	0.015	0.007
870	0.006	0.002	0.001	0.001	0.002	0.001
1020	0.002	0.001	0.001	0.002	0.001	0.001

(c)±5% Error

Aerosol Optical Depth	0.1		0.2		0.3	
Wavelength(nm)	LM	PM	LM	PM	LM	PM
400	0.013	0.007	0.018	0.005	0.027	0.014
500	0.006	0.006	0.006	0.004	0.011	0.01
675	0.003	0.003	0.003	0.003	0.007	0.005
870	0.001	0.001	0.002	0.001	0.001	0.001
1020	0.001	0.001	0.002	0.001	0.001	0.002

In order to investigate the influence of particle size distribution on calibration accuracy, the average and standard deviation of the relative particle size distribution of equation (10) were varied by ± 3% and ± 5%. An example of the particle size distribution recalculated by reanalysis is shown in Fig. 6.

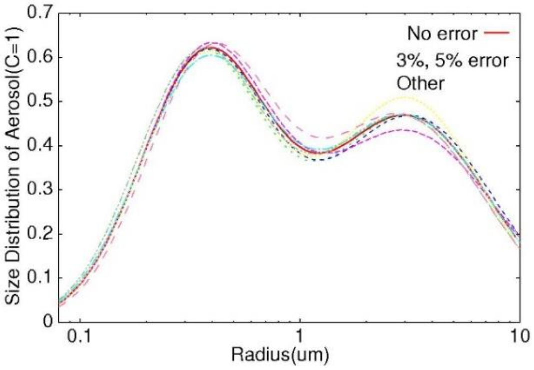
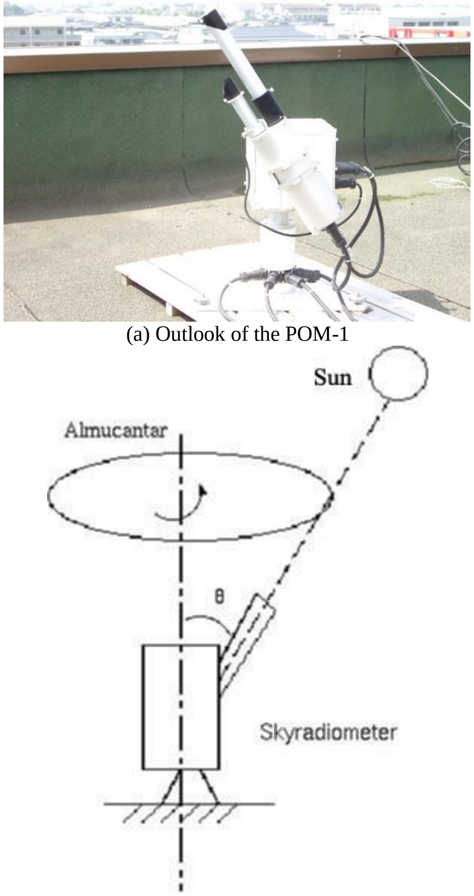


Fig. 6. Size distribution error due to ±3% and ±5% of sky radiance measurement errors.

From this figure, it can be seen that the measurement error of the standard deviation of 3, 5% appears as a difference in volume spectra of 0.61 ± 3.3% at 0.37 µm in the first mode and 0.47 ± 8.5% at 3.06 µm in the second mode. Conversely, calibration accuracy is improved by reflecting this amount on the aerosol optical thickness by this amount.

IV. Experiment

Direct scattering, scattering and marginal light were measured using the sky-radiometer POM - 1 shown in Fig. 7. I am continuing the measurement from September 2003 to the present, but here it is assumed that 15 data sets measured under fine weather conditions (on 16, 17, 23 November 2003, 03, 04, 24 December 2003, 07, 08, 31 January 2004, 20 February 2004, 15 March 2004, 05, 22, 24, 25 April 2004).



(a) Outlook of the POM-1

(b) Observation scheme

Fig. 7. Observation scheme (Almucantar observation for direct, diffuse and aureole measurements) and outlook of the sky-radiometer, POM-1 which is put on the top of the building of the Saga University at 130°29’E, 33°25’N.

The effectiveness of the proposed method was verified. Here, the optical thickness of the atmosphere was 0.3 or less at 0.5 µm. Also, the ratio of the optical thickness ratio at two different wavelengths was within 5%.

The calibration result at the reference wavelength of 0.87 µm by ILM is shown in Fig. 8.

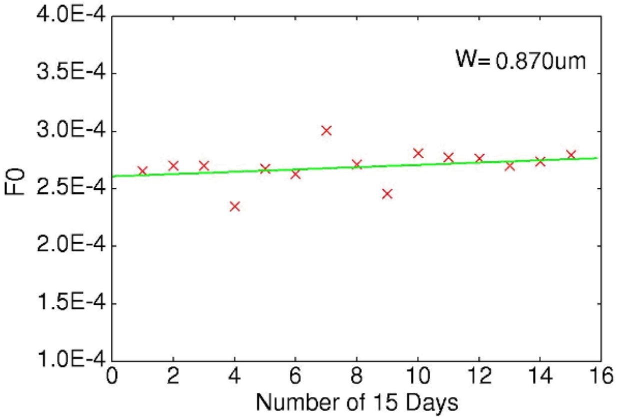


Fig. 8. F_0 measured with Improved Langley method for 15 days of measured solar direct, diffuse and aureole.

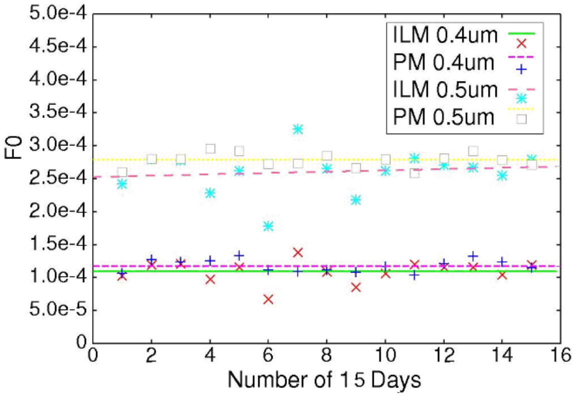


Fig. 9. Comparison between ILM and the proposed method (PM) at bands 0.4 and 0.5μm for 15 days of measured solar direct, diffuse and aureole.

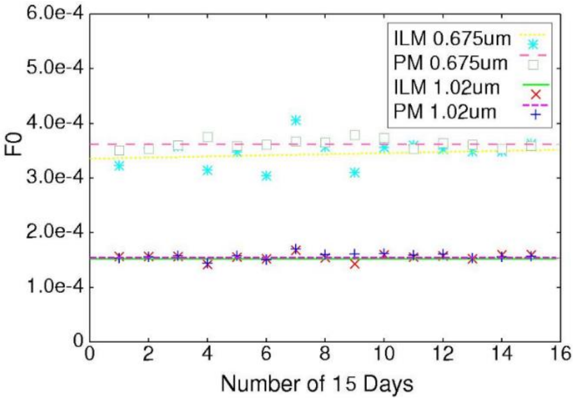


Fig. 10. Comparison between ILM and the proposed method (PM) at bands 0.675 and 1.02μm for 15 days of measured solar direct, diffuse and aureole.

As is apparent from the figure, the standard deviation of the calibration value of F_0 is within 1%, indicating that the accuracy is extremely high. Calibration results of ILM and the proposed method are shown in Fig. 9 (0.4, 0.5μm) and Fig. 10 (0.675, 1.02μm).

As is evident from these figures, it can be demonstrated that the proposed method has higher calibration accuracy than ILM. Whether the calibration accuracy is good or bad is evaluated based on the standard deviation of the calibration value of F_0 is shown in Table IV, and it was confirmed that the calibration accuracy improvement ranged from 8.22% to 47.75%. The effect of improving the calibration accuracy is high in the short wavelength region as can be understood from the principle. From this, it can be said that calibration by the proposed method requires less calibration frequency for the same calibration accuracy requirement.

TABLE IV. COMPARISON OF THE STANDARD DEVIATIONS BETWEEN ILM AND PM

Wavelength(μm)	**Standard Deviation**		**% Improve**
	ILM	**PM**	**Ratio**
0.4	0.03858	0.02016	47.745
0.5	0.02219	0.01691	23.795
0.675	0.01837	0.01295	29.505
1.02	0.01022	0.00938	8.219

V. CONCLUSION

Improved Langley method (ILM): In order to improve the calibration accuracy of the Sky-radiometer by ILM, attention is paid to the fact that the ILM has a high calibration accuracy at a relatively long wavelength. Rati Langley method (RLM): RLM calculates the calibration coefficient at another reference wavelength I proposed a calibration method that improves calibration accuracy in all wavelength bands by utilizing the fact that the calibration coefficient of an arbitrary wavelength can be estimated. Specifically, calibration coefficients of other wavelengths were estimated from the calibration coefficient by ILM at 0.87 μm by the RLM method. The calibration error of the proposed method was intentionally evaluated by numerical simulation based on actual measurement data when the measurement error of ± 3% and ± 5% was superimposed on the measurement value of the marginal light, and the maximum of the error was 0.0014 and 0.0428, and those of ILM were 0.011 and 0.0489. Therefore, the proposed calibration method is more robust to measurement error than ILM, and it was found that highly accurate calibration is possible over all wavelengths

The accuracy of the proofreading method was evaluated based on the standard deviation of the calibration coefficient based on the measured data of the Sky-radiometer for 15 days, which is suitable for calibration, out of the measured data over 4 years, it was 0.02016, and the calibration coefficient by ILM. The standard deviation of 0.03858 of the proof calibration method was confirmed.

Further experimental study is required for validation of the proposed method.

ACKNOWLEDGMENT

The author would like to thank Dr. Teruyuki Nakajima of JAXA and Dr. Xing Ming Liang of NOAA/NESDIS for their valuable comments and suggestions. Also, author would like to thank the fourth group members of the Department of Information Science, Saga University for their contribution to the experiments.

REFERENCES

[1] Kohei Arai, Akira Makoto, Simultaneous estimation of the complex refractive index and particle size distribution of aerosol using scattering and marginal light by simulated annealing, Journal of Japan Remote Sensing Society, Vol. 23, No. 1, .11-20, 2003.

[2] Shaw, G.E., Solar spectral irradiance and atmospheric transmission at Mauna Loa Observatory. Appl. Opt., 21, 2007_2011. 1982.

[3] Holben, B.N., and Coauthors, AERONET-A federated instrument network and data archive for aerosol characterization. Remote Sens. Environ., 66, 1_16. 1998.

[4] McArthur, L.J.B., D.H. Halliwell, O.J. Neibergall, N.T. O'Neill, J.R. Slusser, and C. Wehrli, Field comparison of network sun photometers., J. Geophys. Res., 108, 4596, doi:10.1029/2002JD002964. 2003.

[5] Kohei Arai, Self-Studied Remote Sensing, Mori Kita Publication 2004

[6] Langley, S.P., The bolometer and radiant energy. Proc. Amer. Acad. Arts Sci., 16, 342. 1881.

[7] Schmid, B., and C. Wehrli, Comparison of sun photometer calibration by use of the Langley technique and the standard lamp. Appl. Opt., 34, 4500_4512. 1995.

[8] Slusser, J.R., J.H. Gibson, D.S. Bigelow, D. Kolinski, P. Dister-hoft, K. Lantz, and A. Beaubien, Langley method of calibrating UV filter radiometers. J. Geophys. Res., 105, 4841_4849. 2000.

[9] Adler-Golden, S.M., J.R.Slusser, Comparison of Plotting Methods for Solar Radiometer Calibration J. Atmospheric and Oceanic Technology, 24, 935-938, American Meteorological Society, 2007.

[10] Forgan, B.W., General method for calibrating sun photometers. Appl. Opt., 33, 4841-4850, 1994.

[11] Namajima, T., Tonna, G., Rao, R. et al. Use of sky brightness measurements from ground for remote sensing of particulate polydispersions. Appl. Opt. 35, 2672-2686, 1996.

[12] Tanaka, M., Nakajima, T., Shiobara, M., Calibration of a sunphotometer by simultaneous measurements of direct-solar and circumsolar radiations. Appl. Opt., 25, 1170-1176, 1986.

[13] Schotland, R.M., Lea, T.K., Bais in a solar constant determination by the Langley method due to structured atmospheric aerosol. Appl. Opt., 25, 2486-2491, 1986.

[14] Harrison, L., J. Michalsky, and J. Berndt, Automated multi-filter rotating shadowband radiometer: An instrument for optical depth and radiation measurements. Appl. Opt., 33, 5118_5125. 1994.

[15] Rochford, P.A., and Coauthors, Validation and refinement of hyperspectral/multispectral atmospheric compensation using shadowband radiometers. IEEE Trans. Geosci. Remote Sens., 43, 2898_2907. 2005.

[16] Shaw, G.E., Error analysis of multi-wavelength sunphotometry. Pure Appl. Geophys., 114, 1, 1976.

[17] Kohei Arai, Method for aureole estimation refinement through comparisons between observed aureole and estimated aureole based on Monte Carlo Ray Tracing, International Journal of Advanced Research in Artificial Intelligence, 2, 12, 1-8, 2013.

AUTHOR'S 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 counselor 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 Science Commission "A" of ICSU/COSPAR since 2008 then he is now award committee member of ICSU/COSPAR. He wrote 37 books and published 570 journal papers. He received 30 of awards including ICSU/COSPAR Vikram Sarabhai Medal in 2016, and Science award of Ministry of Mister of Education of Japan in 2015. He is now Editor-in-Chief of IJACSA and IJISA. http://teagis.ip.is.saga-u.ac.jp/index.ht

Wireless Sensor Network and Internet of Things in Precision Agriculture

Farzad Kiani, Amir Seyyedabbasi
Computer Engineering Dept., Engineering and Natural Sciences Faculty,
Istanbul Sabahattin Zaim University, Istanbul, Turkey

***Abstract*—Internet of Things is one of the most popular subjects nowadays where sensors and smart devices facilitate the provision of information and communication. In IoT, one of the main concepts is wireless sensor networks in which data is collected from all the sensors in a network characterized by low power consumption and a wide range of communication. In this study, an architecture to monitor soil moisture, temperature and humidity on small farms is provided. The main motivation for this study is to decrease water consumption whilst increasing productivity on small agricultural farms and precisions on them. This motivation is further propelled by the fact that agriculture is the backbone of some towns and most villages in most of the countries. Furthermore, some countries depend on farming as the main source of income. Putting the above-mentioned factors into consideration, the farm is divided into regions; the proposed system monitors soil moisture, humidity and temperature in the respective regions using wireless sensor networks, internet of things and sends a report to the end user. The report contains, as part of the information, a 10-day weather forecast. We believe that with the above information, the end user (farmer) can more efficiently schedule farm cultivation, harvesting, irrigation, and fertilization.**

Keywords—Wireless sensor network; internet of things; smart agriculture applications; precision agriculture

I. INTRODUCTION

Owing to the increment in population and corresponding decrement in rainfall amount, there is a substantial scarcity of food and water – which are the most basic needs of life. Hence, the importance of precision agriculture [1] has become more pronounced leading to a plethora of researchers being conducted in the field over the recent decades. In most countries, the family economy depends on agriculture, thus, principled and productive agriculture is of paramount importance to them. Unfortunately, in the most countries and regions where farming is done on small farms, primitive methods are still widely being used while in developed countries; statistics show that the use of modern agriculture is on the rise. Modern agriculture can be expressed with this view; reduce agricultural costs and increase productivity.

The main target of this study is small farms, in other words, farms capable of planting several types of products, such as vegetables and fruits in a small area. For instance, consider the cultivation of eggplant, parsley, pepper or tomato on a farm, each of these needs a different irrigation method and scheduling. Thus, the main goal of this study is increasing production with minimal costs and, ultimately, increase net revenue. Technology supported agriculture with concepts of wireless sensor networks (WSN) and the Internet of Things (IoT) can help to gradually shift from primitive agricultural practices. Studies on industrial agriculture ranging from climate information, soil information to intelligent irrigation was based on data collection performed by wireless sensors. The wireless system enhances crop productivity and convenience [2]. WSNs collect information from different sensors in large and small networks so end users can get and process data. These networks can be used in the monitoring of people health, weather conditions, control traffic, and air pollution.

IoT is an environment where objects, animals or people are equipped with unique identifiers capable of data transmission over the internet without the need for human or computer interactions [3]. In this paper, we effort to simulate IOT concept using different sensors. The information collected by the sensors is sent to the server with RASPBERRY PI 3 and WI-FI module that can be understood in a graphical environment (GUI). As days go by, the number of applications in this field increases and many researchers have been focusing on a new idea based on IoT technology concept. One of the popular research areas is agriculture and smart farms. However, they are not comprehensive methods. The most researches are suffered from compatibility and connection between heterogeneous devices [4]. Currently, the farmer can collect data as temperature, air humidity, soil humidity, volumetric water content unit (VWC) and gravimetric water content (GWC) from the different regions of the farmed area thanks to this technology. After collecting the required data, they can be analyzed and give flexible plans to its users in order to different intentions such as cultivation, harvest, irrigation, fertilization. It is also important to note that the collected data by digital sensors must be understandable to the end user.

In our proposed system, we collect different data via sensor nodes. In addition, we use some current information such as weather conditions because analysis of the system that takes external factors into its own parameters is more reliable.

The remainder of the paper is organized as follows. Section 2 is about related works of precision agriculture, Section 3 describes our proposed system architecture and abilities. Finally, Section 4 presents conclusions of the work-study and future works.

II. Related Works

In this section, we focus on how to use the IoT and smart technologies in the agriculture-based applications.

In [2] has proposed a smart wireless sensor network so the authors design an architecture to data collection from nodes in an agriculture environment. They analyze collected data and display their results to its end users. In [5] has presented a method based on wireless sensor networks in potato farming that it monitors and understands individual crop and requirements. Therefore, the farmers can potentially identify the various fertilizers, irrigation, and other requirements. The authors propose an irrigation management model to estimate agricultural parameters using mathematical calculations and intelligent humidity sensors. Devices used for monitoring are laptop computers or PDA. In [6] has introduced a smart system based on wireless sensor network in a red bayberry greenhouse using soil moisture and temperature sensors. This system can collect the temperature, humidity, illumination, and voltage of the greenhouse.

GPRS gateway has been used for transmitting data to monitor the system. The energy of sensors is provided by solar and storage batteries. In [7] has proposed a novel platform to establish energy efficient wireless sensor network in a sugar farm. It provides the energy efficiency by the solar system. In [8] has proposed a method to make agriculture smart using automation and IoT technologies. In this paper, the authors use ZigBee models, camera and actuators to handle smart irrigation on accurate real-time field data. In [9] has focused on automated irrigation using the wireless moisture sensor network and IoT technology in order to smart precision agriculture system. Irrigation by schedule is to supply water to the plant at specific times and automated irrigation by feedback based on temperature, humidity and moisture sensors. Two irrigation methods are compared to the second method has better efficiency in water consumption.

As seen in many studies in the literature, wireless sensor networks are a structure that works nested with IoT technology. Therefore, let us talk a bit about this network.

A. Wireless Sensor Networks

The wireless sensor networks (WSNs) consist of multiple many sensor nodes in a wireless communication-based environment. The sensor node is to detect physical phenomena such as temperature, humidity, and moisture with limited energy and memory [10]. Many current researchers are focused on these limitations by various methods as different routing protocols [11], [14], intelligent based approaches [12], [13] etc. They have many applications area in various environments such as military, medicine, education, agriculture, monitoring systems etc. It is expected that it will be more workspaces by emerging the IoT concept. Fig. 1 shows a sample architecture of a sensor node.

The main components of a sensor node are power unit, computing unit, sensing unit and Communication unit.

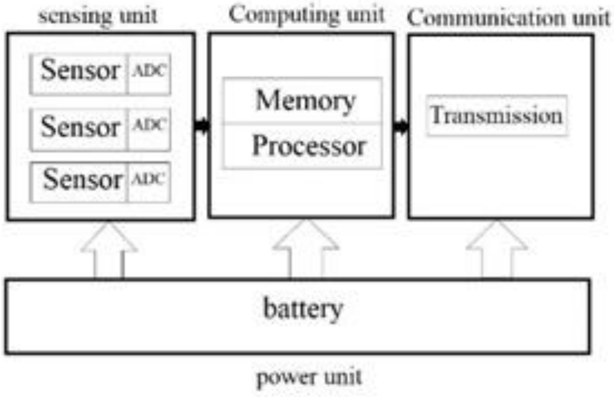


Fig. 1. Node architecture.

The nodes sense a physical event of the phenomenon and convert them to digital signals by sensing unit. Then they handle collected data and process them by computing unit. This unit is consist of processing and memory subunits. They can store necessary data in their storage. In addition, they support necessary their operations from a few hours to months or years via power unit. The sensor nodes are also able to communicate together or server/base station as centralized or decentralized architectures based on various topologies such as mesh, star etc. This is realized via the communication unit. The sensor nodes can communicate together in the point-to-point or multi-hop models [14].

These are the deepest differences with the classic sensor nodes. These nodes furthermore widely deployed in 1m up to 100meters area, low communication bandwidth, limited memory and computation power.

The characteristics of wireless sensor nodes are ease of use, scalability to the large scale of deployment, the mobility of nodes and resilience. Since these networks are interested in information regarding the physical phenomenon instead of information from a single sensor, the failure of a single node should not affect the overall operation of the network. Nevertheless, the fault tolerances is an important design factor in the networks [15]. Need to be aware of other factor is cost. The most basic features of these networks are cheap. Therefore, they are almost used as low-cost sensor nodes. Researchers are introduced new sensor devices with different software and hardware properties. In this paper, we use our designed sensor node. The other factor is energy and network lifetime. The sensor nodes have low battery level (4 Joule or less). Various methods such as optimized routing protocols, topology control, and power management protocols are followed to achieve this aim.

B. Internet of Things

Internet of Things (IoT) is one of the new topics that has been discussed recently that usage areas are growing rapidly. They are consist of many nodes that are equipped with internet output. The existing technologies such as ad-hoc systems, pervasive and embedded systems, wearable technology and machine learning techniques are founded new concept by emerging of IoT. One of the most important requirements in this technology is connection to internet problem [16]. Internet one of the inseparable things in our life that can reduce cost and time. Imagine yourself doing an internet search for your watch you lost somewhere in your house. Therefore, this is the main vision of IoT, an environment where things are able to talk and their data can be processed to perform desired tasks through machine learning techniques [6], [13]. The IoT is built around wireless radio waves that allow to different devices to communicate with each other through the Internet. This platform includes some standards such as Wi-Fi, low-power Bluetooth, NFC, RFID, and so on. Physical object in IoT collects and process data that it can receive it from the environment and/or other objects [17]. These objects are embedded with sensors, actuators, internet output and network connectivity. Truth, not all the devices need to be the connected internet. This property can be increased the cost of the system. In this case, a system designer can use the WSN structure in order to reach to a smart application.

Environmental and personal health monitoring, monitoring and control of industrial processes including agriculture, smart spaces, and smart cities are just some of the examples of IoT applications [3]. According to Fig. 2, it is estimated that by 2025 approximately 75.44 billion devices will be connected to the internet [18]. One of the IoT trends is in agriculture. In agriculture applications based on IoT, the objects communicate together to provide useful information from the farm or greenhouse. In these types of systems, some actual devices do some required operations such as irrigation and prune. In other words, the internet and physical agents have been effective in reducing the human factor involvement, to increase productivity, as well as reduce costs.

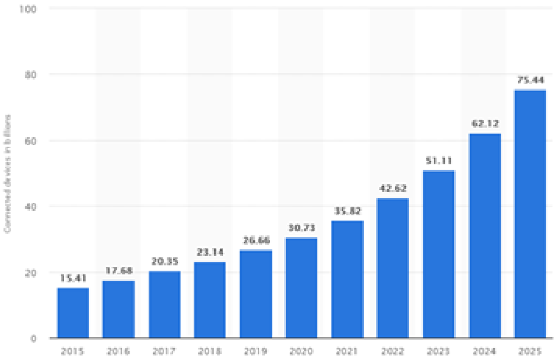


Fig. 2. Internet of Things (IoT) connected devices.

III. Proposed Systems

In this section, the proposed system architecture is introduced. The used sensor types in the system are humidity, temperatures, and soil moisture. In this architecture, the sensor-based devices are deployed in the environment to sense and data collection. The related system has more benefit for farmers so he/she can manage his/her time, energy and costs. Data will be collected by system nodes and it transmits to server finally via the cross from probable other nodes and gateways. The all collected data in server side must be analyzed and be presented to the users in a user-friendly platform. SIXFAB Company and IZU-WSN Research Lab. designed a custom board that is called as KIANI sensor nodes (Fig. 3). This node is used in this case study.

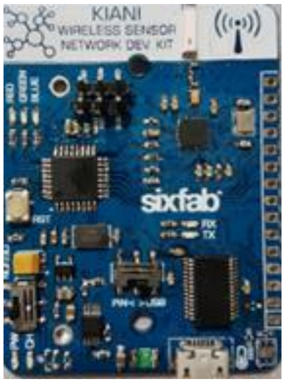


Fig. 3. Our jointly produced custom sensor node.

The sensing unit of the node has SparkFun Soil Moisture Sensor. Furthermore, the board has temperature and humidity sensors that sense different air humidity and temperature. In communication unit, our board transceiver is based on Texas instruments cc1101 Low-Power Sub-1GHz RF Transceiver Computing unit equips with Arduino Nano, ATmega328P used in computing unit. In power unit, we used 1200mah 3.7v Li-ion rechargeable battery. In Second part of the architecture, we use RASPBERRY PI 3 as a gateway to send collected data from sensors to sever for processing and presentation information that users need to it. At last, related information such as temperature, humidity, soil moisture and future air conditions are shown as an application in GUI that can be available in the related website and mobile application. Fig. 4 shows a sample model of the system and node in the agricultural land.

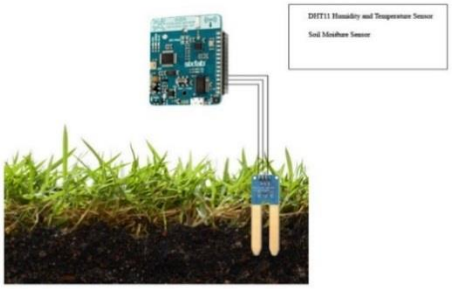


Fig. 4. Sensors based custom nodes in the farm.

Today, the water issue is one of the important problems in our life. We need to avoid the litter and we must use this source correctly, especially in agriculture lands.

On the other hand, the irrigation system has traditionally used in farms and water comes to the point of water entry into the farms. In result, the water has come to the total agricultural land over a given period. To cover the total farm surface, it also depends on the surface of the earth and type of soil. Traditionally, created waterways in the farms could divide the entrance water at least equal. In our proposed system, we consider a farm with 100*100 meters that farmers divided this area into four equal regions as is shown in Fig. 5. Then, in each region, our nodes have been deployed. Data is obtained in every one hour from devices, which forwards it to the gateway. Where it is stored and then transferred to end users through API.

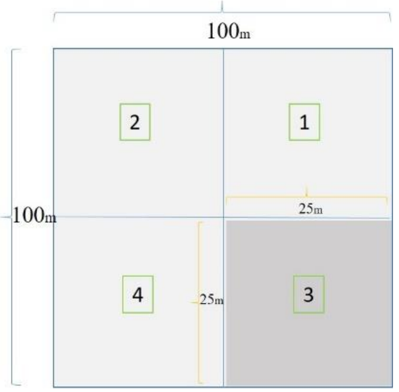
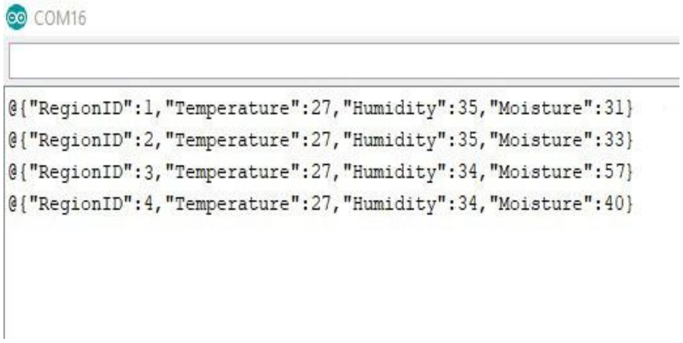


Fig. 5. The farm structure that is divided into 4 equal clusters.

After this, an application will be processed and data will be shown to end user. The working mechanism of the proposed system is presented in Fig. 6. It shows related information in every region in one-hour periodic periods, continuously. The results of the system are presented to end user via a website or mobile application so is shown in Fig. 7. The proposed application uses the current some information as weather condition besides the information that receives from our nodes.

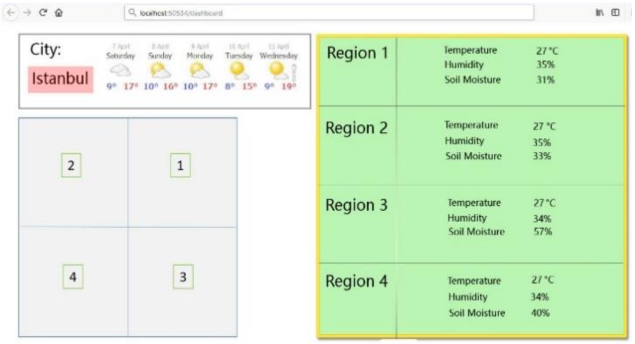
The end user with this comprehensive information will be prepared fields of farms for planting, harvesting, fertilizing and irritating. This general view of the farm, the farmer will be aware of the accurate understanding of land changes, which previously was traditionally used and experimental knowledge was provided. In such systems, technology will provide useful information at the lowest cost. Launching the system may be costly for the farmer for the first time, but over time, with the increase in productivity, this cost will be offset. The work is done automatically and farmer monitors self-land makes some decisions and applies them through automatic devices that are in the farm or near to our region.



```
@{"RegionID":1,"Temperature":27,"Humidity":35,"Moisture":31}
@{"RegionID":2,"Temperature":27,"Humidity":35,"Moisture":33}
@{"RegionID":3,"Temperature":27,"Humidity":34,"Moisture":57}
@{"RegionID":4,"Temperature":27,"Humidity":34,"Moisture":40}
```

Fig. 6. Data comes from Arduino.

This system provides a great convenience to farmers and they can have more and more efficient products, efficient use of water and management of energy, time and unnecessary classic costs. All of them would be one of the reasons why farmers would be welcome. The implemented GUI has a user-friendly interface so the users can benefit from all means. In this system, we tried to have a simple and understandable graphic user interface. The system will update the information hourly, as well as provide a history of the information from every region.



Region 1 Temperature 27 °C Humidity 35% Soil Moisture 31%

Region 2 Temperature 27 °C Humidity 35% Soil Moisture 33%

Region 3 Temperature 27 °C Humidity 34% Soil Moisture 57%

Region 4 Temperature 27 °C Humidity 34% Soil Moisture 40%

Fig. 7. GUI snapshot.

IV. Results and Evaluation

IoT's importance is increasing day by day in our lives. There are a variety of applications in this respect. Because these methods are inspired by WSN technology, they can also be used to differentiate between different algorithms and software methods. They can be more usable in various applications so one of them is agriculture area. In today's world, people are beginning to make use of intelligent devices and system thus; they have a smart assistant thanks these systems. In this paper, an architecture has been proposed so it helps to farmers to manage the irrigation time of their agriculture correctly. The result of it is shown it is efficient in the resource consumption. The resulting of the system that is an application can be of great benefit to its users. The user can directly divide own land into as many regions as desired in this application. Therefore, the users will provide saving in their time and water. In addition, this application is reliable because it provides weather information instantaneously from the central stations.

V. Conclusion and Future Works

In general, water issue and irrigation methods play an important role in efficient water using and increase productivity. So, water consumption reduction that helps farmers economic at the small farms. Furthermore, farmer's information about weather conditions of next days can help to make decisions that are more accurate. In this project, we used the IoT and WSN enables to achieve this goal. In addition to, fertilization, harvesting and cultivation are important as irrigation too. Therefore, with this method, the farmer can schedule his/her next upcoming activities. About future work, we will use a reinforcement learning based system in data collection and processing phases on the farms to give suggestions that are more useful to farmers about normal activities in farms. In addition, some data can be collected from self-farmers.

As future works can design different hybrid architectures for various IoT based applications. It can be realized by software and hardware-based approaches. In addition, the machine learning-based methods such as reinforcement learning, game theory, fuzzy logic and neural networks can be suggested. On the other hand, the energy efficiency issue can be investigated from different perspectives such as routing and communication between devices based on shortest path finding methods.

Acknowledgment

This project has been achieved in Wireless Sensor Networks and IoT Laboratory in Istanbul Sabahattin Zaim University (IZU). Many thanks to our colleagues at this laboratory and SixFab Co. that help us in this project.

References

[1] Z. Naiqian, M. Wang and N. Wang. “Precision agriculture-a worldwide overview.” Computers and Electronics in Agriculture 36.2-3, pp. 113-132, 2002.

[2] G. Mendez, M.A. Yunus and S.C. Mukhopadhyay. “A WiFi-based smart wireless sensor network for monitoring an agricultural environment”. IEEE Conference in Instrumentation and Measurement Technology Conference (I2MTC), pp. 2640-2645, 2012.

[3] A. Gluhak, S. Krco, M. Nati. “A survey of facilities for the experimental internet of things research”. IEEE Communications Magazine, 49(11), pp.58-67, 2011.

[4] T. Ojha, S. Misra and N. S. Raghuwanshi “Wireless sensor networks for agriculture: The state-of-the-art in practice and future challenges”. Computers and Electronics in Agriculture, 118, pp.66-84, 2015.

[5] K. Shinghal and S. Neelam. “Wireless Sensor Networks in Agriculture: For Potato Farming”. Available at SSRN: https://ssrn.com/abstract=3041375, 2017.

[6] X. Jianfa. “An environment monitoring system for precise agriculture based on wireless sensor networks”. IEEE International Conference Mobile Ad-hoc and Sensor Networks (MSN), pp. 28-35, 2011.

[7] P. Milind. “H2E2: A hybrid, hexagonal & energy efficient WSN green platform for precision agriculture”. IEEE 12th International Conference on Hybrid Intelligent Systems (HIS), pp.155-160, 2012.

[8] G. Nikesh, R.S. Kawitkar. “Smart Agriculture Using IoT and WSN Based Modern Technologies”. International Journal of Innovative Research in Computer and Communication Engineering, 4(6), pp. 12070-12076, 2016.

[9] M. Ibrahim, M. Rawidean, M. Kassim and A.N. Harun. “IoT in precision agriculture applications using wireless moisture sensor network”. IEEE Conference on Open Systems, pp.24-29, 2016.

[10] S. Lachure, A. Bhagat and J. Lachure, J. “Review on precision agriculture using wireless sensor network”. International Journal of Applied Engineering Research, 10(20), pp.16560-16565, 2015.

[11] F. Kiani, E. Amiri, M. Zamani, T. Khodadadi, and A. Abdul Manaf, “Efficient intelligent energy routing protocol in wireless sensor networks”, International Journal of Distributed Sensor Networks, 2015, pp.1-13.

[12] F. Kiani, “AR-RBFS: Aware Routing Protocol based on Recursive Best-First Search Algorithm for Wireless Sensor Networks”, Journal of Sensors, 2016, pp.1-11.

[13] F. Kiani, “Reinforcement Learning based Routing Protocol for Wireless Body Sensor Networks”. The 7th IEEE International Symposium on Cloud and Service Computing, pp.71-78, 2017.

[14] K. Sohraby, M. Daniel and Z. Taieb. “Wireless sensor networks: technology, protocols, and applications”. John Wiley & Sons Book, 2007.

[15] F. Akyildiz and M. C. Vuran. “Wireless sensor networks”. John Wiley & Sons Book, 2010.

[16] M. Stoces, J. Vanek, J. Masner and J. Pavlík. “Internet of things (IoT) in agriculture-selected aspects”. Agris on-line Papers in Economics and Informatics, 8(1), pp.83-89, 2016.

[17] L. Atzori, A. Iera and G. Morabito, “The Internet of Things: A survey,” Computer Network, 54(15), pp. 2787-2805, 2010.

[18] S. Yerpude, T. K. Singhal. “Impact of Internet of Things (IoT) Data on Demand Forecasting”. Indian Journal of Science and Technology, 10(5), 1-5, 2017.

Rule Based Artificial Intelligent System of Cucumber Greenhouse Environment Control with IoT Technology

Kohei Arai
Graduate School of Science and Engineering
Saga University
Saga City, Japan

Yoshikazu Saitoh
Takeo Training Farm
Saga Prefectural Training Farm
Takeo City, Japan

***Abstract*—The method proposed here allows control cucumber greenhouse environment based on IoT technology. IoT sensors are to measure the room and air temperature, relative humidity, CO_2 content, water supply, liquid fertilizer, water content. The basic system is rule based system. All the required rules to control the cucumber greenhouse environment are proposed here. Through regressive analysis between IoT sensor data and the harvest cucumber quality, it is found that the proposed rule based system is appropriate to control the cucumber greenhouse environment.**

Keywords—Temperature; relative humidity; CO_2 content; water supply; liquid fertilizer; rule-based system; IoT; artificial intelligence; expert system

I. Introduction

Vitality monitoring of vegetation is attempted with photographic cameras [1]. Grow rate monitoring is also attempted with spectral reflectance measurements [2]. Bi-Directional Reflectance Distribution Function: BRDF is related to the grow rate for tealeaves [3]. Using such relation, sensor network system with visible and near infrared cameras is proposed [4]. It is applicable to estimate nitrogen content and fiber content in the tealeaves in concern [5]. Therefore, damage grade can be estimated with the proposed system for rice paddy fields [6]. This method is validated with Monte Carlo simulation [7]. Also Fractal model is applied to representation of shapes of tealeaves [8]. Thus the tealeaves can be asse3ssed with parameters of the fractal model. Vitality of tea trees are assessed with visible and near infrared camera data [9].

Rice paddy field monitoring with radio-control drone mounting visible and NIR camera is proposed [10] while the method for rice quality evaluation through nitrogen content in rice leaves is also proposed [11]. The method proposed here is to evaluate rice quality through protein content in rice crop with observation of NDVI which is acquired with visible and NIR camera mounted on radio-control drone. Rice crop quality evaluation method through regressive analysis between nitrogen content and near infrared reflectance of rice leaves measured from near field radio controlled drone is proposed and validated successfully [12].

Meanwhile, estimation of protein content in rice crop and nitrogen content in rice leaves through regressive analysis with NDVI derived from camera mounted radio-control drone is conducted successfully [13]. On the other hand, relation between rice crop quality (protein content) and fertilizer amount as well as rice stump density derived from drone data is well investigated [14]. Then, estimation of rice crop quality and harvest amount from drone mounted NIR camera data and remote sensing satellite data is carried out [15]. Furthermore, effect of stump density, fertilizer on rice crop quality and harvest amount in 2015 investigated with drone mounted NIR camera data is well reported [16]. Moreover, method for NIR reflectance estimation with visible camera data based on regression for NDVI estimation and its application for insect damage detection of rice paddy fields is proposed and validated [16].

There is a strong demand for automatic environmental condition control system of green-house in order to improve farmers' labor cost reduction as well as resource reduction. Also, in order to decrease the barriers to entry into agriculture, easy system for automatic environmental condition control system of green-house is highly required. The method proposed here allows cucumber greenhouse environment control based on IoT technology. In artificial intelligence, an expert system is a computer system that emulates the decision-making ability of a human expert [17]. Expert systems are designed to solve complex problems by reasoning through bodies of knowledge, represented mainly as if–then rules rather than through conventional procedural code [18]. The first expert systems were created in the 1970s and then proliferated in the 1980s [19]. Expert systems were among the first truly successful forms of artificial intelligence (AI) software [20]-[24].

An expert system is divided into two subsystems: the inference engine and the knowledge base. The knowledge base represents facts and rules. The inference engine applies the rules to the known facts to deduce new facts. Inference engines can also include explanation and debugging abilities [25]. The most important things are environment control rules which are derived from the acquired environmental data through IoT sensors.

A study on greenhouse automatic control system based on wireless sensor network is recently well reported [26]-[34]. This is the typical system for greenhouse automatic control. Not only wireless sensor network but also knowledge base system and IoT related technologies are required for more efficient feedback control system which refers the product quality for control the environments of the greenhouse in concern.

The proposed method is described in the next section followed by experiments. The experimental results are validated in the following section followed by conclusion with some discussions.

II. Proposed Method

A. IoT Sensor System

The method proposed here allows control cucumber greenhouse environment based on IoT technology. IoT sensors are to measure the room and air temperature, relative humidity, CO_2 content, water supply, liquid fertilizer, water content. The basic system is rule based system. All the required rules for control the cucumber greenhouse environment are proposed here.

Fig. 1 shows IoT sensor system of the proposed rule based artificial intelligent system of cucumber greenhouse environment control system. In the ambient, There are solar illumination sensor, air temperature sensors at the east end and the west end, relative humidity sensors at the both ends, as well as CO_2 sensor in the west end, while room temperature sensors at the both ends as well as room humidity sensors at the both ends, and also CO_2 sensor at the east end. Furthermore, water supply and liquid fertilizer sensor are also equipped. Moreover, water content in the soil, leaf color and size can be monitored with camera images.

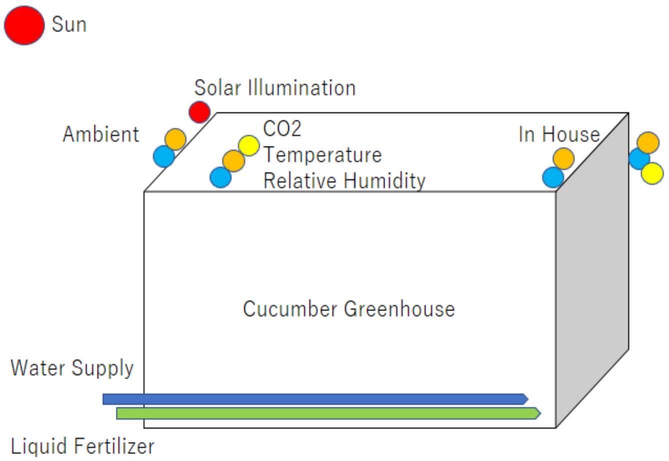


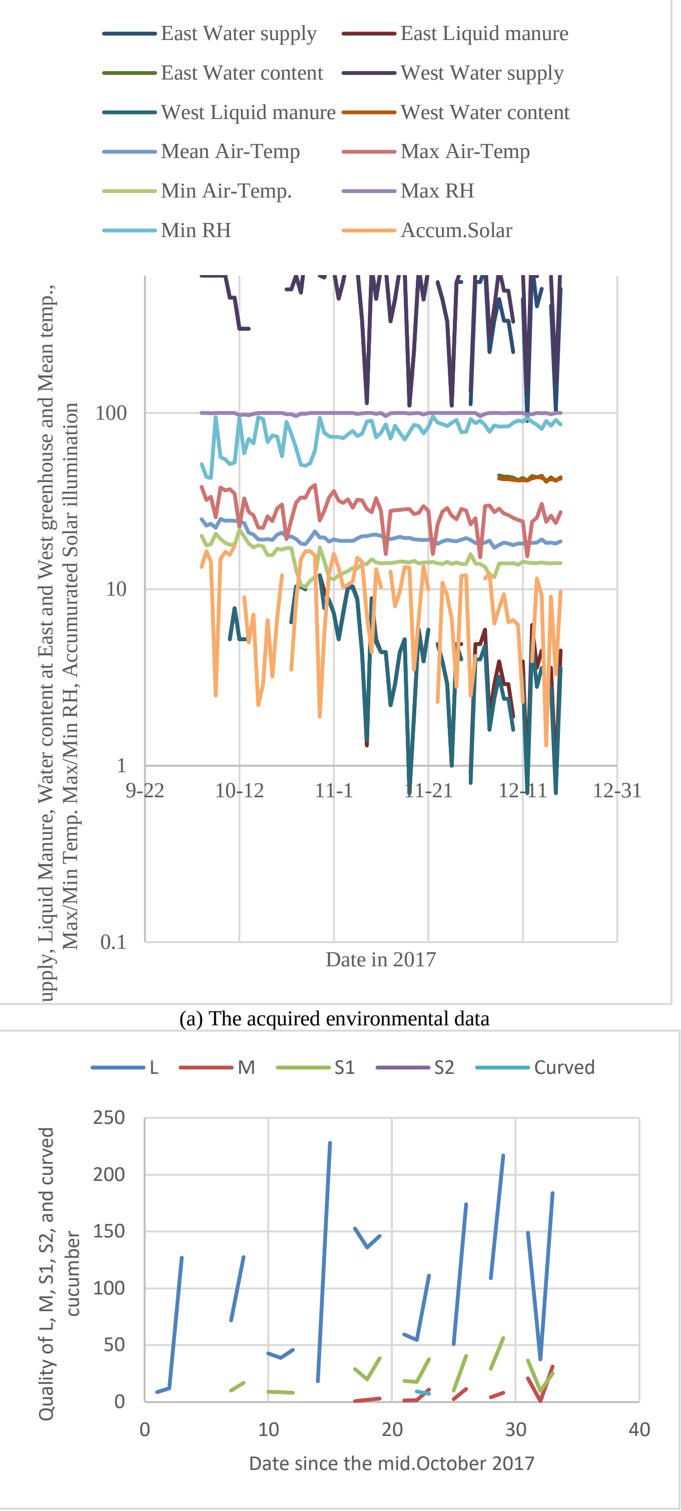
Fig. 1. IoT sensor system of the proposed rule based artificial intelligent system of cucumber greenhouse environment control system.

With the acquired environmental data and cucumber products data of quality and harvest amount, the all the required rules for maximizing cucumber product quality and harvest amount can be derived.

III. Experiment

A. Preliminary Acquired Data

Fig. 2(a) and (b) shows the example of the acquired environmental data and the cucumber products data of quality and harvest amount. By using the environmental and product data which are acquired from October 2017 to February 2018, all the required rules for control room temperature, room relative humidity, water supply, liquid fertilizer, room CO_2, can be derived.



(a) The acquired environmental data

(b) The cucumber products data of quality and harvest amount

Fig. 2. Example of the acquired environmental data and the cucumber products data of quality and harvest amount.

B. Derived Rules

The derived rules of room temperature is as follows:

1) Temperature Control:

It becomes management by principle "saturation difference value". It controls in correlation with humidity.

Means and means for raising the temperature inside the house:

○ October - March

① Heating machine (degree of influence 90%)

② Solar illumination (influence degree 10%)

Because the solar radiation is weak and the cloudy weather days are quite large compared to the Pacific side production area, we mainly keep the temperature by the heating machine.

○ April - September

① Solar illumination (degree of influence 85%)

② Heating machine (influence degree 15%)

From April the solar radiation gradually becomes stronger and the outside air temperature also rises.

However, since it may cool down in the evening too early in the morning around April, heater is often used.

Means and means for raising the temperature inside the house:

○ October - March

① Outside temperature (degree of influence 50%)

② Skylight ventilation (influence degree 50%)

Since the outside air temperature is low, basically the inside temperature of the house drops at the outside air temperature.

In the autumn of October to November, the daytime temperature may rise, so you may open the skylight to lower the temperature inside the house.

○ April - September

① Ventilation with skylight (degree of influence 85%)

② Side ventilation (influence degree 15%)

Because it will be a period of high trend both during the day and at night, it is a point when the temperature inside the house is lowered.

Actually both of ① and ② influence degree is 50% 50%, there are cases where side ventilation is not used due to the invasion of pests in the house.

On the other hand, room humidity control rules are as follows,

2) Relative Humidity Control:

It becomes management by principle "saturation difference value". It controls in correlation with temperature. As shown in the Table I on the right, the value of "saturation value 3 to 6". Control the temperature and humidity to make it. Where the color is attached corresponds to the proper value

In the case of cucumber it is a feeling of 2 ~ 6.

Means and degree of influence for raising humidity in house:

○ Same throughout the year (not related to winter or summer)

① transpiration of cucumber (degree of influence 70%)

② Watershed watering (influence degree 20%)

③ Skylight ventilation "close" (influence degree 10%)

Humidity inside the house basically keeps 80 ~ 90%, depending on the temperature. (This is sudden deviation management). Recently, a small fog cooling facility called "mist" has come out, but farmers who are introducing are rare at the moment. (It is likely to increase in the future.) If we put mist above, the influence will be around 90%.

TABLE I. ROOM TEMPERATURE AND RELATIVE HUMIDITY CONTROL

Room Temp。 (deg.C)	Relative Humidity(%)									
	95	90	85	80	75	70	65	60	55	50
16	0.7	1.4	2	2.7	3.4	4.1	4.8	5.5	6.2	6.7
17	0.7	1.5	2.2	2.9	3.6	4.3	5	5.8	6.5	7.2
18	0.8	1.5	2.4	3.1	3.8	4.6	5.4	6.2	7	7.7
19	0.8	1.6	2.5	3.3	4.1	4.9	5.7	6.5	7.4	8.2
20	0.9	1.7	2.6	3.5	4.4	5.2	6	6.9	7.8	8.7
21	0.9	1.8	2.7	3.7	4.6	5.5	6.4	7.4	8.3	9.3
22	1	2	2.9	3.9	4.9	5.7	6.8	7.7	8.8	9.7
23	1	2.1	3.1	4.2	5.2	6.3	7.3	8.3	9.3	10.3
24	1.1	2.2	3.3	4.4	5.5	6.5	7.7	8.7	9.8	10.9
25	1.2	2.3	3.5	4.7	5.8	6.9	8.1	9.3	10.4	11.5
26	1.3	2.5	3.7	4.9	6.1	7.4	8.5	9.8	10.9	12.2
27	1.3	2.7	3.9	5.2	6.4	7.7	9	10.3	11.6	12.9
28	1.4	2.8	4.2	5.5	6.7	8.2	9.5	10.9	12.3	13.6
29	1.4	2.9	4.4	5.8	7.3	8.6	10.1	11.5	13	14.4
30	1.5	3	4.7	6.2	7.6	9.1	10.6	12.1	13.6	15.2

Means and degree of influence to lower house humidity:

○ Same throughout the year (not related to winter or summer)

① Heating machine (degree of influence 50%)

② Skylight ventilation "opening" (degree of influence 50%)

The above is the way humans intentionally lower. The solar radiation is strong from April to June, and it is very difficult to maintain 70 to 90% if humidity is intentionally raised. Also, if the cucumber leaves get wilted, transpiration will also decrease, so it will become more dry. The humidity is lowered by the above means mainly because the inside of the house at night is high humidity (humidity becomes 100% depending on the day). We will intentionally use skylight ventilation and heating when the humidity becomes 100%.

In addition, there are the following points to note about the saturation value. “harvest start time ~ harvest end time” is used as a guideline to manage while checking the value of the difference. In addition, in the period of “settling ~ harvest start”, the number of leaves is small and transpiration from the leaves is small, so the saturation value cannot be reached. Therefore, temperature and humidity management will be carried out while confirming the withering condition etc. of the cucumber.

When the set concentration value is cut off while always measuring with the CO_2 sensor, the control device issues an

ON command to the carbon dioxide gas generator. After that, it is controlled so that it turns OFF when it exceeds the set upper limit concentration.

Meanwhile, CO_2 control rules are as follows:

3) CO_2 Control:
We change the CO_2 concentration in the house according to the time.

- Winter season

Basically, the skylight does not open (because it warms the inside of the house with heating), so it is set to 550 ~ 600ppm.

We are trying to increase photosynthesis efficiency by making it darker than the concentration of outside air (approximately 400 ppm).

- Spring - summer season

Since the temperature inside the house will rise due to solar radiation, open and close the skylight to control the temperature inside the house.

Due to the nature of CO_2, it cannot be made as high as in winter because it has the property to move from high concentration to low concentration. Therefore, control is made so that 400 to 440 ppm can be maintained so that the state can be maintained almost same as the outside air concentration.

Water supply and liquid fertilizer control rules are as follows:

4) Water supply and liquid manure control:
It is difficult for water, liquid fertilizer. The basic judgment criterion is "looking at the state of cucumber" and "moisture content".

Point to judge

"State of cucumber"
① Leaf color

→ Leaf color is thin · Water is high or there is not enough fertilizer → Action is "Reduce water" or "To increase liquid fertilizer"

→ Leaf color is dense · Water is little or fertilizer is too effective → Action is "increase water" or "thin or cut liquid fertilizer"

② Leaf size
→ Leaves are developing widely · · · Water works well. The action in this case is "as is".

→ Leaves are small · Water is not working → Action increases water.

Leaf color and size can be estimated with the camera images. Example of the acquired camera image is shown in Fig. 3.

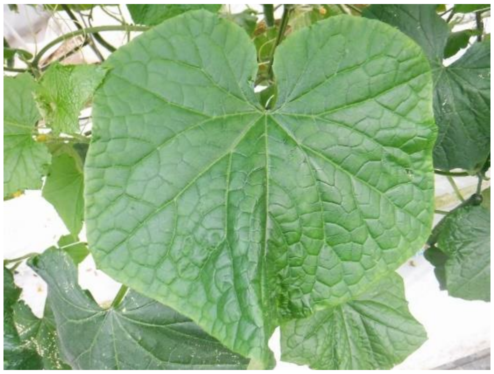


Fig. 3. Example of the acquired camera image of cucumber leaf.

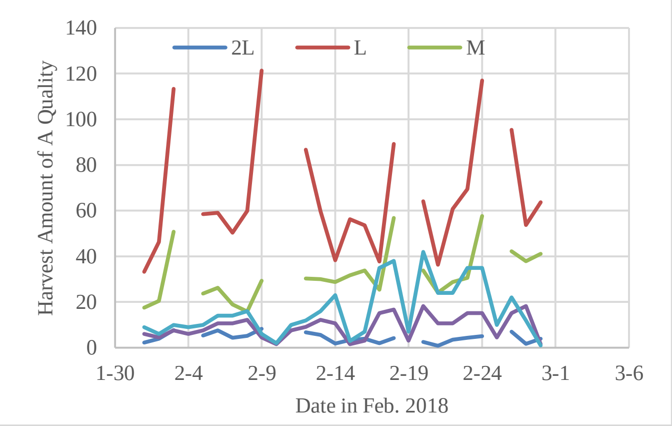
"Moisture percentage"

This is my own way. Confirm the moisture content at the end of the day (around 17 o'clock).

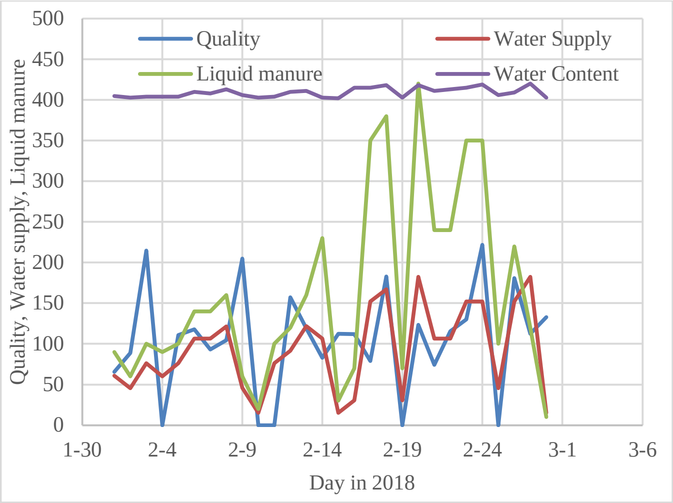
- Moisture content is clearly decreasing → We firmly absorb water but judge that it is not enough to increase irrigation volume
- Moisture content has not changed so much → It is judged that it is roughly irrigated.
- Moisture content is increasing → Multiple irrigation. Set to reduce irrigation setting.

C. *Validation of the Proposed Role Base System*

The proposed rule based artificial intelligent system of cucumber greenhouse environment control with IoT technology is validated with the daily averaged of the environmental data and the product data which are acquired in February 2018. Fig. 4(a) and (b) shows the product quality and the environmental data of water supply, liquid fertilizer and water content in the soil. In the Fig. 4(b), summarized product quality is shown as 2L+0.5L+0.3M (the number of highest quality of cucumber: 2L followed by L and M. therefore, Quality in Fig. 4(b) is calculated weighted sum of these numbers).



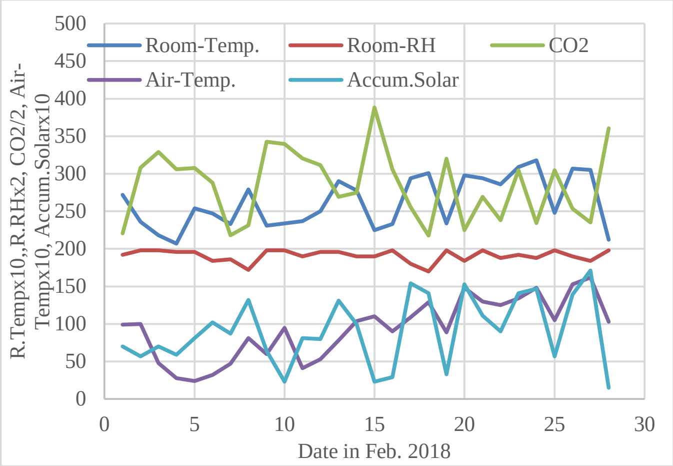
(a)Product quality data



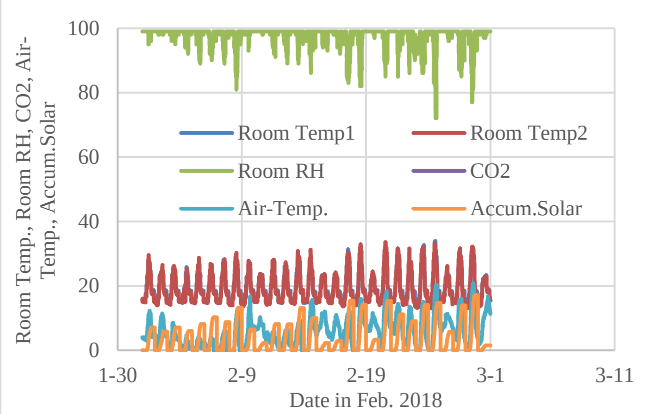
(b)Environmental data

Fig. 4. Product quality and the environmental data of water supply, liquid fertilizer and water content in the soil.

Other daily averaged environmental data of room temperature, room relative humidity (RH). CO_2, Air temperature, accumulated solar illumination acquired in February 2018 are shown in Fig. 5(a) while hourly averaged environmental data is shown in Fig. 5(b), respectively.



(a)Daily average



(b)Hourly average

Fig. 5. Daily and hourly averaged environmental data.

All these product and environmental data acquired in February 2018 is shown in Fig. 6. The correlation coefficients between the products quality and the other environmental data of water supply (WS), liquid fertilizer (LF), water content in the soil (WC), room temperature (RT), room relative humidity (RH), room CO_2, (CO) air temperature (AT) and accumulated solar illumination (AS) are as follows:

WS	LF	WC	RT	RH	CO	AT	AS
0.400	0.339	0.415	0.324	-0.212	-0.160	0.216	0.370

These correlation coefficients are derived from the Fig. 4 of products quality data and the environmental data. Essentially, there is no relation among these parameters, WS LF WC RT RH CO AT AS because we can set these parameters intentionally.

Therefore, the product quality (Q) can be estimated with the following equation,

$$Q=0.237WS+0.2LF+0.245WC+0.192RT-0.125RH-0.094CO+0.127AT+0.219AS \tag{1}$$

The most influencing environmental factor to the product quality is water content in the soil followed by water supply, liquid fertilizer, room temperature, air temperature, room relative humidity, and CO_2 concentration.

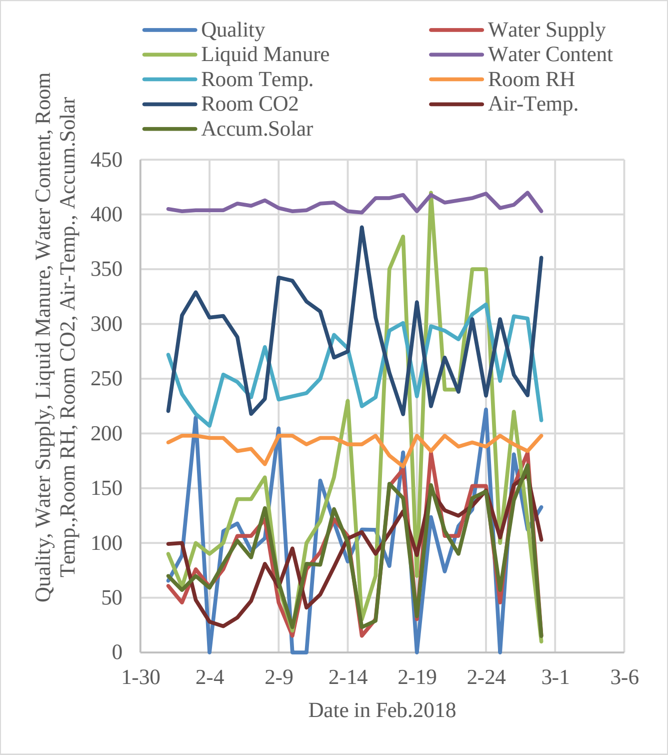


Fig. 6. Product and environmental data acquired in February 2018.

IV. Conclusion

The method which allows control cucumber greenhouse environment based on IoT technology is proposed. IoT sensors are to measure the room and air temperature, relative humidity, CO_2 concentration, water supply, liquid fertilizer, water content in the soil. The basic system is rule based system. All the required rules for control the cucumber greenhouse environment are also proposed. Through regressive analysis between IoT sensor data and the harvest cucumber quality, it is found that the proposed rule based system is appropriate for control the cucumber greenhouse environment. Also, it is found that the most influencing environmental factor to the product quality is water content in the soil followed by water supply, liquid fertilizer, room temperature, air temperature, room relative humidity, and CO_2 concentration.

Further research works are required for improvement of the prediction accuracy of cucumber quality and harvest amount. Also, cost performance evaluation is required for water and fertilizer managements.

Acknowledgment

The author would like to thank Mr. Kazuhiro Nagafuchi of Saga Prefectural Local Government for his effort to conduct the experiments.

References

[1] J.T.Compton, Red and photographic infrared linear combinations for monitoring vegetation, Journal of Remote Sensing of Environment, 8, 127-150, 1979.

[2] C.Wiegand, M.Shibayama, and Y.Yamagata, Spectral observation for estimating the growth and yield of rice, Journal of Crop Science, 58, 4, 673-683, 1989.

[3] Kohei Arai, Method for estimation of grow index of tealeaves based on Bi-Directional reflectance function:BRDF measurements with ground based netwrok cameras, International Journal of Applied Science, 2, 2, 52-62, 2011.

[4] Kohei Arai, Wireless sensor network for tea estate monitoring in complementally usage with Earth observation satellite imagery data based on Geographic Information System(GIS), International Journal of Ubiquitous Computing, 1, 2, 12-21, 2011.

[5] Kohei Arai, Method for estimation of total nitrogen and fiber contents in tealeaves with grond based network cameras, International Journal of Applied Science, 2, 2, 21-30, 2011.

[6] Kohei Arai, Method for estimation of damage grade and damaged paddy field areas sue to salt containing sea breeze with typhoon using remote sensing imagery data, International Journal of Applied Science,2,3,84-92, 2011.

[7] Kohei Arai, Monte Carlo ray tracing simulation for bi-directional reflectance distribution function and grow index of tealeaves estimation, International Journal of Research and Reviews on Computer Science, 2, 6, 1313-1318, 2011.

[8] K.Arai, Fractal model based tea tree and tealeaves model for estimation of well opened tealeaf ratio which is useful to determine tealeaf harvesting timing, International Journal of Research and Review on Computer Science, 3, 3, 1628-1632, 2012.

[9] K.Arai, H.Miyazaki, M.Akaishi, Determination of harvesting timing of tealeaves with visible and near infrared cameradata and its application to tea tree vitality assessment, Journal of Japanese Society of Photogrammetry and Remote Sensing, 51, 1, 38-45, 2012

[10] Kohei Arai, Osamu Shigetomi, Yuko Miura, Hideaki Munemoto, Rice crop field monitoring system with radio controlled drone based near infrared cameras through nitrogen content estimation and its distribution monitoring, International Journal of Advanced Research in Artificial Intelligence, 2, 3, 26-37, 2013

[11] Kohei Arai, Rice crop quality evaluation method through regressive analysis between nitrogen content and near infrared reflectance of rice leaves measured from near field radio controlled drone, International Journal of Advanced Research in Artificial Intelligence, 2, 5, 1-6, 2013.

[12] Kohei Arai, Masanori Sakashita, Osamu Shigetomi, Yuko Miura, Estimation of protein content in rice crop and nitrogen content in rice leaves through regressive analysis with NDVI derived from camera mounted radio-control drone, International Journal of Advanced Research in Artificial Intelligence, 3, 3, 7-14, 2014.

[13] Kohei Arai, Masanori Sakashita, Osamu Shigetomi, Yuko Miura, Relation between rice crop quality (protein content) and fertilizer amount as well as rice stump density derived from drone data, International Journal of Advanced Research on Artificial Intelligence, 4, 7, 29-34, 2015.

[14] Kohei Arai, Masanori Sakashita, Osamu Shigetomi, Yuko Miura, Estimation of Rice Crop Quality and Harvest Amount from Drone Mounted NIR Camera Data and Remote Sensing Satellite Data, International Journal of Advanced Research on Artificial Intelligence, 4, 10, 16-22, 2015.

[15] Kohei Arai, Gondoh, Miura, Shigetomi, Effect of Stump density, Fertilizer on Rice Crop Quality and Harvest Amount in 2015 Investigated with Drone mounted NIR Camera Data, International journal of Engineering Science and research Technology, 2, 2, 1-7, 2016

[16] Kohei Arai, Kenji Gondoh, Osamu Shigetomi, Yuko Miura, Method for NIR Reflectance Estimation with Visible Camera Data Bsed on Regression for NDVI Estimation and Its Application for Insect Damage Detection of Rice Paddy Fields, International Journal of Advanced Research on Artificial Intelligence, 5, 11, 17-22, 2016.

[17] Jackson, Peter (1998), Introduction To Expert Systems (3 ed.), Addison Wesley, p. 2, ISBN 978-0-201-87686-4

[18] "Conventional programming". Pcmag.com. Retrieved 2013-09-15.

[19] Leondes, Cornelius T. (2002). Expert systems: the technology of knowledge management and decision making for the 21st century. pp. 1–22. ISBN 978-0-12-443880-4.

[20] Russell, Stuart; Norvig, Peter (1995). Artificial Intelligence: A Modern Approach (PDF). Simon & Schuster. pp. 22–23. ISBN 0-13-103805-2. Retrieved 14 June 2014.

[21] Luger, George; Stubblefield, William (2004), Artificial Intelligence: Structures and Strategies for Complex Problem Solving (5th ed.), The Benjamin/Cummings Publishing Company, Inc., p. 720, ISBN 0-8053-4780-1, 2004.

[22] Nilsson, Nils. Artificial Intelligence: A New Synthesis. Morgan Kaufmann Publishers. 1998. ISBN 978-1-55860-467-4. 1998.

[23] McCorduck (2004, p. 190-25) discusses Frankenstein and identifies the key ethical issues as scientific hubris and the suffering of the monster, i.e. robot rights, pp. 327–335, 434–435. 2004.

[24] Crevier, Daniel (1993), AI: The Tumultuous Search for Artificial Intelligence, New York, NY: BasicBooks, ISBN 0-465-02997-3, pp. 145–62, 197−203, 1993.

[25] Nwigbo Stella and Agbo Okechuku Chuks, School of Science Education, Expert system: a catalyst in educational development in Nigeria: "Knowledge-based systems collect the small fragments of human know-how into a knowledge-base which is used to reason through a problem, using the knowledge that is appropriated" Proceedings of the 1st International Technology, Education and Environment Conference (c) African Society for Scientific Research (ASS), 566-571, 2010.

[26] Dae-Heon ParkBeom-Jin KangKyung-Ryong ChoChang-Sun ShinSung-Eon ChoJang-Woo ParkEmail authorWon-Mo Yang, A Study on Greenhouse Automatic Control System Based on Wireless Sensor Network, Wireless Personal Communications, January 2011, Volume 56, Issue 1, pp 117–130, 2011.

[27] Alves-Serodio, C. M. J., Monteiro, J. L., & Couto, C. A. C. (1998). An integrated network for agricultural management applications. IEEE International Symposium on Electromagnetic Compatibility. America: Denver, pp. 679–683., 1998

[28] Serôdio C., Cunha J. B., Morais R., Couto C., Monteiro J. (2001) A networked platform for agricultural management systems. Computers and Electronics in Agriculture 31(1): 75–90, 2001..

[29] Sensors Magazine. (2004). Editorial: This changes everything—market observers quantify the rapid escalation of wireless sensing and explain its effects. Wireless for Industry, Supplement to Sensors Magazine, Summer, pp. S6–S8., 2004.

[30] Wang N., Zhang N., Wang M. (2006) Wireless sensors in agriculture and food industry—recent development and future perspective. Computer and Electronics in Agriculture 50: 1–14, 2006.

[31] Kim Y.-S. (2004) Expert development for automatic control of greenhouse environment. Journal of Korean Flower Research Society 12(4): 341–345, 2004..

[32] Huang, Y.-J., Evans, N., Li, Z.-Q., Eckert, M., Chevre, A.-M., Renard, M., & Fitt, B. D. L. (2006). Temperature and leaf wetness duration affect phenotypic expression of Rlm6-mediated resistance to Leptosphaeria maculans in Brassica napus. New Phytologist 129–141, 2006..

[33] Hartman J. R. (1999) Effect of leaf wetness duration, temprature, and conidial inoculum dose on apple scab infections. The American Phytopathological Society 83(6): 531–534, 1999.

[34] Wilks, D. S., & Shen, K. W. (1991). Threshold relative humidity duration forecasts for plant disease prediction. American Meteorological Society, pp. 463–465., 1991.

AUTHOR'S PROFILE

Kohei Aarai 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 and 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 30 books and published 500 journal papers.

Pedestrian Detection Approach for Driver Assisted System using Haar based Cascade Classifiers

M. Ameen Chhajro[1], Kamlesh Kumar[2], M. Malook Rind[3], Aftab Ahmed Shaikh[4], Haque Nawaz[5]
Department of Computer Science
Sindh Madressatul Islam University, Karachi, Sindh, Pakistan

Rafaqat Hussain Arain[6]
Department of Computer Science
Shah Abdul Latif University, Khairpur, Sindh, Pakistan

***Abstract*—Object detection and tracking with the aid of computer vision is a most challenging task in the context of Driver Assistant System (DAS) for vehicles. This paper presents pedestrians detection techique using Haar-Like Features. The main aim of this research is to develop a detection system for vehicle drivers that will intimate them in advance for pedestrian's movement when they are crossing the zebra region or passing nearby to it along the road. For this purpose, dataset of 1000 images have been taken via CCTV camera which was mounted for road monitoring. A Haar based cascade classifiers have been implemented over images. And system is trained for positive (with people) and negative (without people) image samples, respectively. After testing, the obtained results show that it attained 90% accuracy while pedestrian detection. The proposed work provides significant contribution in order to reduce the road accidents as well as ensure the safety measurement for road management.**

***Keywords*—Pedestrian; Haar based classifier; positive and negative samples; computer vision; object detection**

I. Introduction

During past few years, field of computer vision for the object detection and tracking has been a hotspot area for researchers. This has become due to growing interest in visual data; as to acquire hidden patterns that could enable machines for automated decision making tasks. And it might be explored through numerous applications which proves the potential benefits of this research domain, e.g. Road Sign Recognition, Passive surveillance, License Plate Recognition, Face Tracking, Pedestrian Tracking and beside other related area [1]-[6]. Detecting pedestrian remains an ineluctable task for driver assistance system. However, pedestrian movement recognition is also a crucial research problem that cannot be overlooked. Therefore, it is matter of more attention to detect pedestrian as to avoid any adverse accident and to control the speed of vehicle. In this paper, a vision based driving assistant system has been proposed which aims to detect people which are near to or crossing the zebra region. This work provides significant contribution over people safety.

The rest of the paper is arranged as follows. Section II gives literature review related to pedestrian detection. Section III explains training procedure of cascading classifier for tracking objects in an image. Section IV provides simulation results and discussion over zebra crossing people detection. Finally, Section V gives the conclusion and future work.

II. Literature Review

Over last few years, various works has been reported in pedestrian detection for driver assistance system. In this regard, J. Hariyono and K. Hyun Jo suggested [7] a model for detection of pedestrian crossing the road. In which pedestrian pose and lateral speed are recognized using spatial body mass ratio. Whereas, motion trajectory and spatial layout are obtained through centroid of human region and distance of the pedestrian along the road lane boundary. Similarly, B. Riveiro et al. [8] proposed mobile LIDAR data technique for automatic detection of zebra crossing. It uses several successive processes, begins with segmentation of road for curvature analysis in each laser cycle. Then, implements rasterization and Standard Hough Transform techniques, in order to detect zebra crossing. This work is beneficial for road managers that need Geographic Information Systems. Object detection method for advanced driver assistance systems (ADAS) was presented by M. Kang and Y.C Lim [9]. In this research, authors employ fully convolutional network (FCN) in order to detect objects in road environment. In a similar context, B. Wang et al. [10] provided an approach for multiple object tracking and detection for road obstacles. Their system works well when this was tested on different traffic video sequences over public database. Sanchez et al. [11] used Convolutional Neural Network for pedestrian movement and direction recognition, in which histograms of gradients (HOG) and SVM have been implemented for pedestrian detection. Another paper C. Caramuta et al. [12] provided comprehensive survey over pedestrian dynamics detection techniques, simulation and mathematical models. Similarly, Hui Zhou and Wan Hang [13] demonstrated Lane Detection and Tracking for Driver Assistance Systems.

III. Cascading Classifiers

The cascade classifiers are special case of ensemble learning, where several classifiers are combined together [14]. These are mostly used for object detection from images and it was firstly proposed by Viola and Jones in 2001 for face detection. The cascade classifiers are trained for positive and negative images for specific object detection. However, training phase of classifier requires tradeoff between feature selection and its computation. For example, classifier with more features produces higher detection rates as compared to false positive results. But on the other side, it increases its computation time. Therefore, optimization of classifier depends upon selection of framework for training. For this purpose, one must consider following parameters. These are 1) number of classifier stages, 2) the number of features in each stage, and 3) threshold in each stage. In fact, an effective classifier is one that could work well in real time object detection, and such kind of classifiers are very difficult to design. However, cascade classifiers under goes subsequent training stages for target selection. And during each stage in the cascade, it reduces the false positive rate and increases the detection rate. Features are added to a classifier in each stage, until we get targeted object along with reduction of false positives rates. And, accuracy of the classifier is determined by testing it on a ground truth data.

IV. Proposed Methodology

In this paper, Haar Based Cascade Classifier has been implemented for object detection. This classifier has been used successfully for detection and tracking of objects in an image [15]. In order to train the classifier, an image sample is built for positive and negative images respectively for targeted and unwanted objects. The main aim of using classifier is to generate most optimized targeted values for detecting and tracking the object through varying the size of window for Haar feature selection as highlighted in Fig. 1. It filters the features of positive images and then creates specific target values through separation of black and white areas in the image features. And these are placed in an image where we want to perform object detection.

In an image there are different objects, but we are interested in finding the specific object from it. For this purpose, a training is carried out over dataset for positive and negative images for object location and detection. Testing of the classifier has been achieved for positive (with people) and negative (without people) images as shown in Fig. 2 and 3 respectively. These images are obtained via video sequence using a fixed CCTV Camera. And it is mounted on the road for monitoring the traffic. We have made dataset of 1000 images, each having 500 images for positive and negative samples, respectively. Thus classifier executes as aforementioned manner, however, speed of locating the objects in the image highly depends upon classifier training.

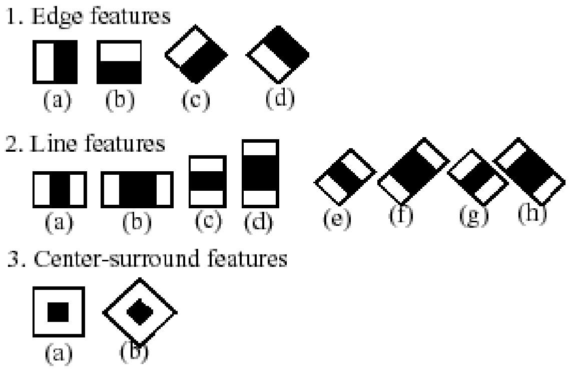


Fig. 1. Types of Haar features.

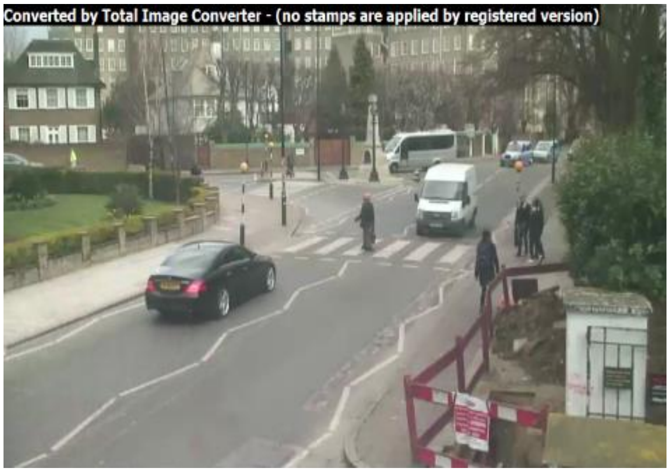

Converted by Total Image Converter - (no stamps are applied by registered version)

Fig. 2. Positive sample image.

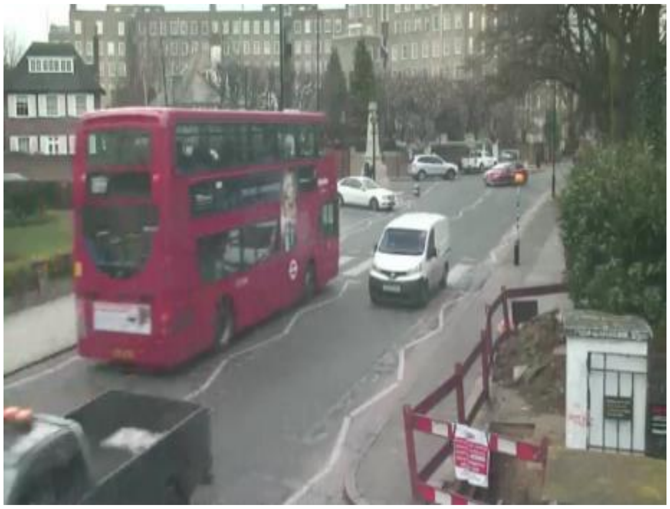


Fig. 3. Negative sample image.

V. Simulation Results and Discussions

In this section, simulation results are shown for proposed system and discussion is made over it. In this regard, testing is performed for some positive and negative images for people detection. However, research aim is to detect people which are near or over the zebra crossing region. For this purpose, we have taken 300 images for training, and then performed testing over 100 images. The performance of proposed system has been evaluated using correct and mis-detection metrics.

Table I depicts that system recognize 90 images with people out of 100 and leaving only 10 mis-identified images. It can be observed from Fig. 4 and 5, respectively, there are different objects namely: people, vehicles, poles, buildings, barriers and trees. But proposed system is trained to detect people when they are crossing the zebra or approaching near to it. However, in some cases it yields mis-detection, when both the people and beside other surrounding object are targeted as shown in Fig. 6. The graphical results in Fig. 7 shows that proposed system provides higher recognition rates as compared to mis-detection as it achieves accuracy of 90% when tested over dataset.

TABLE I. RESULTS OUTCOMES OF PROPOSED MODEL

No. of Images in Dataset	No. of Images for Training	No. of Images for Testing	Correctly Detected	Mis-Detected	Rate (%)
1000	300	100	90	10	90%

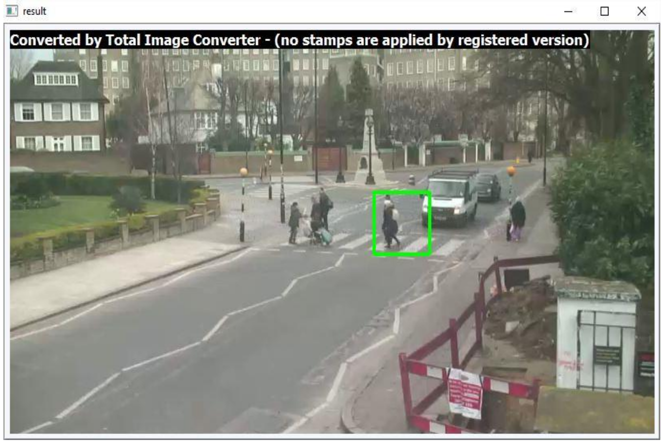

result
Converted by Total Image Converter - (no stamps are applied by registered version)

Fig. 4. Pedestrian detection along zebra crossing.

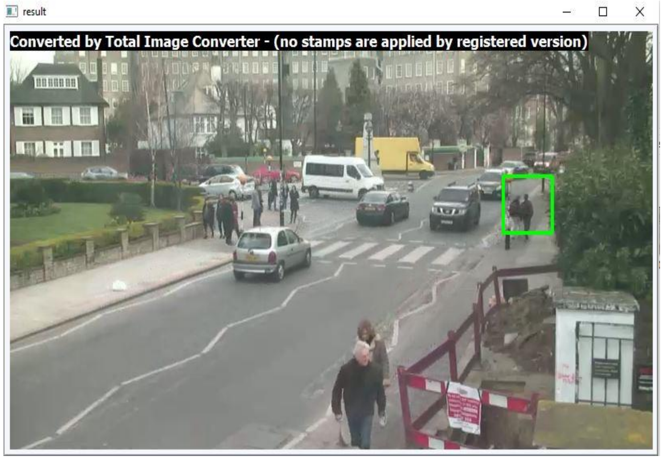

result
Converted by Total Image Converter - (no stamps are applied by registered version)

Fig. 5. Pedestrian detection near zebra crossing.

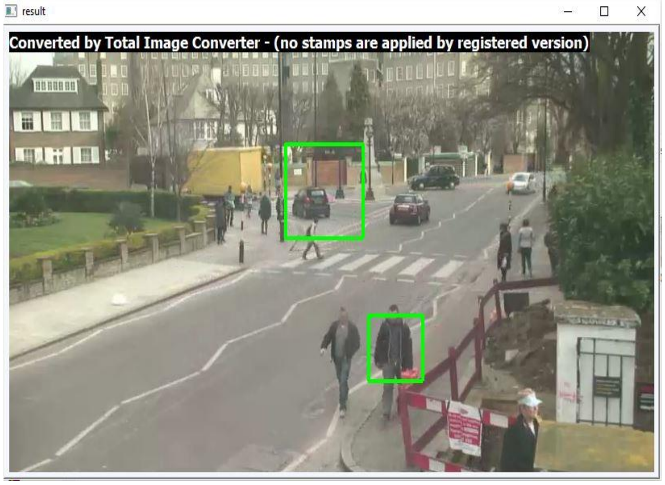

result
Converted by Total Image Converter - (no stamps are applied by registered version)

Fig. 6. Pedestrian mis-detection.

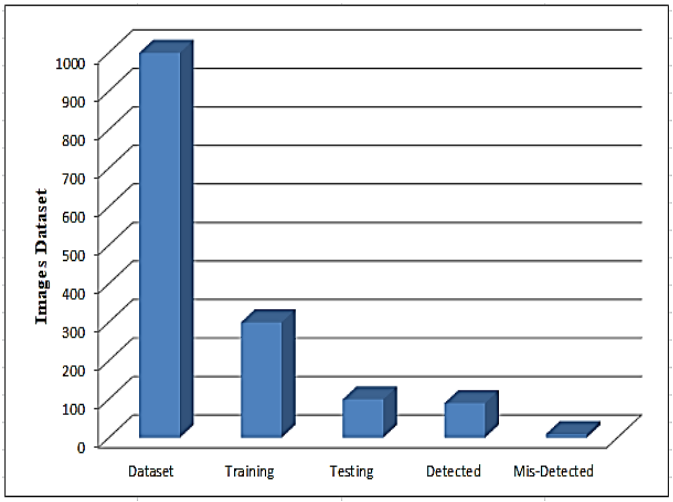


Fig. 7. Graphical results representation.

VI. Conclusion and Future Work

Accurate pedestrian detection and tracking remains a big dilemma in computer vision for driver assisted system. In this paper, people detection near to or crossing the zebra region was proposed for vehicle drivers. A Haar based cascade classifiers were used for feature detection for positive and negative samples respectively. The system was trained over 1000 images obtained via CCTV camera. After testing of 100 images, results showed that out of them 90 images were correctly identified leaving only 10 mis-identified images. However, overall system attained accuracy of 90%. In future, we shall improve the performance of this algorithm using deep learning technique for classification in order to improve its accuracy.

References

[1] Chen, Y., Zhao, D., Lv, L. and Zhang, Q., 2017. Multi-task learning for dangerous object detection in autonomous driving. Information Sciences.

[2] Dominguez-Sanchez, A., Cazorla, M. and Orts-Escolano, S., 2017. Pedestrian Movement Direction Recognition Using Convolutional Neural Networks. IEEE Transactions on Intelligent Transportation Systems, 18(12), pp.3540-3548.

[3] Zhong, Z., Lei, M., Cao, D., Fan, J. and Li, S., 2017. Class-specific object proposals re-ranking for object detection in automatic driving. *Neurocomputing*, *242*, pp.187-194

[4] Talib, H., Ismail, K. and Kassim, A., 2016. A robust approach for road users classification using the motion cues. *Transportation Research Part C: Emerging Technologies*, *73*, pp.77-90.

[5] Wang, B., Florez, S.A.R. and Frémont, V., 2014, December. Multiple obstacle detection and tracking using stereo vision: application and analysis. In Control Automation Robotics & Vision (ICARCV), 2014 13th International Conference on (pp. 1074-1079). IEEE.

[6] Wang, S., Pan, H., Zhang, C. and Tian, Y., 2014. RGB-D image-based detection of stairs, pedestrian crosswalks and traffic signs. *Journal of Visual Communication and Image Representation*, *25*(2), pp.263-272.

[7] J. Hariyono, K. H, Jo. Detection of Pedestrian Crossing Road A Study on Pedestrian Pose Recognition [J]. Neurocomputing, 2016, 234, 2017, 144-153

[8] B. Riveiroa, H. G. Jorge, J. M. Sánchez, L. D. Vilariño, P. Arias. Automatic detection of zebra crossings from mobile LiDAR data[J]. Optics & Laser Technology, 2015, 70, 63-70

[9] M. S. Kang, Y. C. Lim. High Performance and fast objection in road environment[C]. International Conference on Image Processing Theory, Canada, 2017, 1-6

[10] Wang, B., Florez, S.A.R. and Frémont, V., 2014, December. Multiple obstacle detection and tracking using stereo vision: application and analysis. In *Control Automation Robotics & Vision (ICARCV), 2014 13th International Conference on* (pp. 1074-1079). IEEE.

[11] Dominguez-Sanchez, A., Cazorla, M. and Orts-Escolano, S., 2017. Pedestrian Movement Direction Recognition Using Convolutional Neural Networks. *IEEE Transactions on Intelligent Transportation Systems*, *18*(12), pp.3540-3548.

[12] Caramuta, C., Collodel, G., Giacomini, C., Gruden, C., Longo, G. and Piccolotto, P., 2017. Survey of detection techniques, mathematical models and simulation software in pedestrian dynamics. *Transportation research procedia*, *25*, pp.551-567.

[13] Zhou, H. and Wang, H., 2017, November. Vision-based lane detection and tracking for driver assistance systems: A survey. In *Cybernetics and Intelligent Systems (CIS) and IEEE Conference on Robotics, Automation and Mechatronics (RAM), 2017 IEEE International Conference on* (pp. 660-665). IEEE.

[14] Kuo, Y.C., Yang, Z.Y. and Yen, C.H., 2012. Fast pedestrian detection system with a two layer cascade of classifiers. *Computers & Mathematics with Applications*, *64*(5), pp.1311-1323.

[15] Mohamed, A. Issam, A. Mohamed, B. and Abdellatif, B., 2015. Real-time detection of vehicles using the haar-like features and artificial neuron networks. *Procedia Computer Science*, *73*, pp.24-31.

MapReduce Performance in MongoDB Sharded Collections

Jaumin Ajdari, Brilant Kasami
Faculty of Contemporary Sciences and Technologies
South East European University (SEEU)
Tetovo, Macedonia

***Abstract*—In the modern era of computing and countless of online services that gather and serve huge data around the world, processing and analyzing Big Data has rapidly developed into an area of its own. In this paper, we focus on the MapReduce programming model and associated implementation for processing and analyzing large datasets in a NoSQL database such as MongoDB. Furthermore, we analyze the performance of MapReduce in sharded collections with huge dataset and we measure how the execution time scales when the number of shards increases. As a result, we try to explain when MapReduce is an appropriate processing technique in MongoDB and also to give some measures and alternatives to take when MapReduce is used.**

Keywords—NoSQL; big data; MapReduce; sharding; MongoDB

I. Introduction

We live in the era of the Information Age. Everything is connected and online services are more and more oriented to user data gathering. Major companies process hundreds of petabytes daily at their servers and the computations have to be distributed across hundreds or thousands of machines in order to finish it in a reasonable amount of time. The issues of how to parallelize the computation, distribute the data, and handle failures obscure the simple computations within a large amounts of complex code in dealing with them. With these problems in mind engineers try to borrow ideas from functional programming languages by using the map and reduce primitives as an abstraction that allows to express the simple computations, and hide the complex details of parallelization, fault-tolerance, data distribution and load balancing, hence MapReduce was introduced. The main purposes of this paper are:

- Analyzing MongoDB sharding capabilities
- What is MapReduce and why use it
- Presentation of the results using MapReduce in sharded collections by number of shards used.

In this paper we measure MapReduce time performance through MongoDB, and try to find out how the MapReduce execution time changes with increased number of MongoDB shards. We have described the environment, defined a mini cluster of three virtual machines on which MongoDB is run and we have experimented with a collection of relatively large number of documents. And at the end, the results and conclusions are shown, tried to answer some questions such are the use of MapReduce within MongoDB when is a good option, what needs to be done to speed up the processing and what alternatives to consider.

The rest of this paper is organized as follows: Section 2 presents a summary of some related work in this area. Section 3 contains a short description of the MongoDB where the main point is the shard techniques and possibility of sharding. Section 4 provides the testing results and MapReduce performance evaluation implemented on MongoDB by use different number of shards. Finally, Section 5 provides some conclusions.

II. Related Work

Big companies started facing issues on how to handle the huge amount of data they were receiving and how to process those. Google as the pioneer in search technologies needed computations that process a large amounts of data such as crawled documents, web request logs, graph structure of web documents, etc. According to authors Jeffrey D. and Sanjay G., Google needed a simple solution that was easy to understand, fault tolerant, cheap and reliable. In their paper [1] they analyze MapReduce in large clusters that are highly scalable where hundreds of programs are run and thousands of MapReduce jobs are executed, what is Google on a daily basis.

The authors Smita A. et al., in their paper [2] have introduced an explanation of the MongoDB, its features, advantages and disadvantages. Especially, they address the MongoDB features such as MapReduce, Auto – sharding, MongoDump, etc. They continue with their analyses in case of dealing with small and large amount of unstructured/semi structured data and at the end the conclude that if the amount of the data is big and permanently increases, and high performance and availability are required then MongoDB should be considered as options to use as database.

The authors Zeba Khanam and Shafali Agarwal, in their paper [3], explore large scale data processing using MapReduce and its various implementations to facilitate the database, researchers and other communities in developing the technical understanding of the MapReduce framework. They continue with exploring different MapReduce implementations; most popular Hadoop implementations and other similar implementations using other platforms and compare those based on different parameters.

The authors A. Elsayed et al., in their paper [4], look back to the MapReduce and try to find out the strengths and weaknesses, dealing with failures and enhancements that could be made to it. Furthermore, they argue that MapReduce doesn't show the expressiveness of query languages like SQL and it needs improvement of limitations such as collocation of related data, implementing efficient iterative algorithms, and managing skew of data.

Another study shows an attempt to analyze user data with MapReduce in real time [5]. The authors Ian B. and Joe D., in their paper show a system which uses the information state collected during a person-machine conversation and a case-based analysis to derive preferences for the person participating in that conversation. They use MapReduce in their processing model to achieve a near real – time generation of user preferences regardless of total case memory size.

Authors Michael T. G. et al., in their paper [6], study the MapReduce framework from an algorithmic standpoint and demonstrate the usefulness of approach by designing and analyzing efficient MapReduce algorithms for fundamental sorting, searching, and simulation problems.

In time when not only big data but also fast data are exploded in volume and availability, authors Wang L. et. al. in their paper [7], address the key challenges that MapReduce is not well suited for and provide solutions with MapUpdate use which is a framework like MapReduce and specifically developed for fast data.

Into the researches [8]-[10] are analyzed the MongoDB, NoSQL databases and reasoning behind choose of them, query optimizations, and comparisons between NoSQL and SQL databases are shown.

Authors Shuai Z. et al., in their paper [11], analyze the MongoDB clusters and introduce how to partition spatial data to distributed nodes in the parallel environment, using its spatial relationships between features.

Mohan and Govardhan, in the papers [12], [13], have analyzed MapReduce as a paradigm and combine it with online aggregation used in MongoDB. Online aggregation, according to them is useful when the data collected from massive clusters and can be very advantageous when the data are collected and estimated from sensors, various social media or Google search. Combining those two area (MapReduce and Online Aggregation) they introduced a new methodology that uses MapReduce paradigm along with online aggregation.

Dede et al. in their paper [14], have evaluated the combination of the MapReduce capabilities of Hadoop with the schema – less database MongoDB, as implemented by the mongo – Hadoop plugin. This study provides insights into the relative strengths and weaknesses of using the MapReduce paradigm with different storage implementations, under different usage scenarios. They have concluded that, in general, if the workload uses MongoDB as a database that needs to be occasionally used as a source of data for analytics then MongoDB is appropriate solution, however, it is not appropriate when using MongoDB as an analytics platform that sometimes must act like a database. Also, they show that using Hadoop for MapReduce jobs is several times faster than using the built-in MongoDB MapReduce capability and this is due to Hadoop file management system (HDFS).

III. MONGODB

Document oriented databases are designated to work without of use of SQL, and instead of it, they use a different language to communicate. A document can have any number of fields listed in any order, like in a relational database. Unlike to relational databases, a row inside a document oriented database, not need to have the same number and types of fields as any other row inside the database. This is due to the fact that there is no schema that restricts a row to be identical in number and the sequence of fields. While there are many document based databases, MongoDB stands out due to its high performance and ease of setup.

MongoDB as document based database uses BSON to store the data, which is the binary – encoded serialization of JSON format. JSON currently supports the following data types: string, number, Boolean, array and object. BSON supports: string, int, double, Boolean, date, byte array, object, array and others. BSON's only restriction is that data must be serialized in little – endian format. Since BSON is a format that the data are sent/retrieved and stored, there is the need of decode those to text. In an analogy with the relational database, a table into MongoDB is a collection of the documents and a database is a group of collections. A document is the most basic entity where MongoDB stores information, similar to a row inside a table in relational database, except the data structure is schema – less. One of the best features of a document is that it may contain other documents embedded inside.

Indexes in MongoDB work almost the same as in relational databases. MongoDB uses B-Tree to implement the indexes and also allows two – dimensional geospatial indexing which is very useful when dealing with location based services.

A. Sharding

The problem of huge amount of data, MongoDB solves in an effective fashion with use of the horizontal data distribution, known as horizontal scaling. Horizontal scaling is shown as very well solution and means a distributed and balanced work across the machines. This way of work in MongoDB is known with the name sharding. Sharding in MongoDB is designed to partition the database into smaller pieces accommodated to different machines, so that no single machine has to store all the data or handle with all the load. MongoDB handles sharding very easily and transparently which means that the interface for querying a sharded cluster is exactly the same as the interface for a single MongoDB instance.

Usually, there are collection which needs to be together and the others which allow or might be need to be distributed across some machines. So, no all collections need to be sharded, but only some collections that need data to be distributed over some shards to improve read and/or write performance. All un – sharded collections will be held in only one shard that is called primary shard (e.g., Shard A in the Fig. 1). The primary shard can also contain sharded collections.

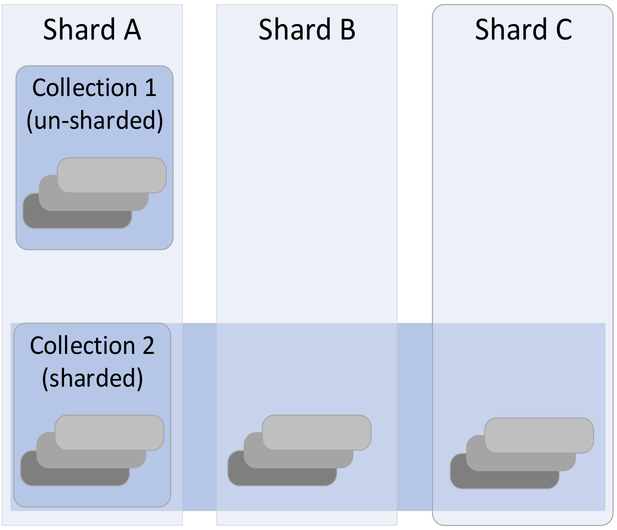


Fig. 1. Example of sharding a collection across multiple shards.

In case of more complex application, we should shard only the collections that would benefit from the added capacity of sharding while leaving the smaller collections unsharded for simplicity. Because sharded and unsharded collections are possible to be accommodated into a same system, all of this will work together, completely transparently to the application. In fact, if later we find that one of the collections that is not sharded, becomes larger and larger, we can shard it, so, it is allowed, at any time, to enable the shard and make a sharding [15].

Manual sharding can be done with almost any database software. Manual sharding is when an application maintains connections to several different database servers, each of which are completely independent. The application manages storing different data on different servers and querying the appropriate servers how to get data back. This approach works well, but there are difficulties when adding or removing nodes to/from the cluster is needed or in face of changing data distributions or load patterns.

MongoDB supports autosharding, and by use of this tries to avoid the abstract architecture from the application and simplify the administration of such a system. MongoDB allows to application to ignore the fact that it isn't talking to a standalone MongoDB server, to some extent. On the operations side, MongoDB automates the data balancing across shards and makes it easier to add and remove capacity.

A sharded cluster consists of shards, mongos routers, and config servers, as shown in Fig. 2.

B. Shared Key

To shard a collection, we have to choose at least one field which will be used to split up the data. This field(s) is called a shard key. In case, when there are a few shards, it's almost impossible to change the shard key, so, it is important to choose a correct one. To choose a good shard key, a good knowledge of the workload and how the shard key is going to distribute the application's requests are needed. And it is often difficult to imagine.

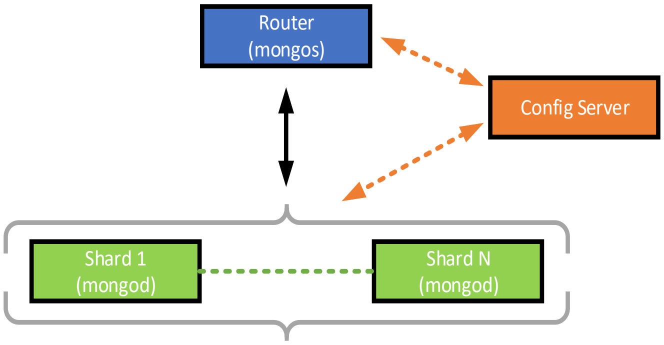


Fig. 2. Components in a Mongodb sharded cluster.

There are three most common distributions ways of splitting the data, which are: ascending key, random, and location – based. Also there are other types (with other key types) but most of those fall into one of the mentioned categories:

Ascending key distribution: The shard key field is usually the data type of Date, Timestamp or ObjectId. With this pattern, all writes are routed to one shard. MongoDB keeps distribution and spends a lots of time migrating data between shards to keep data distribution relatively balanced across the shards. This pattern shows weaknesses in the write scaling.

Random distribution: This pattern is more appropriate in case of when the fields (taken for shard key) do not have an identifiable pattern within dataset. For example, if shard key includes any of the following field username, UUID, email address, or any field which value has a high level of randomness. This is a preferable pattern for write scaling, since it enables balanced distribution of write operations and data across the shards. However, this pattern shows weak performances in case of query isolation, if the critical queries must retrieve large amount of "close" data based on range criteria in which case the query will be spread across the most of the shards of the cluster.

Location – based distribution: The idea around the location-based data distribution pattern is that the documents with some location – related similarity will fall into the same range. The location related field could be postal address, IP, postal code, latitude and longitude, etc.

MongoDB supports three types of sharding strategies:

Range – based sharding: MongoDB divides dataset into ranges determined by the shard key values.

Hash – based sharding: MongoDB creates chunks via hash values it computed from the field's values of the shard key. In general, range – based sharding provides better support for range queries that need query isolation while the hash – based sharding supports more efficiently write operations.

Tag – aware sharding users associate shard key values with specific shards. This type of sharding is usually used to optimize physical locations of documents for location – based applications.

On the below table (Table I) is shown the guidance how to select the shard key.

TABLE I. KEY CONSIDERATIONS FOR A SHARD KEY SELECTION REGARDING THE QUERY ISOLATION AND WRITE SCALING REQUIREMENTS

Query isolation importance	Write scaling importance	Shard Key Selection
High	Low	• Range shard key. • If the selected key does not provide relatively uniformly distribution of data, we can either use a compound shard key or add a special purpose field to our data model that will be used as a shard key. Or for location – based applications we can manually associate specific ranges of a shard key with a specific shard or subset of shards.
Low	High	• Hashed shard key with high cardinality. • If a selected key does not provide relatively uniformly distribution of data, we can add a special purpose field to our data model that will be used as a shard key.
High	High	• A shard key enabling mid – high randomness and relatively uniformly distribution of data. • Determine which shard key has the less performance effect on the most critical use cases. • Special purpose field to our data model that will be used as a shard key.

C. MapReduce

MapReduce is a programming model which is capable to process a huge amount of data with a parallel and distributed algorithm on a cluster. It is a programming paradigm that allow for massive scalability across hundreds or thousands of servers in a Hadoop cluster. MapReduce also is a powerful and flexible tool for aggregating data, solves some problems that are too complex to express by use the aggregation framework's query language. In our case we use MapReduce with JavaScript as its "query language" to express arbitrarily complex logic.

MapReduce processes different problems across large datasets using a large number of computers (or computing nodes) in parallel. Basically, it takes a set of input key/value pairs and produces a set of output key/value pairs [15] and this operations is executed in three steps: Map is the first step, takes the input pairs and to each node applies the "map" function and finally writes the temporary output. To prevent same data being processed a master node ensures that only one copy of redundant input data is processed. Shuffle is the second step, where the shards redistribute that data based on the output keys and reaches a stage that all data with the same key value belonging to the same shard. And finally, reduce is the final step which takes the shuffled data and processes each group of data per key.

MapReduce uses a finalize function to clean the temporary results and to manipulate with the MapReduce output, which are given from the last reduce phase. The finalize function is called before the MapReduce output is saved to a temporary collection. Returning large result sets is less critical with MapReduce, so the call of the finalize function is a good chance to take averages or remove the temporary or unnecessary information in general [16]. MongoDB allows the user to define which shards will execute the map function, the shuffle and reduce and also we can use the same shards for map function execute and as well as reducer function or define other shards that will do that job.

By default, MongoDB creates a temporary collection while MapReduce processes with the data and the temporary collection name is unlikely chosen from a collection name, but, it is a dot – separated string containing MapReduce, the name of the collection which is in MapReduce process, a timestamp, and the database job's ID. It looks something like mr.geonames.1525765769.2. MongoDB automatically destroy this temporary collection when the job is finished and /or MapReduce connection is closed. To keep the temporary collection after the job finishing and connection closed we have to set keeptemp in true as an option parameter. In case that the temporary collection is used, MongoDB allows naming the output collection with the out part option, which is a combined name and out string. To address the last issue MongodB contains an optional parameter called as out and which needs to be set to true, if out parameter is set to true, then there is no need to specify keeptemp, since it is implied. Even if a name for the temporary collection is specified, MongoDB again uses the autogenerated collection name for MapReduce further intermediate steps. When the computations have finished, the temporary collection automatically and atomically will be renamed from the autogenerated name to our set or chosen name. This means that if MapReduce is run multiple times with the same target collection, it will never use an incomplete collection in performing operations. The output collection created by MapReduce is a normal collection, which means that there is no problem with doing a MapReduce on it or a MapReduce on the results from that MapReduce.

IV. MAPREDUCE PERFORMANCE ANALYSIS

To analyze the MapReduce performances, used in MangoDB circumstances, we have created a mini cluster of few virtual servers on which is run MongoDB and the geonames database. Geonames database is an open source database and is taken as an example. Geonames database contains detailed information to world countries such are population, size, geolocation, rivers, villages, capital, etc. [17]. It contains around 11 Million records, rendered on tab separated text file. To manipulate on a better way, we converted those data to a csv format, that could easily be exported to mongo. We scaled down the database only to documents with population larger than zero and the number of those documents was 469660. From the csv file we took into consideration and imported only geonameid as id, asciiname, country and population.

Next, we built a sharded cluster to which was run MongoDB 3.2 under Ubuntu 16.04, based on Fig. 2, through three Virtual Machines, named as mongo-c1, mongo-c2, and mogno-c3, one VM for the configuration server and one query server VM. We indexed the id with "hash" that allows us to create a shard key with the hashed id which makes sure the equally distribution of our geoname collection documents. The hostnames and ip addresses of each VM was set as follow:

- mongo-config: 192.168.157.132
- mongo-router: 192.168.157.130
- mongo-c1: 192.168.157.129
- mongo-c2: 192.168.157.128
- mongo-c3: 192.168.157.131

In our tests a simple map function was set, which finds the country code and returns a value of 1, and reducer function which iterate through the values to count the number of documents in the collection which belongs to each country. The number of documents included in our tests was 469660 and the id was used as a shard key to shard the documents to the different shards.

On the above-mentioned architecture, we executed three tests. In our first test we used only one shard (mongo-c1 was used). The number of documents was 469660 and the total import time was 11.28. In the second test, we used the same number of documents but sharded into two shards (in this case was added the second shard mongo-c2). The total import time in this case was 08.25. The collection was sharded successfully and after sharding the achieved distribution was as follow: into first shard (mongo-c1) 234349 documents and into second shard (mongo-c2) 235311 documents. And in our third test we included another shard (mongo-c3), the same number of documents was included but distributed into three shards. For this case the total import time was 8.47 and the collection was successfully sharded as follow 156646 documents into first shard (mongo-c1), 156693 documents into second shard (mongo-c2) and 156321 documents into third shard (mongo-c3). We executed the same MapReduce job (with the same map and reduce functions) three times to each test and the results are shown on Table II.

TABLE II. MAPREDUCE JOB EXECUTION TIME EXECUTED ON ONE, TWO AND THREE SHARDS USED. NUMBER OF DOCUMENTS USED IS 469660

Execution		1	2	3	Average time
Num. Of Shards	Import time				
1	11.280	9.470	9.238	9.878	9.529
2	8.250	5.773	5.771	5.800	5.781
3	8.470	4.848	4.874	4.922	4.881

To better express the dependence between MapReduce job execution time and the number of nodes used, so the dependence on the number of shards to which the documents are distributed, on Fig. 3, the curve which clearly expresses the decrease of the time with increasing the number of shards is shown.

Next, we again performed the last test, but this time with a little complex shard key. We chose the pair (id, population) as a shard key. Total import time was 8:08. Since the shard key cannot be changed afterwards, so, we drop the before collection shards and recreate a new shard by use of the new shard key. By use of the new shard key the sharding was 349699 documents to the first shard, 119947 documents to the second shard and only 14 documents to the third shard. So, it is produced ununiform distribution. We executed the same MapReduce job as in previous tests and the results are shown on the following table (Table III).

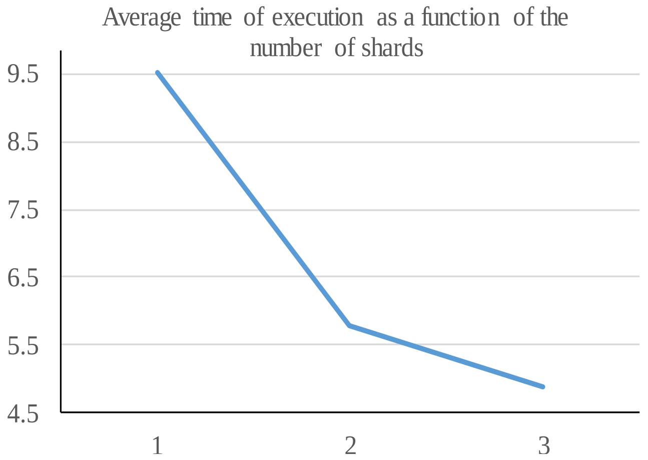


Fig. 3. Average time of mapreduce job execution per number of shards. executed on 469660 documents..

TABLE III. MAPREDUCE JOB EXECUTION TIME EXECUTED ON THREE SHARDS USED. NUMBER OF DOCUMENTS USED IS 469660, SHRD KEY (ID, POPULATION)

Execution		1	2	3	Average time
Num. Of Shards	Import time				
3	8.470	6.685	6.841	6.638	6.721

Regarding to the above tests, clearly we can conclude that when the number of shards increases MongoDB MapReduce performs better and faster. The only trouble as shown in last test is that we should be very precautious in chose of the shard key, so, we need to choose an appropriate (a good shard key which provides as far as possible uniform documents distribution) that will not slow down MapReduce.

V. CONCLUSIONS

Big data has indeed reshaped the way we deal with data. The problems that arise when trying to manipulate huge amounts of data are growing every day and solutions are found from both scientists and companies alike. MongoDB and other NoSQL databases alike has seen growth by providing an alternative to SQL databases. Their design, high availability and fault tolerance have attracted usage in projects where SQL databases cannot be used such as handling unstructured raw huge amount of data.

MapReduce as a framework is designed to solve many problems with huge amount of data, so, MapReduce has a little significance when dealing with small data, but it has an impact when the collections grow. Also it is clear and our tests show that as the number of shards scales up, MapReduce jobs are executed faster especially if we take precautions and use a good shard key. However, the Mongo 3.2 documentation [18] recommends the avoid of the MapReduce use and instead of MapReduce the Aggregation Pipelining is preferred for better performance.

References

[1] Jeffrey Dean and Sanjay Ghemawat. 2008. MapReduce: simplified data processing on large clusters. Commun. ACM 51, 1 (January 2008), pp. 107-113

[2] Smita Agrawal, Jai Prakash Verma, Brijesh Mahidhariya, Nimesh Patel and Atul Patel. Survey on Mongodb: An Open-Source Document Database. International Journal of Advanced Research in Engineering and Technology, 6(12), 2015, pp. 01-11.

[3] Zeba Khanam and Shafali Agarwal, MapReduce Implementations: Survey and Performance Comparison, International Journal of Computer Science & Information Technology (IJCSIT) Vol 7, No 4, August 2015, pp. 119 - 126

[4] A. Elsayed, O. Ismail, and M. E. El-Sharkawi, MapReduce: State-of-the-Art and Research Directions, International Journal of Computer and Electrical Engineering, Vol. 6, No. 1, February 2014.

[5] Beaver, I. and Dumoulin, J., Applying mapreduce to learning user preferences in near realtime. In: Case-Based Reasoning Research and Development, ICCBR 2014, Berlin, Springer (2014) pp. 15–28.

[6] M. T. Goodrich, N. Sitchinava, and Q. Zhang. Sorting, searching, and simulation in the MapReduce framework. Proceedings of the 22nd international conference on Algorithms and Computation (ISAAC'11), 2011, Takao Asano, Shin-ichi Nakano, Yoshio Okamoto, and Osamu Watanabe (Eds.). Springer-Verlag, Berlin, Heidelberg, pp. 374-383.

[7] W. Lam, L. Liu, S. Prasad, A. Rajaraman, Z. Vacheri and A. Doan, Muppet: MapReduce-style processing of fast data, Proceedings of the VLDB Endowment, Volume 5 Issue 12, August 2012, pp. 1814-1825

[8] Freire, S., Teodoro, D., Wei-Kleiner, F., Sundvall, E., Karlsson, D. and Lambrix, P. (2016). Comparing the Performance of NoSQL Approaches for Managing Archetype-Based Electronic Health Record Data. PLoS ONE 11(3): e0150069. https://doi.org/10.1371/journal.pone.0150069.

[9] Marwa, E. and Jemili, F., Using MongoDB Databases for Training and Combining Intrusion Detection Datasets. International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, SNPD 2017. pp. 17-29, Studies in Computational Intelligence, vol 721. Springer, Cham.

[10] Parker, Z., Poe, S. and Vrbsky, S., Comparing NoSQL MongoDB to an SQL DB. In Proceedings of the 51st ACM Southeast Conference (ACMSE '13). ACM, New York, NY, USA, Article 5, 6 pages. 2013, DOI: https://doi.org/10.1145/2498328.2500047.

[11] Shuai Z., Bolei Z., Zhenjie C., Sanglu L, "Point collection partitioning in MongoDB Cluster", Research Foundation of Graduate School of Nanjing University (2013CL09), http://www.geog.leeds.ac.uk/groups/geocomp/2013/papers/97.pdf

[12] B. Rama Mohan, A Govardhan, Online Aggregation Using MapReduce in MongoDB, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 9, September 2013, pp. 1157-1165

[13] B. Rama Mohan, A Govardhan, Sharded Parallel Mapreduce in Mongodb for Online Aggregation, International Journal of Engineering and Innovative Technology (IJEIT), Volume 3, Issue 4, October 2013, pp. 119-127

[14] Elif Dede, Madhusudhan Govindaraju, Daniel Gunter, Richard Shane Canon, and Lavanya Ramakrishnan. 2013. Performance evaluation of a MongoDB and hadoop platform for scientific data analysis. In Proceedings of the 4th ACM workshop on Scientific cloud computing (Science Cloud '13). ACM, New York, NY, USA, pp. 13-20.

[15] Li, Feng & Chin Ooi, Beng & Özsu, M. Tamer & Wu, Sai. (2013). Distributed Data Management Using MapReduce. Journal of ACM Computing Surveys (CSUR) Volume 46 Issue 3, January 2014, Article No. 31.

[16] Kyle Banker, MongoDB in Action, 2nd ed., Manning Publications Co. 2016, pp. 333-375.

[17] http://www.geonames.org/, Retrieved 2018-01-31

[18] MongoDB. https://docs.mongodb.com/manual/ , Retrieved 2017-12-16

Demand Response Programs Significance, Challenges and Worldwide Scope in Maintaining Power System Stability

Muhammad Faizan Tahir, Chen Haoyong*, IEEE Senior Member, Idris Ibn Idris, Nauman Ali Larik
School of Electric Power, South China University of Technology, Guangzhou, P.R. China

Saif ullah Adnan
School of Electronics and Information Engineering, South China University of Technology,
Guangzhou, P.R. China

***Abstract*—In order to cope up the continuously increasing electric demand, Governments are forced to invest on Renewable Energy (RE) sources due to scarcity of fossil fuels (such as coal, gas and oil), high costs associated with it and emission of greenhouse gases. However, stochastic nature of RE sources like wind and PV threaten the reliability and stability of power system. Demand Response (DR) is an alternative solution to address the issues of economic constraints, integration challenges of RE, and dependency on fossil fuels. It is an aspect of Demand Side Management (DSM) that converts consumer's passive role to active by changing energy consumption pattern to reduce peak load. DR plays the role in deferring the investment on building new power plants, eliminating transmission losses and making the society green. This work analyzes initialization of different DR programs due to slumping technology costs and recognition of users' behavior in electricity market. Moreover, this paper points out the problems associated with DR and its project implementation across USA, China and developed cities of Europe.**

***Keywords*—Demand side management (DSM); demand response (DR); renewable energy (RE); DR programs; wind; PV**

I. Introduction

DSM also known as energy demand management was first coined in 1973 due to high prices of fuel and energy crisis [1], [2]. It was first premised by Electric Power Institute in 1980s. It is the process of implementing, planning, monitoring and controlling user end activities to match the balance between supply and demand [3], [4].

Other than increasing power generation to meet up the increase load, demand can be controlled at user end by demand response, energy efficiency and load conservation techniques. Peak reduction, shifting peak load to valley hours and turning on users own generating units are popular DR techniques and these techniques helps in deferring the installation of new power plant, decreasing peak demand and improving the load factor. In addition, Energy Efficiency (EE) methods encourages the customer to use energy efficient devices that consume less power such as Compact Fluorescent (CFL) lights. Strategic load conservation ideas like building home in such a way that it requires less cooling during summer and less heat during winter play the vital role in reducing electric demand at user end [5]-[7]. However, strategic load growth like electrification increases the energy consumption but its objective is to increase electricity sale and local resource consumption to find the alternate of fuel. Flexible load curve gives the option that load can be interrupted by grid operators when needed to reduce peak demand but varying reliability and quality of service. All these DSM techniques not only serve to reduce power consumption but also put its effort in building green society [8]-[10].

Making the society green requires the abundant use of RE sources such as PV and wind. However, integration of these stochastic sources threatens the grid reliability. Consequently, DR proves to be effective alternate to supply reserves, reducing peak and providing other ancillary services to mitigate the integration challenges. Increasing trend of RE generation till 2017 and DSM classification are shown in Fig. 1 and 2, respectively.

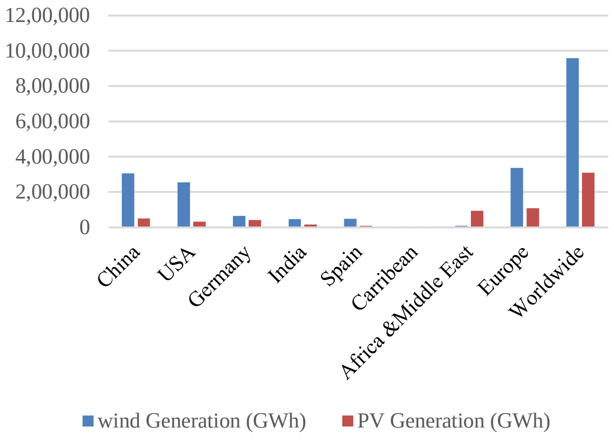


Fig. 1. PV and wind generation (GWh) till 2017.

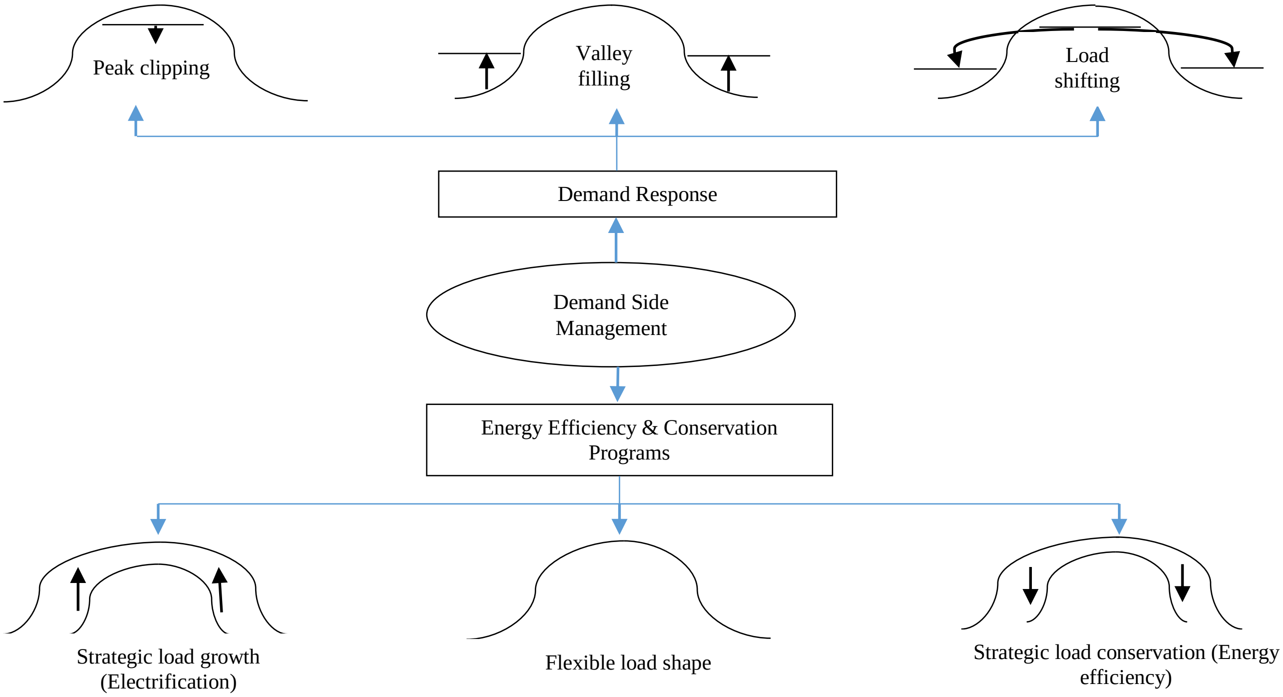

Peak clipping
Valley filling
Load shifting
Demand Response
Demand Side Management
Energy Efficiency & Conservation Programs
Strategic load growth (Electrification)
Flexible load shape
Strategic load conservation (Energy efficiency)

Fig. 2. Classification of DSM explaining load shaping techniques.

A. *Need of Demand Response*

With each passing year the power consumption of residential, commercial and industrial users is increasing exponentially. It has been estimated that till 2030 power demand will be increased up to 40 percent. In order to overcome this problem the first solution come in mind is to generate more power. Amount of power generation can be increased by installing more small, medium or large size power plants in proportion with the increase in demand. More fossil fuels generating units will be required which will produce greenhouse gases and pollute the environment that is already suffering with extreme global warming issues [11]. Furthermore, after power generation, distribution and transmission networks need to expand. These all set ups from generation to transmission to distribution not only increases the cost but also increases the occurrence of unwanted events like brown outs or black outs [12]. Our present grid system are prone to faults and often suffer with problem of brown outs when load increases beyond the peak load. Third world countries that do not have enough sources to catch up the demand often have to shed the load. In case of failure of shedding proper amount of load power system leads to cascaded failure that causes the whole grid to black out [13]. Number of cases of these black outs has been increased significantly since past few years that effected millions of consumers.

The effective and less costly solution other than firing up the generating plants is to control or reduce the demand. There have been few hours in a day where demand shoots up and this time period is known as peak hours or peak period. During these peak hours if consumer shifts their load to off peak hours or reduce their energy consumption by using devices that consume less power: system will not suffer the problem of reliability. Moreover the need to build new generating plants will be reduced that will save a lot of capital cost. In order to achieve this, continuous monitoring of power consumption should be noted that needs two way communication between utility operators and consumers. The different possibilities of coping up increase in demand are shown in Fig. 3:

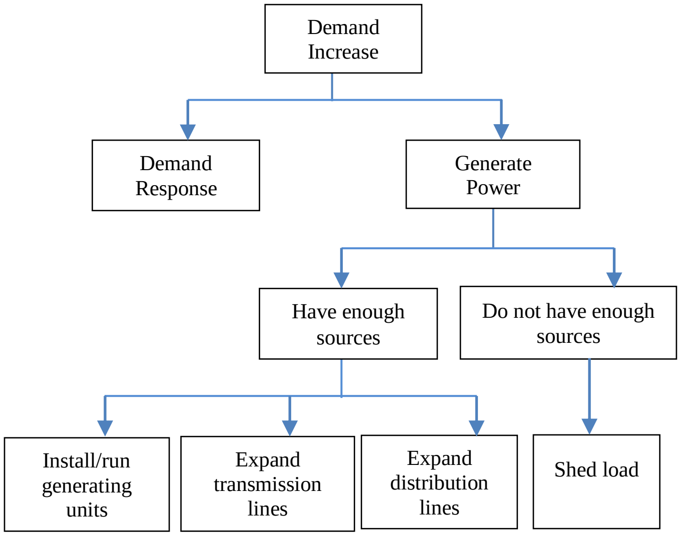

Demand Increase
Demand Response
Generate Power
Have enough sources
Do not have enough sources
Install/run generating units
Expand transmission lines
Expand distribution lines
Shed load

Fig. 3. Different schemes to meet the increase demand.

The popular method of reducing load from 1970s to 1990s was curtailable and interruptible programs but customers were never exposed to real-time whole price signals. Integration of more renewable energy sources in the electric grid causes uncertainty in electric supply [14], [15]. This forces the utility managers to work on demand response programs that base on price response in order to convince customers to participate in these programs. Rest of the paper is organized as follows: Sections 2 and 3 enlist customers load reduction methods and DR programs respectively while Section 4 elaborates DR measurement phenomena by calculating self and cross elasticities. Section 5 elaborates the merits and demerits of DR schemes. Section 6 analyze the DR significance across USA, Europe and China and also identify technical barriers to implement these programs. Section 7 concludes the paper and also discusses the future work.

II. Customers Load Reduction Methods

DR can be done at three levels.

A. *Residential level*

B. *Commercial level*

C. *Industrial level*

There are three ways of reducing the load for residential customers during peak hours [15], [16].

- First one is to completely shed the load during peak hours when price is considerably high but it will involve the loss of comfort. Turning off the AC, heaters, dryers or other household appliances during these hours helps to reduce peak demand.
- Second one is not to shut the high consuming devices but reduce its consumption level such as turning on the AC at 26 or 27 instead of 18 will reduce the electricity consumption.
- Third option is to shift load to off peak hours. Daily house hold activities like washing clothes, dryers and cooking chores can be shifted from afternoon or evening to night when power consumption is relatively less.

Prior to save electricity and reduce the monthly bill; residential customers also receive special incentives from electric suppliers in participating demand management.

Unlike residential customers it is difficult for industrial customers to shed load completely. This entails commercial customers to reduce load and provide stability to power system when demand increases. Demand can be mitigated considerably without being effected too much by just dimming the lights of lobby of big hotel for 20 minutes or changing the thermostat setting of freezers of big plazas.

The best way for industrial users are to shift load on their own power generating units because industry will not able to bear the loss of shutting or reducing the load. These two customers can save electricity costs by DR and can earn back as much as 5 to 25%. Prior to the customers benefit to opt DR programs, utilities also has their reasons in convincing customers to participate in these programs [17].

III. Demand Response Programs

There are many demand response programs [18]-[22] and its types and sub types are shown in Fig. 4.

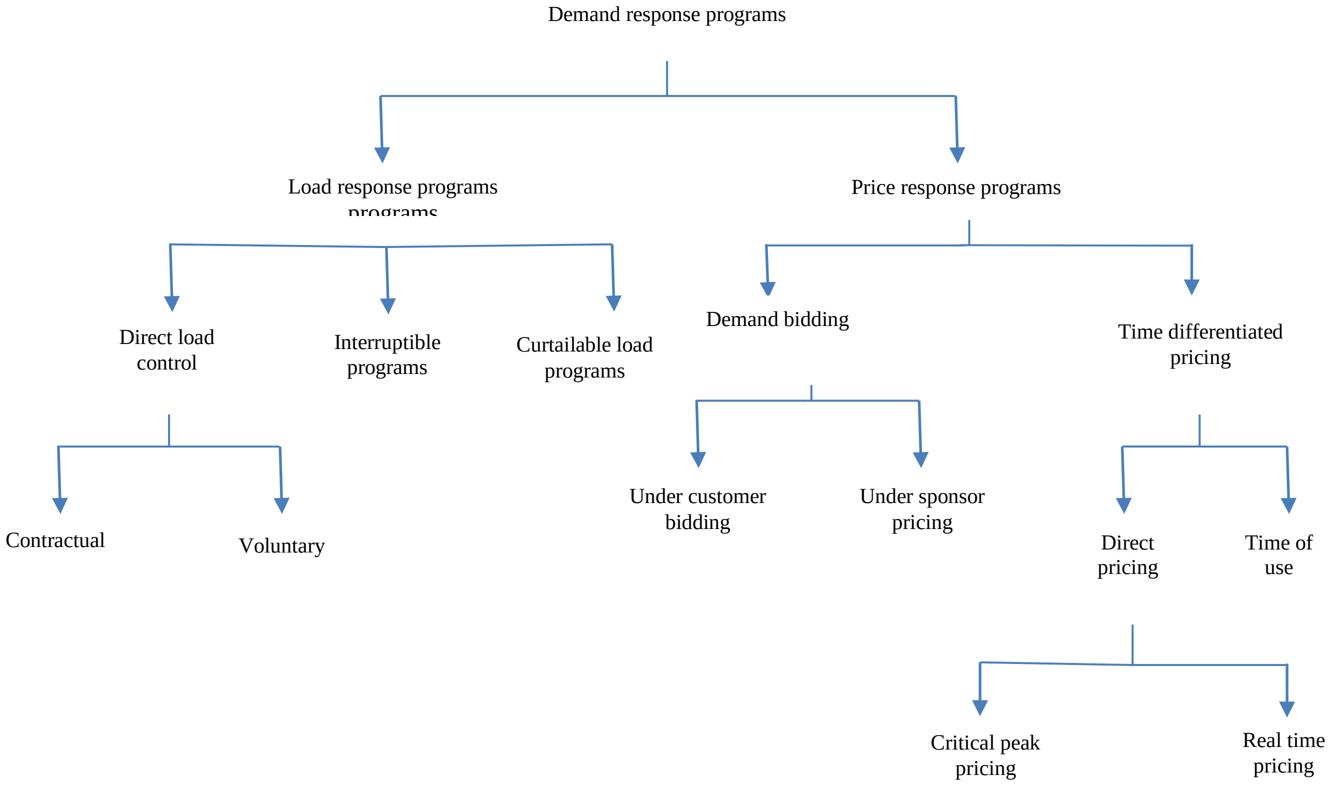


Fig. 4. Different schemes to meet the increase demand.

A. Load Response Programs

This type of demand response also called event based program, reliability based or dispatchable. In this kind of program, some agreement takes place between customers and utility. Some incentive or reduction in bills are offered to customers by reducing load when utility asked them to do so. Grid operators install some control technologies in consumer's premises to control and monitor their electricity usage. This load response program can be referred as contractual and voluntary load programs.

- According to contractual load programs agreement customers must curtail their load during peak hours and receive guaranteed payment otherwise being penalized.
- However, in voluntary programs agreement customer decides by himself when and when not to reduce or shed the load.

This kind of incentive based scheme sometimes also known as explicit demand response.

1) Direct load control: As the name indicates in direct load control program utility directly control the customer's energy consumption. Grid operators install some remote control devices such as wireless communication, wired communication, radio control, GSM control and so on. With the help of these devices they can remotely monitor and control the consumer's appliances. Mostly this type of program is suitable for residential customers.

2) Interruptible programs: This type of demand response mostly targets big commercial and industrial customers. These customers can shut down electricity for short time interval or can switch back to their own generating units. Participating in this program customers receive electricity rates that are much lower than industrial rates and these rates are usually known as "interruptible rates".

3) Curtailable load programs: Those customers that cannot shut their supply can participate in DR programs by just reducing their load and this type of program is known as curtailable load programs. Mostly industrial and commercial customers signed to this program and they are generally notified about shedding the load between 30 minutes to two hours.

B. Price Response Programs

Price response program also known as market-based, non-dispatchable or non-event based programs is totally based on voluntary action of customers. Showing willingness to take part in this program customers are offered some economic incentive or pricing choice. Generally, this scheme focuses in reducing the wholesale market price.

This price based voluntary scheme also refers as implicit demand response.

1) Demand bidding programs: Demand bidding or buy-back programs that sometimes also known as economic response programs mostly targets commercial and industrial facilities. These customers can minimum shed load of 100 kW per event. This program further divides in two branches depending upon how bid is structured.

a) Under Customer Bidding

In this case, customers bids the price that is lower than market price for reducing specific load at specific time at most a day ahead or in some cases an hour ahead.

b) Under Sponsor Pricing

In this scheme, Customers are being notified the price of per kWh of load reduction by the market administrators. Upon showing willingness to this method customers receive reduced electricity price depending on how much reduction in load occurs.

2) Time differentiated pricing: Electricity suppliers expose customers to time-varying electricity prices that show the price of electricity at different time periods. It may vary from flat day and night price to high dynamic price depending on hourly wholesale prices. Therefore customers can shift load from high prices intervals to low prices intervals.

Electricity price does not remain constant; it varies significantly according to months of the year, days of the week and hours of the day. During peak periods the market prices are considerably high as compared to off peak periods. There are several structures for this scheme that are mentioned below:

a) Time-of-Use Rates

Time-of-use rates do not follow the single flat rate for energy consumption instead prices are high when electric demand is higher. Usually in summer period, afternoon 6 hours are considered as peak hours while remaining other hours are considered off peak hours. TOU is applied to these two blocks of hours where price is predetermined and remains constant. This program gives consumers chance to reduce electricity bill by shifting load from peak hours to mid peak hours or valley hours.

b) Dynamic Pricing

In contrast to TOU rates in which electric suppliers gave forecast of load a day ahead but in dynamic pricing it is as closely correlated as one or two hours ahead. It reflects hourly variation of wholesale market prices. Increasing trend of smart grids that includes smart meters, advanced communication and control technologies creates novel options for dynamic rate structures. The possible dynamic pricing mechanisms are mentioned below.

c) Real Time Pricing

Electricity prices not only vary weekly or daily, but also vary hourly or sub hourly. Real time pricing (RTP) reflects the prices that varies on hourly basis. It informs customers of price variation as little as of 5 minute interval.

d) 5.2.4 Critical Peak Pricing

Critical peak pricing (CPP) is a hybrid of TOU and RTP. TOU has two main standard periods on and off periods but CPP adds the third block called "critical peak period." TOU blocks have fixed price and specific time frame unlike CPP block that may or may not occur on any specific day.

This program may come in account during emergency condition in power system or when electric suppliers anticipate high wholesale electric prices. In these cases grid operators might invoke critical events during any specific time interval such as it can be from 6 p.m. to 10 p.m. on any cold day of a winter or between 3 pm to 6 pm of summer afternoon. During this critical time interval electricity prices will shoot up exponentially.

IV. Demand Response Measurement

Amount of peak reduction is an indicator for the success of DR. Besides peak reduction dynamic pricing act as an important factor to determine sensitivity of electric users to price of electricity. Users sensitivity is determined by demand price elasticity which is outlined as ratio of change in demand to ratio of change in price and is calculated as [23]:

$$E = \frac{\frac{\partial d(i)}{d(i)}}{\frac{\partial p(j)}{p(j)}} = \frac{\partial d(i)}{d(i)} * \frac{p(j)}{\partial p(j)} = \frac{p(j)}{d(i)} * \frac{\partial d(i)}{\partial p(j)} \qquad (1)$$

E = elasticity of demand

∂d(i) = Change in demand (MWHr)

d(i) = Initial demand (MWHr)

∂p(j) = Change in price ($ or RMB/MWHr)

p(j) = Initial price ($ or RMB/MWHr)

Highly elastic demand occurs when huge changes in demand results only slight changes in prices. Price demand elasticities can be divided in two types counting on users' response to price demand elasticities:

A. Self-Elasticity/Single Period Elasticity

Loads like incandescent lamps that can only be turned on and off and cannot be shifted to some other period has sensitivity during a single period only and known self-elasticity or single period elasticity and always has negative value [24]. Therefore, single period elasticity is defined as ratio of relative change in demand during t_{th} hour to relative change in its day-ahead price during same hour. It can be calculated as:

$$E(i,j) = \frac{p_o(j)}{d_o(i)} * \frac{\partial d(i)}{\partial p(j)} \qquad (2)$$

or can be expressed as:

$$E^{tk} = \frac{{p_o}^t}{{d_o}^t} * \frac{\partial {d_d}^t}{\partial {p_d}^t} \qquad (3)$$

$p_o(j)$ or ${p_o}^t$ = change in price during t_{th} hour

$d_o(i)$ or ${d_o}^t$ = initial demand during t_{th} hour

B. Cross-Elasticity/Multi Period Elasticity

Loads like processing loads that has the ability to operate in more than one mode and can switch its mode to off peak periods has cross elasticity or multi period elasticity [23], [25]. Consequently, cross-elasticity is termed as ratio of relative change in demand during t_{th} hour to relative change in its day-ahead price during kth hour and its value is always positive.

$$E^{tk} = \frac{{p_o}^k}{{d_o}^k} * \frac{\partial {d_d}^t}{\partial {p_d}^t} \qquad (4)$$

${p_o}^k$ = change in price during kth hour

${d_o}^k$ = initial demand during kth hour

It reveals that DR schemes performance are comprised of below factors.

1) *Peak demand reduction*
2) *Demand elasticity*

DR can work efficiently only when automated response technologies are enabled. The complete architecture of these technologies is shown in Fig. 5.

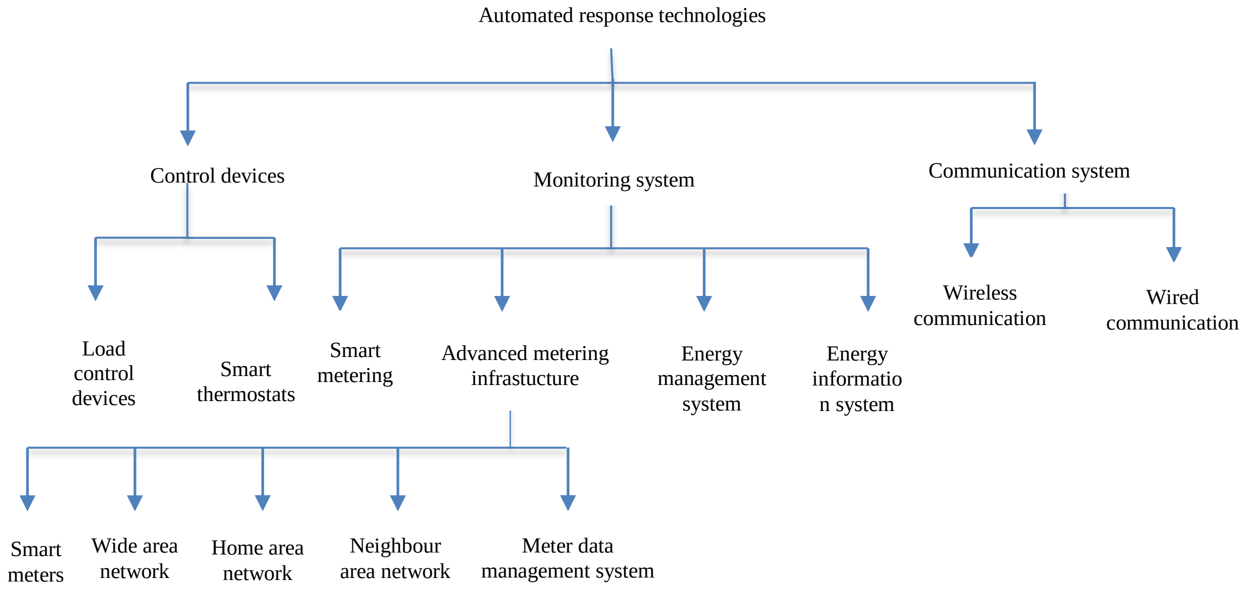


Fig. 5. Automated technologies needed to implement DR.

V. Merits and Demerits of Demand Response

Electric power system has three main features:

1) Electric energy cannot be stored economically for that demand and supply must be in balance all the time.

2) Due to the need of increase in power generation because of the continuous increase in demand, grid conditions change drastically from day to day or even hour to hour. It may cause the mismatch between supply and demand that will jeopardize the stability of system.

3) Electric grid that includes power generating units, hundreds of kilometers long transmission lines network and distribution network makes it highly capital-incentive.

These problems enhance the importance of demand response. Merits of DR can be viewed as power grid and electric customer benefits which are discussed below.

A. Power Grid Benefits

Prior to the customers benefit to opt DR programs, utilities also have their reasons in convincing customers to participate in these programs. In case of sudden increase in demand power system has to use their stand by generating units such as hot spinning reserve and cold spinning reserve. First, hot spinning reserves comes in action and if it is not sufficient to fulfill the demand then has to start the cold spinning reserves as well.

This all can be avoided just by opting to DR programs. These programs can eliminate the need of building new power plants such as in New York alone industrial and commercial customers save up to 543 MW that is about the capacity of medium size power plants.

- Increase in demand during peak period forces to run the standby generating units that may only run few hours in the entire year. This problem can be dealt easily by reducing peak demand during these hours.
- DR eliminates the problem of integration of renewable energy sources to the grid. The grid uncertainty also increased drastically due to the varying nature of these sources.
- During sudden increase in load causes the frequency to decay and if generation does not match up with load, it will cause the generators to shut down. This problem can lead to cascaded failure and whole system will suffer black out. Demand response can become important to eliminate this problem and keep the balance between demand and supply.
- Governor that control the amount of fuel in generators come in action as soon as demand increase. In case governor action is not sufficient to fulfill the load requirement power system has to switch to hot spinning reserve and then to cold spinning reserves. There will be no need of these ancillary services by following the demand response programs.
- After getting awareness of market price (by following price response program) if customers reduce or shift load, it will lower the wholesale market prices. In this scenario customers not only save bill but also help other customers in reducing their bill as well. These advantages are summarized in Fig. 6.

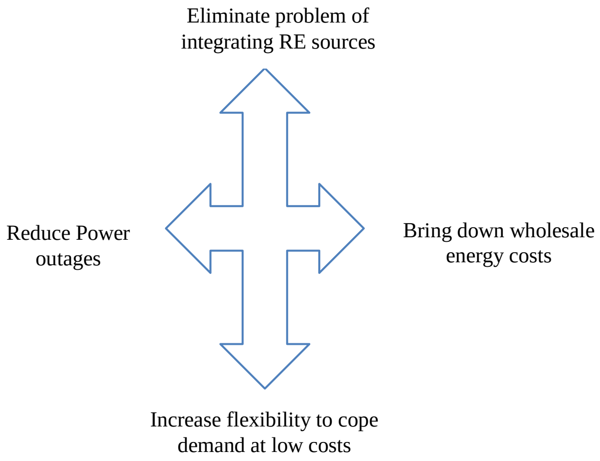

Eliminate problem of integrating RE sources
Reduce Power outages
Bring down wholesale energy costs
Increase flexibility to cope demand at low costs

Fig. 6. Power grid benefits.

B. Electric Customer Benefits

- Customers can reduce electricity bill by shutting down high load devices during peak periods.
- Number of increasing fossil fuels generating units has adverse effect on climate because of the release of greenhouse gases. Demand response helps to make the environment cleaner and healthier.
- Customers can continuously monitor consumption and prior to their own financial management can also play the role towards grid stability that reflects the positive impact on society as depicted in Fig. 7.

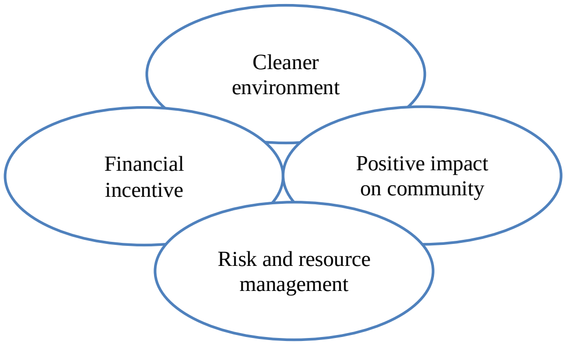


Fig. 7. Electric customer benefits.

As DR programs are gaining importance by each passing day but still there are number of problems and demerits associated with it are discussed below:

- Demand response programs need two way communication that will require some changes in current electric grid. Control, monitoring and communication devices will increase the initial cost of the utility.
- The first and foremost thing for demand response is the participation of customers. Sometimes customers are

hesitate to take part and think these programs will benefit only utility. Furthermore, they don't want to waste time to monitor and control the load.

- The current market structures lack the appropriate market mechanism. Beforehand planning for DR causes doubt in the response that can be accomplished in real time.
- Smart meters and smart thermostats are needed to install at residential customers' premises to participate in DR programs. Customers feel reluctant to share their privacy and control to utility. Furthermore they got confused that in spite of reducing the same amount of load why the price of electricity varies on daily basis?
- DR programs focuses more on commercial and industrial consumers instead of residential because of the high success and income involved with these customers.
- Few energy sellers are reluctant to participate in DR activities because it will reduce the sale of energy efficient devices.
- Monetary funds of DR programs are beyond the customers controls. Policies of contracts and incentives for DR programs may vary year to year or might eliminate it. These little uncertainties causes customers observe DR as not a "sure bet."

VI. World Wide Demand Response Experience

A. DR in USA

The invention of Air conditioning system by a New York student Willis Carrier in early 20th century turned in to central heating/cooling unit till 1970s. This central air conditioning cause the electricity demand to grow and load factor to decline. Simultaneously, oil crises and scarcity of natural gas increases the electricity price. Ultimately in 1980s oil prices collapsed and natural gas used to address the price of new capacity [26], [27]. These scenarios and "integrated resource planning" force the USA energy policy makers to work on load management and during 1980s and 1990s interruptible DR programs quiet become popular. 1992 Energy Policy Act of USA allowed independent power producer to participate in market price mechanism [28] that encourages to invest heavily on DR programs that reach up to 2.7 billion dollar nationwide in 1993. However, this value declined to half (1.3 billion dollar) till 2003 due to industry restructuring. Merely 22,904 MW peak load reduced in 2003 due to DSM activities.

Restructuring of USA electric power system during 1990s unbundled generation, transmission and distribution network. Transmission operations in many regions have evolved into Regional Transmission Organizations (RTOs) under the supervision of the Federal Energy Regulatory Commission (FERC) that manage transmission systems as well as wholesale power markets. Therefore, same restructuring or deregulation that swept during 1990s used as a mean to promote DSM activities by developing competitive market price and announcing "public benefit programs" [20].

In 2008, 38,000 MW and 2700 MW peak load reduction caused by incentive and price based DR programs respectively. Contribution of DR towards reducing load in 2010 reach up to 31,702 MW. Installation of Advanced Metering Infrastructure (AMI) has increased from 8.7% to 23% [29] within three years span (2010-2012) and the number reaches to 38 million in 2012 [30]. This number continues to grow up to 65 million in 2015 and furthermore increased integration of distributed generation push to maximize DR potential. FERC estimated that by the year 2019, amount of load reduction by DR may reach up to 138,000 MW that makes 14 percent peak demand of total load and its details are listed in Table I.

TABLE I. Potential Peak Reduction from DR Programs by North American Electric Reliability Region

RTO/ISO	**2013**		**2014**	
	Peak reduction (MW)	**Percent of peak demand**	**Peak reduction (MW)**	**Percent of peak demand**
California (CAISO)	2180	4.8%	2316	5.1%
Electric Reliability Council of Texas (ERCOT)	1950	2.9%	2100	3.2%
ISO New England, Inc. (ISO-NE)	2100	7.7%	2487	10.2%
Midcontinent Independent System Operator (MISO)	9797	10.2%	10356	9.0%
New York Independent System Operator	1307	3.8%	1211	4.1%
PJM Interconnection, LLC (PJM)	9901	6.3%	10416	7.4%
Southwest Power Pool, Inc. (SPP)	1563	3.5%	48	0.1%
Total ISO/RTO	28798	6.1%	28934	6.2%

TABLE II. PEAK REDUCTION OF DIFFFERENT DR SCHEMES

Sr. No:	DR Program	Peak Reduction (%)
1	Capacity Resource	29
2	Interruptible load	24
3	Direct load control	15
4	Time of use	12

DR in USA can be considered a source of generation as it covers 10 % demand of country. Moreover, USA is interested in promoting more of DR programs in next 5 years as only 4 types of DR schemes contributes to 80 % of peak reduction as shown in Table II.

DR offers a win-win situation both for participants and utility and is expected to become more mainstream as it saves money, reduce the need to build more power plants and reduce the CO_2 emission to protect the environment. The total peak reduction by following DR and EE techniques during 1992-2008 are summarized in Fig. 8.

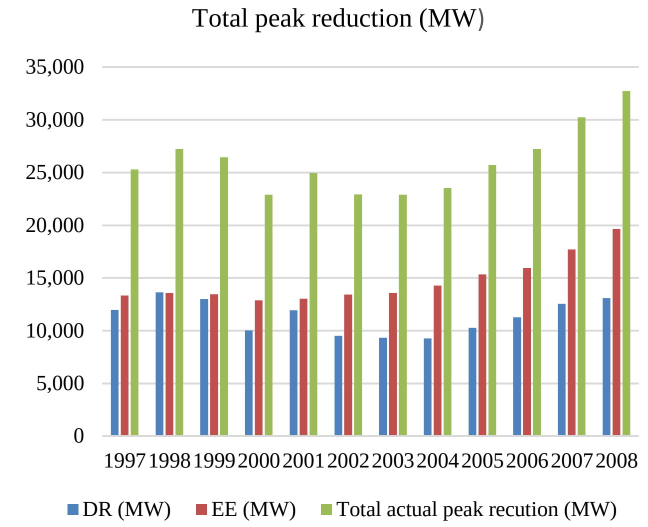


Fig. 8. Total peak reduction during 1997-2008.

B. DR in Europe

Unlike USA, summer peak periods are not a major concern in Europe. Though, during winter heating load increases electric demand but the bigger drivers are the increasing share of renewable energy such as wind and solar power. Large part of USA load management is via capacity markets but Europe has just initiated to develop the structures that allow DR resources to participate effectively [31].

Since past 20 years, electric demand reduction in Europe has been done by different forms of load shedding mechanisms [32]. These techniques did not base on precise price signals to give the dynamic pricing option to customers. In 2008, Union for the Co-ordination of Transmission of Electricity (UCTE) estimated DR forecasts in European countries that is shown in Table III. Except Germany and Hungary, DR potential continues to rise and up to 2015 and UCTE forecast has accomplished desire goal [32].

TABLE III. DR FORECASTS TILL 2020

	Year wise DR forecast (GW)				
Country	**2008**	**2010**	**2013**	**2015**	**2020**
Italy	4.00	4.00	4.00	4.00	4.00
France	3.60	3.00	3.00	3.00	3.00
Spain	2.00	2.30	2.50	2.70	3.00
Netherland	1.00	1.00	1.10	1.25	1.50
Greece	0.40	0.60	0.80	1.00	1.30
Germany	0.20	0.30	0.40	0.50	0.05
Belgium	0.20	0.20	0.20	0.20	0.20
Hungary	0.00	0.05	0.08	0.10	0.20
Monte negro	0.03	0.03	0.05	0.05	0.05
Luxemburg	0.02	0.02	0.02	0.02	0.02
UCTE countries	11.45	11.50	12.15	12.82	13.32

The most positive step towards consumer participation was the inclusion of DR in Network codes during 2014-2015. Network codes are the list of rules blueprinted by European network of transmission system operators for electricity (ENTSO-E) to facilitate integration and efficiency of European electricity market and its objective till 2020 is [33], [34].

- Reduce 40% greenhouse gas emission in comparison to 1990
- 20% involvement of RE consumption
- 27% energy saving

Summary of DR potential till to date is shown in Fig. 9:

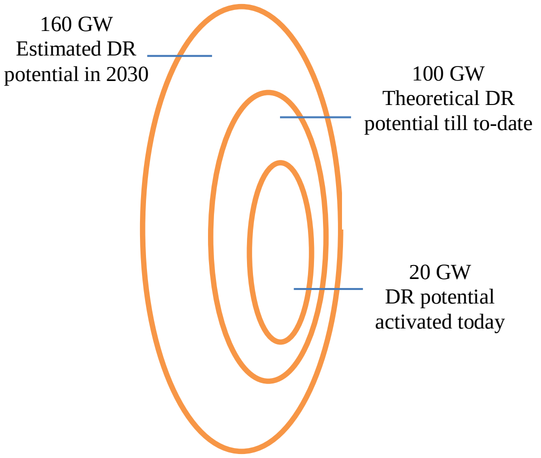


Fig. 9. DR potential in Europe.

Recently, potential of electric demand response, integration of renewable energy to match the balance between supply and demand has been discussed by Industry Research and Energy Committee (ITRE) held on 30th May 2017. Vice Chairman of the committee address the problem of continuously increasing load and said "electricity generation follows consumption" paradigm is no longer applicable [35]. This phrase emphasizes the importance of DR in addressing this critical issue across the whole Europe. Peak reduction by different companies across few developed European countries and list of technical and policy barriers in implementing DR across various countries are shown in Table IV and Fig. 10.

TABLE IV. DR FORECASTS TILL 2020

Country	Restrictions/Barriers in implementing DR
Austria	EMC (Electricity Market Code) assign strict requirements for participating in DR programs.
Germany	No penalties against producing a negative reserve
Denmark	Involvement of DR activities is unclear on balancing markets.
Finland	Participation for primary reserve is limited.
France	Aggregators can bid separately but unable to bid as one single block
Ireland	Ancillary services not available for DR activities.
Sweden	DR aggregators cannot participate independently in market.
Poland	DR programs are limited to only Emergency demand side reserve program.
Spain	Only interruptible load program is available.
Italy	Interruptible load program is available only for industrial loads.

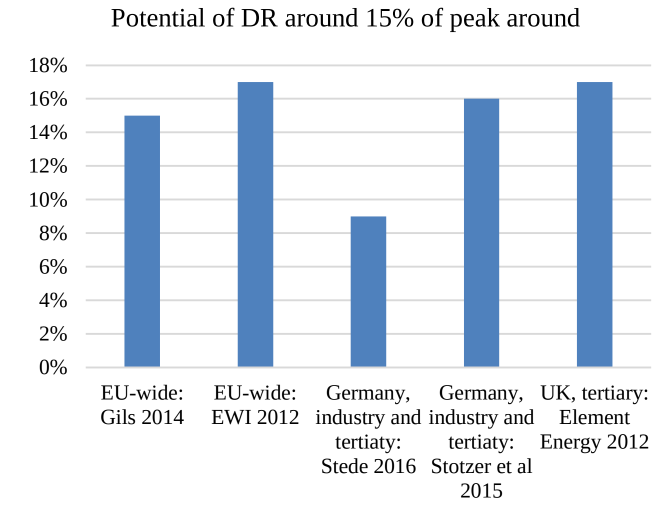


Fig. 10. Peak reduction by DR schemes across Europe.

Therefore, fostering DR is an effective way to meet the energy goals of Europe without increasing the GHG emission gases. Energy goals of Europe to use more RE and decarbonize the environment till 2030 are elaborated in Fig. 11.

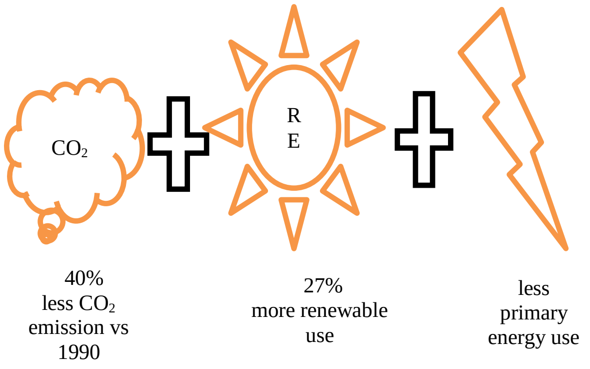


Fig. 11. Europe energy goals till 2030.

C. DR in China

China represents 44% of total world coal production and 65% of its power demand is fulfilled by these fossil fuel sources [36] as it is world's largest electricity producer and consumer. Load management started in 1990s in china but load shedding and black outs issues due to increase load during 2003-2008 fasten the implementation of DSM projects especially DR programs [37]. DR programs enlisted in Table V are actively working in China.

TABLE V. ACTIVE DR PROGRAMS IN CHINA

Price-based program	Incentive-based program	Policy-guided program
Time-of-use rates	Interruptible/curtailable load	Power rationing
Critical peak pricing (CPP)	Direct load Control	Orderly power
Two-part pricing		

Jiangsu is the first province that issue its own DSM regulations in April, 2002 [38]. Up till 2003, it used TOU scheme for industrial customers to implement DR and about 95% of users enrolled in it. Nanjing, Yangzhou and Xuzhou are the cities that actively participated in DR events. Complete list of peak reduction in different provinces during 2003 are presented in Fig. 12.

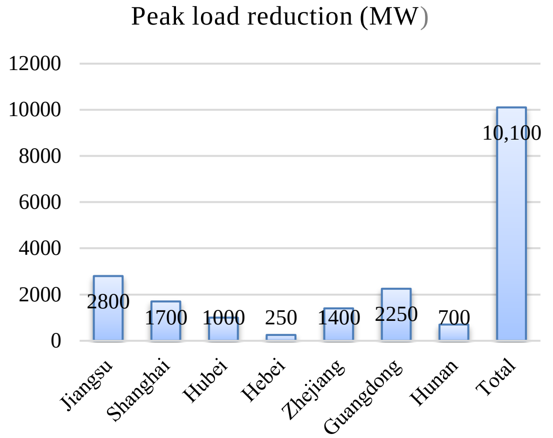


Fig. 12. Peak reduction in pilot cities during 2003.

10 GW increase in load in Foshan city (Guangdong province) in 2014 prompt to follow DR schemes and resulted in 450 MW peak reduction in following two years [39]. During 1997-2003, Beijing's DSM project focus was just to improve load factor above 80% that falls below 76% [37], [40]. However, this focus shifted to other DR schemes like increasing the difference of peak and valley prices, involving industrial users to participate in interruptible DR schemes, increase in the usage of energy storage devices. Beijing was pioneer city that took part in DSM pilot projects of National Development and Reform Commission (NDRC) and Ministry of Finance (MoF). Being part of three year plan (2013-2015), Beijing reduce load up to 800 MW that composed of 4 to 5% of city's entire load. List of targets of all pilot cities of three year plan are enlisted in Table VI.

TABLE VI. PILOT DR PROGRAMS TILL 2015

City	**2015 Load Reduction Goals (MW)**			**Targeted End Users**
	Temporary reduction	***Permanent reduction***	***Total load reduction***	
Beijing	150	650	800	Commercial buildings, Industry and Municipal facilities
Suzhou	200	800	1000	Industry and Municipal facilities
Foshan	90	360	450	Industry and Municipal facilities
Tangshan	Total load reduction-400			Industry

Load reduction, less reliance on fossil fuels and achievement of 40 % renewable energy generation till 2030 are the main goals of China. Investing in the latest control and communication technologies, China is making progress by implementing pilot projects across industrial provinces. However, lack of competitive market prices, reasonable incentive allotment, customer education programs and Government involvement and supervision impedes the growth of DR in China [39], [41]-[43].

VII. CONCLUSIONS AND FUTURE WORK

DR can play a vital role in balancing the supply and demand without introducing more generation capacity and threatening environment. 100 GW DR potential in Europe, 14% peak demand reduction using DR in USA till 2019 and 40% RE generation by China till 2030 highlights the significance of DR across world-wide. Current developments within developed countries regarding to a wide rollout of advance metering infrastructure technology will enable the consumer for prompt action during peak hours or in case of contingency.

In spite of the effectiveness in mitigating peak demand and addressing RE integration issues there are still some barriers in using DR to its full potential. Educating customers, introduction of aggregators, competitive market price and two way communication between utility and customers need to be done. Although DR in World has not reached its entire potential, however collecting information about potential targets, proposing an action plan and educating the customers can prove successful to attract remaining energy sectors as well as industrial, commercial and residential customers.

The problem with existing DR programs is that inelastic loads like lamps, refrigerators cannot participate unless users are willing to loss of comfort. Therefore, to solve this issue, novel concept Integrated Demand Response (IDR) has been proposed recently. IDR deals with multi energy carries and provides the option of switching energy source. It guarantees all users participation without loss of comfort and can be further explored for future work.

REFERENCES

[1] Lampropoulos, I., et al. History of demand side management and classification of demand response control schemes. in 2013 IEEE Power & Energy Society General Meeting. 2013.

[2] Joseph, H.E., The Past, Present, and Future of U.S. Utility Demand-Side Management Programs. 1996, LBNL: Berkeley.

[3] Auffhammer, M., C. Blumstein, and M. Fowlie, Demand-Side Management and Energy Efficiency Revisited. The Energy Journal, 2008. 29(3): p. 91-104.

[4] Gupta, P., Impact of Demand Side Management Programs on Peak Load Electricity Demand in North America, 1992 to 2008.

[5] Strbac, G., Demand side management: Benefits and challenges. Energy Policy, 2008. 36(12): p. 4419-4426.

[6] Palensky, P. and D. Dietrich, Demand Side Management: Demand Response, Intelligent Energy Systems, and Smart Loads. IEEE Transactions on Industrial Informatics, 2011. 7(3): p. 381-388.

[7] Gelazanskas, L. and K.A.A. Gamage, Demand side management in smart grid: A review and proposals for future direction. Sustainable Cities and Society, 2014. 11: p. 22-30.

[8] Zhao, H. and Z. Tang. The review of demand side management and load forecasting in smart grid. in 2016 12th World Congress on Intelligent Control and Automation (WCICA). 2016.

[9] Song, L., Y. Xiao, and M.v.d. Schaar, Demand Side Management in Smart Grids Using a Repeated Game Framework. IEEE Journal on Selected Areas in Communications, 2014. 32(7): p. 1412-1424.

[10] Mahmood, A., et al., A New Scheme for Demand Side Management in Future Smart Grid Networks. Procedia Computer Science, 2014. 32: p. 477-484.

[11] Solomon, S., et al., Persistence of climate changes due to a range of greenhouse gases. Proceedings of the National Academy of Sciences of the United States of America, 2010. 107(43): p. 18354-18359.

[12] Tahir, M.F., et al., Optimal Load Shedding Using an Ensemble of Artificial Neural Networks. International Journal of Electrical and Computer Engineering Systems, 2016. Volume 7, (2): p. 39-46.

[13] Tahir, M.F., H. Tehzeeb ul, and M.A. Saqib, Optimal scheduling of electrical power in energy-deficient scenarios using artificial neural network and Bootstrap aggregating. International Journal of Electrical Power & Energy Systems, 2016. 83(Supplement C): p. 49-57.

[14] Chen, H., et al., Key Technologies for Integration of Multitype Renewable Energy Sources&#x2014;Research on Multi-Timeframe Robust Scheduling/Dispatch. IEEE Transactions on Smart Grid, 2016. 7(1): p. 471-480.

[15] Verzijlbergh, R.A., et al., Institutional challenges caused by the integration of renewable energy sources in the European electricity sector. Renewable and Sustainable Energy Reviews, 2017. 75: p. 660-667.

[16] Boyd, P.A., G.B. Parker, and D.D. Hatley, Load Reduction, Demand Response and Energy Efficient Technologies and Strategies. 2008: United States.

[17] Bradley, P., M. Leach, and J. Torriti, A review of the costs and benefits of demand response for electricity in the UK. Energy Policy, 2013. 52: p. 312-327.

[18] Paridari, K., et al. Demand response for aggregated residential consumers with energy storage sharing. in 2015 54th IEEE Conference on Decision and Control (CDC). 2015.

[19] Liu, Y., Demand response and energy efficiency in the capacity resource procurement: Case studies of forward capacity markets in ISO New England, PJM and Great Britain. Energy Policy, 2017. 100(C): p. 271-282.

[20] Goldman, C., et al., Coordination of energy efficiency and demand response. 2010, Ernest Orlando Lawrence Berkeley National Laboratory, Berkeley, CA (US).

[21] Qdr, Q., <DOE_Benefits_of_Demand_Response_in_Electricity_Markets_and_R

ecommendations_for_Achieving_Them_Report_to_Congress.pdf>. US Department of Energy, February 2006: p. 1-122.

[22] Qdr, Q., Benefits of demand response in electricity markets and recommendations for achieving them. 2006.

[23] Aalami, H.A., M.P. Moghaddam, and G.R. Yousefi, Demand response modeling considering Interruptible/Curtailable loads and capacity market programs. Applied Energy, 2010. 87(1): p. 243-250.

[24] Shayesteh, E., et al. Congestion Management using Demand Response programs in power market. in 2008 IEEE Power and Energy Society General Meeting - Conversion and Delivery of Electrical Energy in the 21st Century. 2008.

[25] Gill, C.P.S., Y.S. Brar, and K.S. Singh. Incentive based demand response program: An effective way to tackle peaking electricity crisis. in 2012 25th IEEE Canadian Conference on Electrical and Computer Engineering (CCECE). 2012.

[26] York, D. and M. Kushler. Exploring the Relationship Between Demand Response and Energy Efficiency: A Review of Experience and Discussion of Key Issues. 2005.

[27] Ruff, L., Economic principles of demand response in electricity. 2002.

[28] Hurley, D., P. Peterson, and M. Whited, Demand response as a power system resource. 2013.

[29] Lee, M., et al., Assessment of demand response and advanced metering. Federal Energy Regulatory Commission, Tech. Rep, 2013.

[30] Cappers, P., C. Goldman, and D. Kathan, Demand response in US electricity markets: Empirical evidence. Energy, 2010. 35(4): p. 1526-1535.

[31] Hamidi, V., F. Li, and F. Robinson, Demand response in the UK's domestic sector. Electric Power Systems Research, 2009. 79(12): p. 1722-1726.

[32] Torriti, J., M.G. Hassan, and M. Leach, Demand response experience in Europe: Policies, programmes and implementation. Energy, 2010. 35(4): p. 1575-1583.

[33] Coalition, S.E.D., Mapping demand response in Europe today. Tracking Compliance with Article, 2014. 15.

[34] Eid, C., et al., Time-based pricing and electricity demand response: Existing barriers and next steps. Utilities Policy, 2016. 40: p. 15-25.

[35] Trinomics, The Potential of Electricity Demand Response Committee on Industry, Research and Energy (ITRE) 2017: p. 1-66.

[36] Zhang, Q., et al., External costs from electricity generation of China up to 2030 in energy and abatement scenarios. Energy Policy, 2007. 35(8): p. 4295-4304.

[37] Ming, Z., et al., Historical review of demand side management in China: Management content, operation mode, results assessment and relative incentives. Renewable and Sustainable Energy Reviews, 2013. 25(Supplement C): p. 470-482.

[38] Zhong, J., et al., Demand side management in China. IEEE Power and Energy Society General Meeting, 2010: p. 1-4.

[39] Wang, J., et al., Demand response in China. Energy, 2010. 35(4): p. 1592-1597.

[40] Zhao, Z., et al., Demand side management: A green way to power Beijing. Vol. 7. 2015. 041505.

[41] Dong, J., G. Xue, and R. Li, Demand response in China: Regulations, pilot projects and recommendations – A review. Renewable and Sustainable Energy Reviews, 2016. 59: p. 13-27.

[42] Guo, P., V.O.K. Li, and J.C.K. Lam, Smart demand response in China: Challenges and drivers. Energy Policy, 2017. 107: p. 1-10.

[43] Yang, C.-J., Opportunities and barriers to demand response in China. Resources, Conservation and Recycling, 2017. 121: p. 51-55.

Intelligent System for the use of the Scientific Research Information System

Khaoula Benmoussa, Majida Laaziri, Samira Khoulji
Information System Engineering Research Group
National School of Applied Sciences
Abdelmalek Essaadi University
Tetouan, Morocco

Mohamed Larbi Kerkeb
Information System Engineering Research Group
Faculty of Sciences
Abdelmalek Essaadi University
Tetouan, Morocco

***Abstract*—As part of the digital governance of scientific research of Moroccan universities and national research institutions, the Ministry of Higher Education, Scientific Research and Executive Training has shown great interest in setting up the Moroccan Information System for Scientific Research (SIMARECH). Despite a great effort that was made for the implementation of SIMARECH in Moroccan universities, difficulties appear in the use of this information system. This prompted Abdelmalek Essaadi University to consider developing an intelligent system to provide remote training for users to use SIMARECH to facilitate the learning process, reduce administrative costs for displacement to national universities and save time training, etc. This article is in the context of a new rapidly expanding learning paradigm in the field of artificial intelligence for education. It encompasses the design and development of a SIMARECH Intelligent Learning System of Global Use and Features to provide customized learning and adapt to different environments as well as the types of learning scenarios for user training of SIMARECH, this system is named ISSIMA (intelligent system for the use of SIMARECH).**

Keywords—Moroccan Information System for Scientific Research (SIMARECH); intelligent system; E-learning systems; learning process; interactive learning environments; intelligent tutoring systems

I. Introduction

The academic world is full of enormous potentialities and skills, unfortunately still hidden. Universities are facing change and ICTs are at the heart of this ongoing process. These technologies reconfigure the teaching environment by introducing "forms of innovation". ICTs make it possible to enrich pedagogical content, to stimulate interactions between learners and between learners and teachers, to further individualize the training and reach populations previously excluded from university training; hence, the challenges that have been set by Moroccan universities, particularly in the field of scientific research.

The 2013-2016 action plan [1] has devoted an axis to the promotion of scientific, technical and innovation research, which is based on several projects, among them we find the development of a computer system for scientific research in universities. Indeed, a Moroccan information system for scientific research (SIMARECH) is set up by Abdelmalek Essaadi University and generalized to the various universities after having tested it at this university [2]. This system tracks research activities at the national level and the information needed to manage research. In order for the Abdelmalek Essaadi University to ensure the application, integration and efficient and effective use of the latter by the users at the various national universities, the idea of developing an intelligent system appeared.

This intelligent system named ISSIMA will be devoted to train teacher-researchers, those who are responsible of structures, deans, and administrative staff of universities. In order to facilitate the use of SIMARECH, reduce administrative costs (mainly due to displacement and commitment of trainers, booking rooms and materials, etc.) also save training time (meetings, telephone exchanges, etc.) and avoid the problem of the availability of SIMARECH users in a specific time.

The approaches implemented for the design and development of ISSIMA are derived from both knowledge engineering and software engineering methods. Abdelmalek Essaadi University has taken great care in the study and the choice of the methods used, according to the field and the objectives of the training, to guarantee as result a robust and technically and scientifically valid system.

II. Cross-border Digital Services from Abdelmalek Essaadi University

The experience and lessons learned from the Ministry's emergency program, the desire for consolidation of autonomy of Moroccan universities as well as the new instruments made available to respond effectively to the expectations of stakeholders (Students, Companies, Public Authorities) lead to the definition of a new approach to university governance (modernizing its governance and structures). both in its organization and in the field of its competences inevitably going through the modernization of its information system [3].

In the context of autonomy and international, the proliferation of digital services (intranet, e-mail, forums, website publication, online courses, online document collections, e-administration, etc.) has experienced an unprecedented boom in the history of the Moroccan university. Corollary to this massive deployment of ICT, new issues are now posed with insistence: problems relating to functional consistency, more others are technical relating to the interoperability between applications, others finally relating to the cultural and organizational aspect [4].

Defying these difficulties, technological evolution and feasibility helping, Tempus projects and emergency program supporting, Abdelmalek Essaadi university (UAE) set itself the objective [4], [5], to adapt its organization and equip itself with the tools and methods to carry out its missions and meet the challenges it has set itself. Adapt Its Information System Department has entrusted, missions of design, evolution and urbanization of the information system (business applications and digital services); integration, studies and development of digital applications and services, and administration and maintenance of the information system [6].

This is how the UAE moved to digital governance by taking several actions, complementary and of equal importance, namely [3], [6]:

- The implementation of an original information system dedicated to scientific research, "SIRech", renamed today "SIMARECH" following its adoption by all Moroccan universities.
- The implementation of a management information system for university cooperation SIMACoop.
- The implementation of an integrative single access system for all digital services at the university: a Digital Work Environment (ENT).
- The implementation of an integration portal dedicated to the winners of the university and companies called "KHIRIJ".
- Development of the Information and Management System (GIS).
- Involvement in the information system "JAMIATI".

It goes without saying that the UAE continues, beyond its priorities, to devote efforts and resources to other important missions in the continuity of previous projects. This is particularly the international influence of the university; its ambition in this area is to focus on the international research component by strengthening exchanges and scientific collaborations with EU countries and Latin American countries. aware that the trend has already begun with results showing that research in the UAE is oriented internationally with 82% of the indexed publications made in the context of cooperation [6], [7].

III. Moroccan Information System for Scientific Research (SIMARECH)

In the context of the mobilization of all the actors around the emergency program of the Moroccan Ministry of Higher Education, and in the spirit of careful planning of the strategies of the valorization of the research, that the University Abdelmalek Essaadi presents the platform SIMARECH. It is not only a tool for steering research in universities, but also in its institutions and structures, with the ultimate aim of supporting researchers and enhancing their scientific output.

SIMARECH is indeed the result of a collaborative work, its creation was at the Abdelmalek Essaadi University in 2008 [2]. The first version was in fair testing phase at the UAE institutions. Later, other Moroccan universities and national research institutions were gradually included in the implementation of other versions of SIMARECH in order to offer new features that meet the expectations and needs expressed by users.

SIMARECH is an IT platform that facilitates the collection and management of data on research activities and results within an organization [2]. It aims to present in a coherent way the establishments, the research units within the university, including the personnel and their scientific activities (publications, patents, events, equipment, etc.), as well as the follow-up of the financial means and the international activities (partnership agreements, congresses, etc.). Researchers, administrators, and evaluators have all the tools they need to monitor research results, improve visibility, and efficiently allocate available resources (Fig. 1).

Initially, the system was structured to present and include a range of information services (programs, projects, events and products resulting from the activities of the research developed, etc.) for the use of users (teacher-researchers, those who are responsible for structures , deans, and administrative staff) according to their role (Fig. 2). Moreover, at the strategic level, it aims to have a set of indicators and available research statistics. Modular nature of the system and flexibility of its data model allowing different users to access certain information and functionalities throughout the process, and facilitate the processing, organization and transmission of information in accordance with international standards [2], [8].

SIMARECH is developed in stages with incremental improvements to the usability of the tool for users. The UAE team has involved users from beginning to end, in order to i) be familiar with the information system, ii) have a vision of what the SI is aiming for and therefore want to provide data to achieve this, and iii) that their wishes regarding what the IS could do are reflected in its design. This avoids that there are major differences between the expected SI and that which is realized.

Ensuring the proper use of SIMARECH system has always been a real concern of the UAE within Moroccan universities and national research institutions. A huge investment is already being made to provide necessary training for the implementation of SIMARECH by displacement and organizing meetings, personal visits or telephone and e-mail exchanges, etc. to help users of SIMARECH (the teacher-researchers, those who are responsible for structures ,deans and administrative staff) to understand the features and use of the system. But despite all the efforts made by UAE and that SIMARECH integrates an intuitive and usability interface for users to adapt with the system, difficulties and problems arise (especially teachers who are not used to using a such a system, faculty of humanities and human sciences, the elderly, etc.) to use and access the various features of the system.

A solution to be effective should be able to adapt to the user by providing effective learning tools, and at the same time reduce administrative costs and save training time. Indeed, each individual according to his abilities has his own strategies and pace of learning. For this, the university has thought to develop an intelligent system to meet this need. Intelligent tutoring system (ITS) are ideal solutions in these conditions because they are equipped with environments that provide individualized, adaptive and quality learning [9].

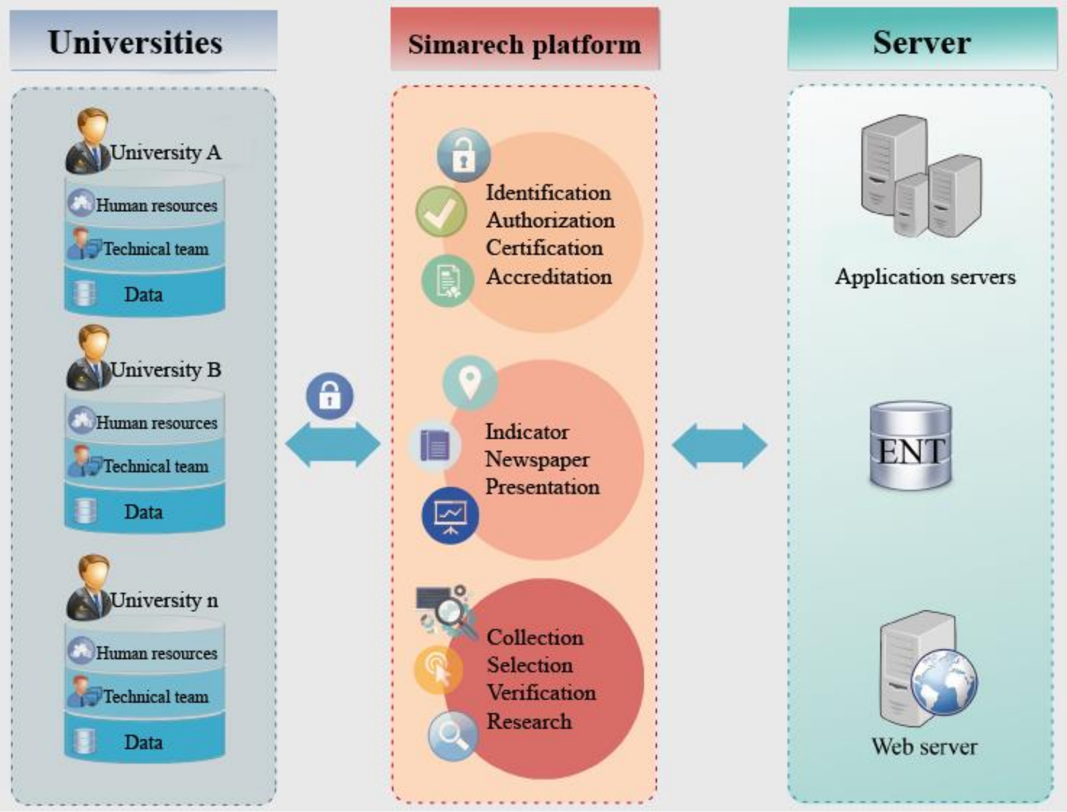
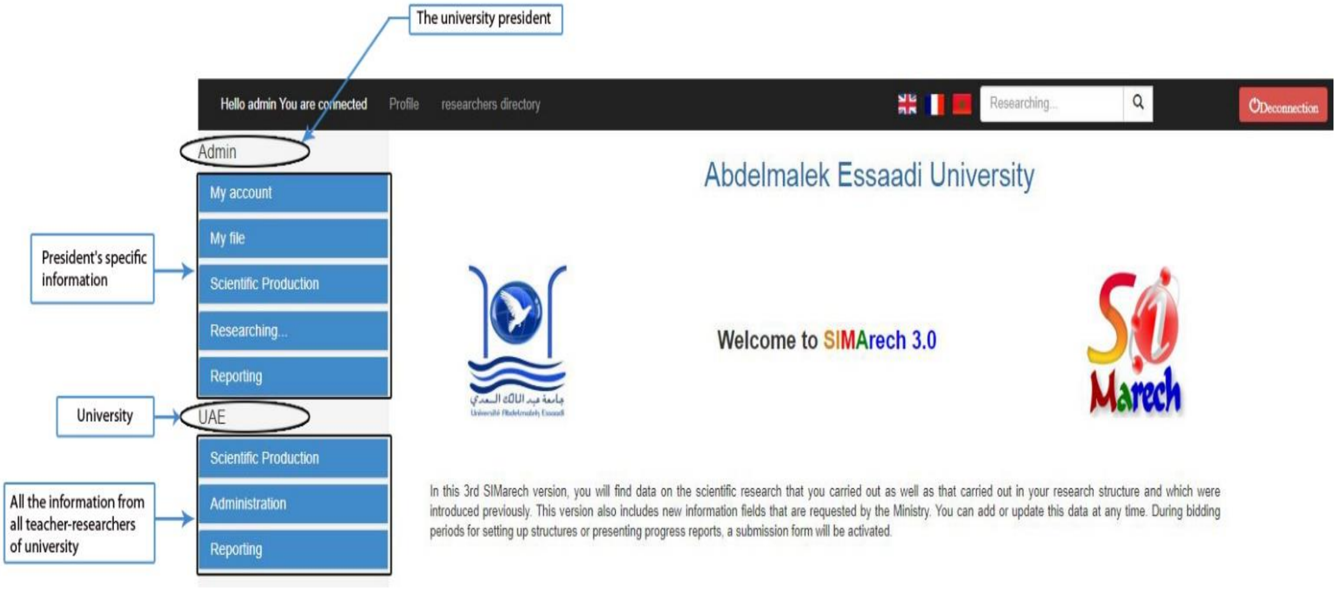


Fig. 1. Integration of SIMARECH in universities [2].



In this 3rd SIMarech version, you will find data on the scientific research that you carried out as well as that carried out in your research structure and which were introduced previously. This version also includes new information fields that are requested by the Ministry. You can add or update this data at any time. During bidding periods for setting up structures or presenting progress reports, a submission form will be activated.

Fig. 2. SIMARECH web application (administrator space) [8].

IV. Development of the Intelligent System for Learning How to Use the Moroccan Information System for Scientific Research (ISSIMA)

A. Overview

E-Learning concerns a method of acquiring knowledge or building knowledge using interactions (actor - actor or actor - resource) relayed by a telematics system (electronic, computer connected by networks) [10]. E-learning broadly refers to the use of technologies for learning. This term includes both e-education, e-learning, e-training, virtual learning, use of ICT; in fact, it is learning using the computer alone or with other multimedia tools [11].

E-learning today nevertheless benefits from international recognition, to find its place among the public. Its effectiveness has been demonstrated several times, sometimes impressively. It represents a reliable alternative to costly face-to-face training and makes it possible to react quickly to the constant evolution of trades [12].

In e-Learning, two levels of complexity can be distinguished for the applications: on the one hand, the systems derived from the computer-assisted instruction (CAI) do not offer any adaptation according to the learner; it is exclusively interested in the data (transmission of the information to be learned). on the other hand those derived from intelligent computer-assisted instruction (ICAI) seeking to understand the learner to provide the best possible response to his or her learning needs [13]–[15]. The ICAI helps to orient learning according to the learner: understand the reasoning of the learner, the reasons for his mistakes, validate and quickly assess his knowledge, his pedagogical preferences to adapt the teaching according to particularities specific to this individual. To achieve this goal, ICAI has quickly established itself as a highly multidisciplinary field. These latter systems are commonly referred to as Intelligent tutoring system (ITS) [13], [16].

The purpose of ITS is to provide the benefits of individual training. It enables learners to practice their skills by performing tasks in highly interactive learning environments [16], [17]. According to Murray [18], ITSs are computer-based teaching systems that have content in the form of a knowledge base (which specifies what should be taught), teaching strategies (which specify how to teach this content) as well as knowledge about the learner's level of content, in order to dynamically adapt their teaching.

The design of such systems involves specialists in AI, the field taught and the teaching, each of whom has a role to play in the components of the system. In general, ITSs consist of four main models [17]: the domain model, the student model, the tutoring model and the interface.

Fig. 3 shows the different components of a ITS [17], [19].

- The domain model, representing the expert's knowledge of the field;
- The student / learner's model, allowing the state of one's knowledge to be established at a given moment;
- The tutoring or pedagogical model in which the knowledge of one or more pedagogical experts is represented, making it possible to make teaching choices according to the behavior and the model of the student / learner;
- The interface model, allowing the exchange of information between the system and the user.

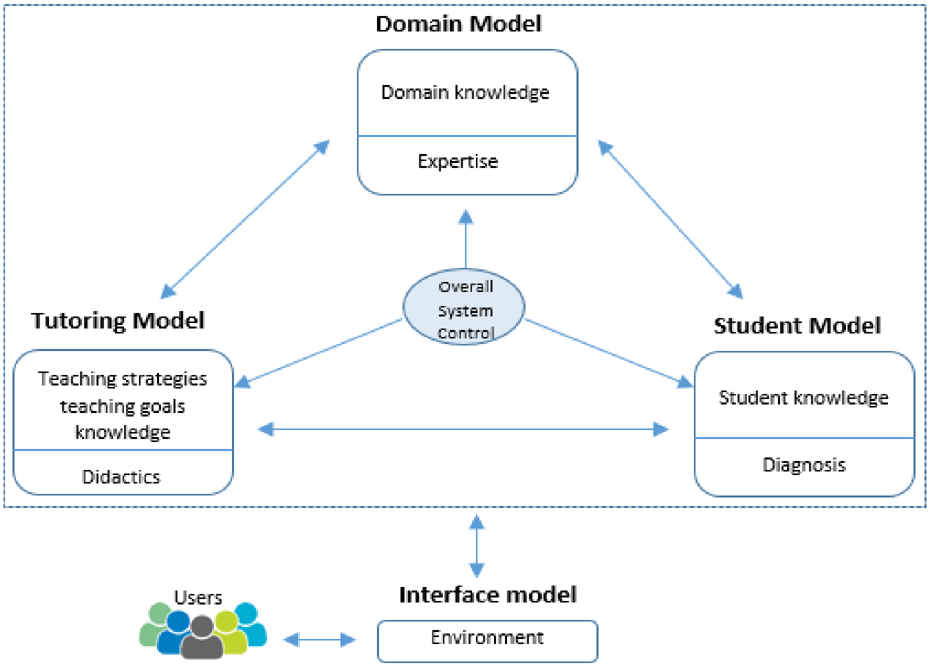


Fig. 3. The components of an intelligent tutoring system [17].

ITS includes both a student model and a domain model, both interacting with the tutoring model, which is available to users in the user interface.

There are three main approaches to the development of ITSs [19], [20] :

- The program sequence is made so that the student can easily adapt himself, in addition to the material is demonstrated to the student just when he / she needs it [16].
- ITS gives detailed feedback to the learner about the imperfect or false solution, helping to learn of his mistakes (If the feedback between ITS and the learner is provided, better training results are obtained [21], [22]).
- Problem-solving methods, little help is provided to the learner so that the right solution is reached (if ITS instantly corrects the mistakes made by the learner or shows the correct way to avoid the error , the knowledge of the student becomes deeper and broader [21], [22]).

B. The Objectives of the ISSIMA Training

In this context, the objectives of the university in this training are to acquire knowledge, skills and Attitudes.

1) Acquisition of knowledge

- Train learners on the principles of SIMARECH.
- Provide learners with the means and knowledge necessary to handle SIMARECH.
- Train learners in operating procedures and management of research units, etc. at SIMARECH.

2) Acquisition of Skills

- Develop learners with the skills they need to perform their tasks.
- Develop skills to improve the usage performance of SIMARECH.
- Provide reminder training when necessary (the case where a skill has not been sufficiently mastered).

3) Acquisition of Attitudes

- Personal effectiveness.
- Allow learners to know each other better at the national level.
- Coach teams / researchers.

Naturally, through these objectives, the training aims at the development of SIMARECH's competence to use, the training also helps to motivate learners to use SIMARECH.

C. Development of the ISSIMA Intelligent System

1) The development of the ISSIMA prototype: The development of the intelligent system prototype for learning the use and functionality of SIMARECH, breaks down as follows [23]:

- Design a system architecture based on classical ITS architecture (Fig. 3).
- To elicit raw knowledge expressed by the experts (to identify and bring out the important concepts of the field and the significant relations between these concepts).
- Build a model of learner knowledge and performance representation including diagnostic tools to identify errors and misconceptions and provide mechanisms to help correct them.
- Establish tutorial rules based on tutoring strategies that have been proven effective and are widely used in learning environments [22].
- Implement and evaluate a first stable ISSIMA prototype and offering the basic features.

The long-term goal is to create an intelligent learning system in which it will be possible to access different environments (dependent on connected actor) and to define learning scenarios for the training of the learner. The system will integrate all SIMARECH usage and feature learning scenarios to meet the needs of SIMARECH actors. Its configuration will have to be simple, intuitive and does not require a great competence in computer programming.

2) The operating cycle of ISSIMA: The operating cycle we present here is the one we designed for the ISSIMA system as a whole, i.e. it applies to all the scenarios defined for the training. It is important to note that ISSIMA is designed to be a learning system that can support and offer SIMARECH usage and functionality training (principles, objectives, role, etc.).

It is important to mention that SIMARECH has four types of actors; it is accessible to all teacher-researchers, those who are responsible for structures, deans, and administrative staff at the level of the presidency of the university. Their knowledge, skills or expertise of using the SIMARECH platform differs, so the use of the system does not present the same difficulties for them.

The SIMARECH platform has four spaces. Each actor of SIMARECH has different access from the other. Our ISSIMA system should offer the possibility of integrating different environments and defining adaptive learning scenarios for each SIMARECH actors. Here we describe the overall course of a training session in ISSIMA. The entire training session takes place in the learning environment on which the learner can act, and through which the tutor interacts with the latter.

When the learner enters the ISSIMA system he will be asked to select his profile: trainer or user / learner (for all actors of SIMARECH); then he will have to choose his username from the list of learners of the selected profile and enter his password. If the learner does not have a username, he will be asked to contact the platform administrator.

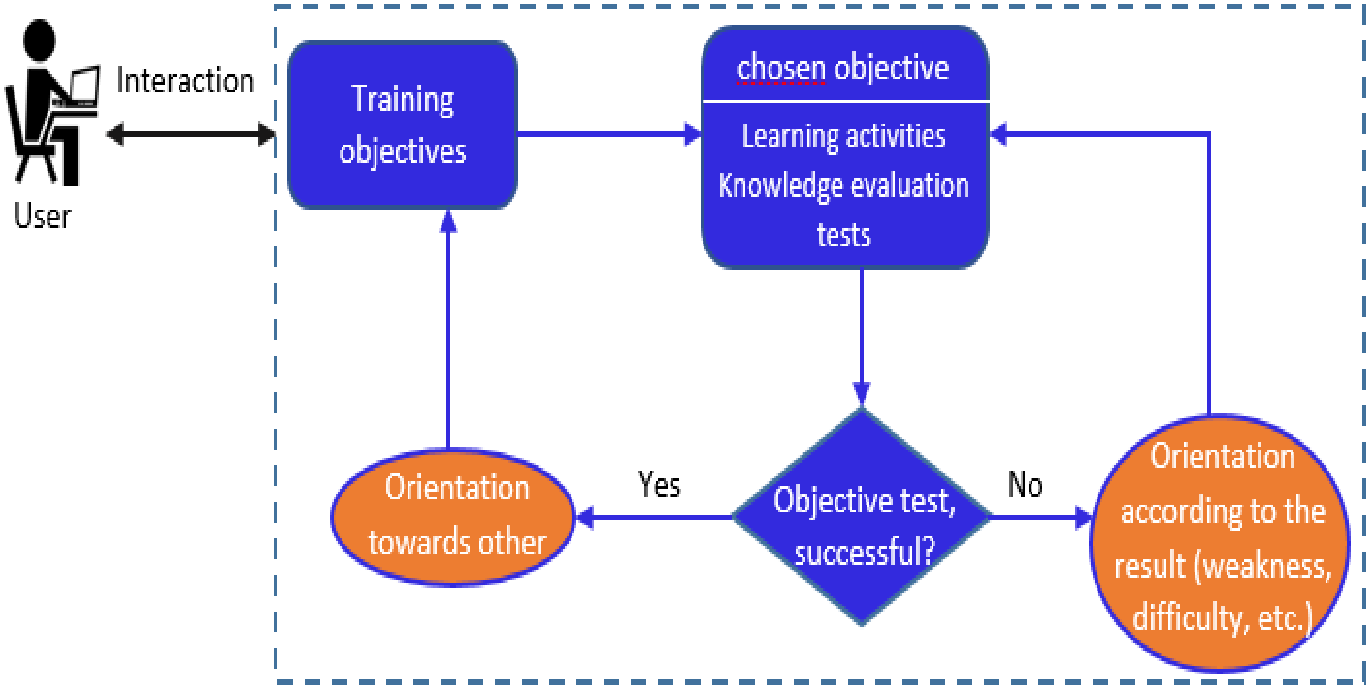

Interaction
User
Training objectives
chosen objective
Learning activities
Knowledge evaluation tests
Objective test, successful?
Yes
No
Orientation towards other
Orientation according to the result (weakness, difficulty, etc.)

Fig. 4. Course of a training session.

At the very first access to his account, the system will offer him a questionnaire. The answers to the quiz questions will be used to create the psychometric model that will be used to initialize the learner model, allowing the system to establish a first cognitive model that will provide the basic information for the learner's personalized follow-up [24]. After accessing his account, and passing the entrance test, the learner will be asked to choose a scenario (an objective) from those available in the learning environment for which the training was defined. Once the objective is chosen, the learner can begin the training under the supervision of the intelligent tutor who observes his movements in his environment and his behavior, note each of his actions and give him positive or negative feedback.

Each interaction of the learner in his / her learning environment or with the tutor, in case he / she will ask for help for example, will be taken into account by the system and the resulting information will be used to update the learner model as it evolves in training. The tutor asks the expert to determine whether the actions taken by the learner in the training are relevant to the current scenario, and gives feedback to the learner if necessary.

At the end of the training, the model of the learner will be presented to him as well as his degree of reaching the chosen objective that will allow him to see at what level he is compared to other learners having the same user profile (actor) either at his university or at national level. Fig. 4 illustrates in a simplified way the progress of a training session.

3) Technical Implementation Options: In the e-Learning projects, the question of the choice of the platform is essential, Indeed, According to the pedagogical scenario that one wants to use at the level of the architecture, some platforms will be more adapted than of other. As the platform on which our system will be based, we have chosen the Moodle platform that is an extremely powerful tool that we have used to ensure the reliability, efficiency and robustness of our solution. It is the most suitable for our needs, enjoys a good reputation as well as numerous Moodle platform are revealed a great success in the teaching field.

Moodle is the contraction of Modular Object-Oriented Dynamic Learning Environment, is a platform of e-learning of constructivist philosophy, free and open source. In other words, it is an accessible system with a web browser, which allows managing online courses (space for filing documents, online activities with students, registration management and access rights, remote tutoring, etc.), its ergonomic interface and its way of being make it easy to handle and make you want to use it [25].

Moodle is a learning platform designed to provide teachers, administrators, and learners with a single, robust, secure, and integrated system for creating personalized learning environments [26]. Indeed, Moodle gives the possibility to organize courses in the form of the sector; it is also a content management system to which is added educational and communicative interaction tools creating an online learning.

The application thus makes it possible to create through its network interactions between pedagogues, learners and educational resources [25], [26].

Based on the features of Moodle [25], we will develop a system for learning how to use SIMARECH by using best practices to achieve the educational objectives, functional expectations of the training. This system will be fully automated, will facilitate learning without the intervention of a human tutor and adaptable to the needs and current knowledge of each actor of SIMARECH, which will give this system the main features of the intelligent computer system.

Our automated system will help us increase the effectiveness of training and achieve significant results in terms of learning. This system is open:

- To all actors of SIMARECH, They can access it by their login and password.
- Only trainers (teachers or others) from the University can create one or more objectives on Moodle.
- An administrator who manages the technical part of the system.

V. Conclusion

We have a strategy to develop an intelligent learning system for the use of Moroccan Information System of Scientific Research (SIMARECH) at national universities, which will be fully automated and adaptable to the needs and current knowledge of each actor of SIMARECH formed in it, which gives this system the main features of the intelligent computer system. This solution will facilitate the use of SIMARECH and reduce administrative costs, save training time, avoid the problem of the availability of SIMARECH users in a specific time.

The system design presented in this article is mainly offered to individual learning situations. The learner's interactions with the learning environment in which he evolves are through a set of actions observable by the system. Throughout the learning process, the learner benefits from the intervention of a machine tutor to guide him in the acquisition of the skills that will enable him / her to complete the training, to be able to understand the functionalities and to use SIMARECH easily.

The development of the SIMARECH intelligent learning system will be a real solution to the difficulties encountered in implementing SIMARECH at the national level.

Acknowledgment

This work was supported in part by the Ministry of Higher Education, Scientific Research and Management Training to support national projects in the management of scientific research.

References

[1] M. de l'enseignement supérieur de la recherche scientifique et de la formation des Cadres, "Projet de Performance," 2015.

[2] K. Benmoussa, M. Laaziri, S. Khoulji, and M. Kerkeb, "SIMARECH 3: A New Application for the Governance of Scientific Research," First Int. Conf. Affect. Comput. Mach. Learn. Intell. Syst. Sch., vol. 5, pp. 776–784, 2017.

[3] R. Maroc, "Gouvernance de l'Université Marocaine : Acquis, contraintes et perspectives de développement," 2013.

[4] "ENSAT." [Online]. Available: http://www.ensat.ac.ma/guide/08servicesnumeriquestransversaux.htm.

[5] A. Azirar, A. Belalia, A. Bellamine, A. Ibenrissoul, A. Driouchi, and D. Zejli, "Comment faire du Maroc un hub régional en matière de recherche scientifique et d'innovation ?," 2015.

[6] H. Ameziane, "Projet de developpement 2015-2018," 2015.

[7] J. Gaillard and A.-I. Afifi, "Appui au système national de la recherche (SNR) au Maroc pour une intégration à l'Espace européen de recherche (EER) : jumela....," 2012.

[8] "SIMARech 3." [Online]. Available: http://simarech.uae.ac.ma/.

[9] P. Phobun and J. Vicheanpanya, "Adaptive intelligent tutoring systems for e-learning systems," Procedia - Soc. Behav. Sci., vol. 2, no. 2, pp. 4064–4069, 2010.

[10] E. Kanninen, "Learning styles and e-learning," Master Sci. Thesis, Tampere Univ. Technol., vol. 12, pp. 1–76, 2008.

[11] L. Changlin and W. Kebao, "Integrated e-learning," in 2009 International Conference on E-Business and Information System Security, EBISS 2009, 2009.

[12] J. L. Moore, C. Dickson-Deane, and K. Galyen, "e-Learning, online learning, and distance learning environments: Are they the same?," Internet High. Educ., vol. 14, no. 2, pp. 129–135, 2011.

[13] R. M. Bottino and M. T. Molfino, "From CAI to ICAI: An educational technical evolution," Educ. Comput., vol. 1, no. 4, pp. 229–233, Jan. 1985.

[14] P. Duchastel, S. Doublait, and J. Imbeau, "Instructible ICAI," J. Inf. Technol., vol. 3, no. 3, pp. 162–168, Sep. 1988.

[15] A. Durey and D. Beaufils, "L'ORDINATEUR DANS L'ENSEIGNEMENT DES SCIENCES PHYSIQUES : QUESTIONS DE DIDACTIQUE," 8èmes Journées Inform. Pédagogie des Sci. Phys. - Montpellier, 1998.

[16] J. Dāboliņš and J. Grundspeņķis, "The Role of Feedback in Intelligent Tutoring System," Appl. Comput. Syst., vol. 14, no. 1, pp. 88–93, 2013.

[17] R. Nkambou, J. Bourdeau, and R. Mizoguchi, Advances in Intelligent Tutoring Systems. 2010.

[18] T. Murray, "Authoring Intelligent Tutoring Systems: An analysis of the state of the art," Int. J. Artif. Intell. Educ., vol. 10, pp. 98–129, 1999.

[19] I. Padayachee, "Intelligent tutoring systems: Architecture and characteristics," Inf. Syst. Technol., pp. 1–8, 2002.

[20] A. Latham, K. Crockett, D. McLean, and B. Edmonds, "A conversational intelligent tutoring system to automatically predict learning styles," in Computers and Education, 2012, vol. 59, no. 1, pp. 95–109.

[21] F. Gutierrez and J. Atkinson, "Adaptive feedback selection for intelligent tutoring systems," Expert Syst. Appl., vol. 38, no. 5, pp. 6146–6152, 2011.

[22] I. Roll, V. Aleven, B. M. McLaren, and K. R. Koedinger, "Improving students' help-seeking skills using metacognitive feedback in an intelligent tutoring system," Learn. Instr., vol. 21, no. 2, pp. 267–280, 2011.

[23] P. Joel Arthur Karol Djeumo, "Implémentation d'un système tutoriel intelligent de type jeu sérieux pour l'apprentissage des pratiques sécuritaires," 2017.

[24] R. Mislevy, M. Wilson, K. Ercikan, and N. Chudowsky, "Psychometric Principles in Student Assessment," Int. Handb. Educ. Eval., pp. 489–531, 2003.

[25] D. R. Tobergte and S. Curtis, Using Moodle, 2nd Edition, vol. 53, no. 9. 2008.

[26] J. Matijašević-Obradović, J. Dragojlović, and S. Babović, "The Importance of Distance Learning and the Use of Moodle Educational Platform in Education," Proc. Int. Sci. Conf. - Sint. 2017, pp. 236–241, 2017.

Generating Relational Database using Ontology Review

Issues, Challenges and Trends

Christina Khnaisser
Département d'informatique
Université de Sherbrooke
Sherbrooke, Canada

Anita Burgun
INSERM UMR 1138 team 22, CRC
Université Paris Descartes
Paris, France

Luc Lavoie
Département d'informatique
Université de Sherbrooke
Sherbrooke, Canada

Jean-Francois Ethier
Département de médecine / informatique
Université de Sherbrooke, Sherbrooke, Canada
INSERM UMR 1138 team 22, CRC
Paris, France

***Abstract*—A huge amount of data is being generated every day from different sources. Access to these data can be very valuable for decision-making. Nevertheless, the extraction of information of interest remains a major challenge given a large number of heterogeneous databases. Building shareable and (re)usable data access mechanisms including automated verification and inference mechanisms for knowledge discovery needs to use a common knowledge model with a secure, coherent, and efficient database. For this purpose, an ontology provides an interesting knowledge model and a relational database provides an interesting storage solution. Many papers propose methods for converting ontology to a relational database. This paper describes issues, challenges, and trends derived from the evaluation of 10 methods using 23 criteria. Following this study, this paper shows that none of the methods are complete as well as the conversion process does not use the full expressivity of ontology to derive a complete relational schema including advanced constraints and modification procedures. Thus, more work must be done to decrease the gap between ontologies, a relation database.**

***Keywords*—Ontology; relational database; database modeling; knowledge model; ontology to relational database**

I. Introduction

Information systems are now at the center of decisions in many areas (healthcare, economic, industrial, manufacturing and so on). A huge amount of data is being generated every day from different sources. Organizations desire to reuse this data for many kinds of analyses to enhance decision-making. The correctness of decision-making depends on the quantity but also on the quality of data collected. Nevertheless, data is stored in different sources that are structured (structural heterogeneity) and encoded (terminological heterogeneity) in different ways. On the one hand, to use data efficiently and correctly from these various heterogeneous sources, experts would ideally be able to express their queries according to a unified knowledge model that represents their domain without the need to know the structure of each database nor to manually extract data from many sources each time. The resulting data should also be available in a unified format reflecting the knowledge model used to define the query. On the other hand, data managers must be able to create, manage, and maintain data with the least possible resources while ensuring its fidelity, integrity, and traceability of its evolution. For this purpose, data integration is the mechanism used for combining data from different sources in a unique unified model offering a single access point. In the database field, the two main and widely used techniques to represent a data model (let's call them conventional techniques) are the entity-relationship model [1] and the object-oriented model [2] which are otherwise mutually convertible [3]. However, these conventional techniques do no longer provide expressivity that is sufficiently complete to semantically interpreted and widely reused data outside a restricted field of application [4]. The current trend in data integration is, therefore, the use of knowledge to enhance the process [4], [5].

In many heterogeneous environments, a knowledge model seems very useful to decipher source structure, and isolate interesting data elements to extract and combine [4]. In the early 80s, the computer science community adopted the ontologies as a knowledge model with reasoning abilities [6] to provide a shared conceptualization of some domain of interest [7]. Since then, ontologies (such as those expressed using the OWL language [8]) have been used in different ways: database modeling, data integration, data mapping, data exchange, data annotation, information retrieval, knowledge discovery, and so on [9]. In particular, ontologies are becoming an important tool for data integration because they handle semantic interoperability by describing a common understanding of data without preoccupation with the underlying layer [7], [4]. In addition, sound integration cannot be done without handling data integrity when a large amount of data sources is highly fragmented and heterogeneous. In light of the recent technology, Relational Database Management System

(RDBMS) with vertical representation [10] and in-memory databases [11] are performing significantly better than "triple stores" sometimes used in conjunction with ontologies [12]. Thus, a RDBMS is needed to allow multiple users to store, modify, and interrogate a large volume of data in a concurrent, reliable, and secure manner [13]–[15].

Yet, since ontologies offer excellent data integration support for disparate systems, a relational database derived directly from an ontology can be hypothesized to be the best way to ensure data integrity and a unique data access point for a large volume of data. First, modeling a relational database using ontologies (rather than entity-relationship) allows the reuse of ontology in many tasks facilitating queries expression as well as ensuring semantic and structural uniformity [16], [17]. Second, with the axiomatic model underlying the ontology, data storage and verification can be automated, increasing data integrity. Reasoning mechanisms can also be leveraged to allow for knowledge discovery. Finally, both models (ontological and relational) must be used together to benefit from the expressiveness of ontologies and the maturity of databases [16].

Several methods have been proposed in the literature to convert an ontology into a relational database [5]. Nevertheless, methods differ in terms of the various ontological constructs covered to generate an enriched relational model. Therefore, 10 methods using 23 specific criteria were described and analyzed.

The rest of the paper is organized as follows: Section II describes the study methodology including conversion issues and challenges as well as criteria definition and the selection process. Section III presents the analysis of the evaluated methods. Section IV elaborates on the advantages of using ontologies and gives an overview of current trends. Section V concludes the paper. Finally, in the appendix, the results of each criterion for the 10 methods are presented in a table.

II. Study Methodology

In this section, first, some ontology and relation database constructs are briefly presented. Then, some noteworthy conversion issues and, challenges are described. Twenty-three evaluation criteria were then derived to evaluate different methods, analyze the challenges, and underline the trends. Using the criteria, the evaluation of 10 methods is presented in the form of a table in the appendix.

A. Issues and Challenges

An ontology and a relational database (relational theory) represent facts using different modeling construct [18]. An ontology is defined using classes, individuals, axioms, properties (object properties and data properties), datatypes, and annotations. For more detail please refer to [8], [19]. A relational database is defined by a set of relational variables (a.k.a *relvar* in relational theory or table in SQL), each relational variable is defined by a set of attributes (pairs of a unique name and a datatype) and constraints. For more detail please refer to the references [20], [21]. Both models also share common foundations: the set theory and the first order logic. Thus, a conversion from one to the other is possible, at least in part, but the main challenge is to maintain the richness of ontological definitions in the resulting relational database by converting uniformly and consistently ontology constructs into a relational construct. Some related challenges are now presented.

1) Preserve property cardinalities: An axiom is an expression that links classes or individuals using properties and cardinalities. A cardinality represents the number of individuals of the related entities that can participate regarding a property. A cardinality is represented as a participation with a range having a minimum value and a maximum value such as [0..1], [1..1], [0..n], [1..n], [0..*], [1..*] and [n..n] where n is a positive integer representing the exact participation value, and a "*" (star) is an undefined participation. In a relational database, an axiom can be represented as candidate keys (primary keys, unique keys), referential keys (foreign keys), and general constraints (functions and triggers). A conversion process that handles cardinality constraints increases data integrity and automatic inconsistency detection.

2) Handle missing information: In an ontology, properties of some class can be optional (as attributes in a relational database). That is, the value can be unknown (in some cases) or inapplicable (in some other cases). This can be implemented by the use of null values, but this implementation is not semantically complete. Moreover, null values can introduce data inconsistency in query results. Thus, it is recommended to avoid them by using other modeling techniques as described in [22], [23].

3) Preserve axioms expressivity: Axioms can be defined using different forms: simple or composite. A simple axiom is defined by atomic entities including a property, one or two classes, and a datatype. Several simple axioms can be combined by set operators (intersection and union) and logical operators (conjunction and disjunction) thus forming composite axioms. Let be the following composite axiom **A OP qt (B ∩ C)** where **A**, **B**, and **C** are classes, **OP** is an object property, and **qt** is a quantifier: some, only, min, max, exactly; it is possible to express it in 3 simple axioms: **A OP qt Z ; Z isa B ∧ Z isa C** where **Z** is a new class subclass of Thing. In a relational database, such operators are defined in the conceptual model and may have different transformations [3] generating different structures and constraints. This may be an issue when using popular medical and biomedical ontologies (see ontologies in OBO foundry repositories [24]) because composite axioms are frequently used. There is a need for a conversion process that uniformly converts such axioms by preserving the complete semantic.

4) Maintain Ontology-SQL type compatibility: A type (datatype) is a set of values. On the one hand, in computing, these sets are necessarily finite but this restriction is not automatically applicable on a data property. On the other hand, for some ontology-type (e.g. owl:rational, xsd:positiveInteger) there is no direct mapping to an SQL-type. Furthermore, SQL-types compatibility and exact handling depend on the target RDBMS. This must also be considered when converting types.

5) Enable structural and tuple reversibility: One of the objectives of using ontologies is the reuse of the model (described entities and axioms) in several tasks. Once data is integrated, a sound mechanism for linking tuples to entities can be very valuable for knowledge discovery and ontology-based data access [7]. In this context, the reversibility between the components of an ontology and those of a relation schema is an essential point to take into account when defining the conversion process. Two reversibility features can be defined: the structure reversibility, the ability to reconstruct or identify an ontological construct from a relational construct; the tuples reversibility, the ability to reconstruct individuals (i.e. as RDF triples) from tuples.

6) Handle knowledge and schema evolution: Knowledge is in constant evolution. This implies changes to ontologies with ensuing repercussions on the related relational schema. The schema must, therefore, cope with it while maintaining earlier knowledge interpretations and preserving coherent data [17]. Moreover, with the opportunity to easily access data, new needs will emerge, and existing needs may change. The impact when integrating knowledge change that implies schema evolution can be very large. Consequently, the conversion process needs to be defined in a way to facilitate structure modification and extensibility.

7) Maintain schema documentation: Relational schema documentation is often ignored despite the added value in many tasks such as data querying, mapping definition, application development and so on [25]. Without a clear definition of the data or the database structure, several tasks cannot be verified or even done [4]. Part of the semantics of the class is carried by the axioms but complementary information can be found in annotations. An annotation in the ontology represents a text in a specific language that defines various aspects of an entity. Using annotations can be beneficial for schema generation and documentation. Examples of interesting types of annotations are: labels (e.g. rdf:label) and comments (e.g. rdf:comment), and definitions (e.g. definition from IAO ontology[1]) which propose a common language description of classes, individuals, and properties.

B. Criteria Definition

Based on the issues and challenges described above, criteria were defined and grouped into four categories: the ontology criteria, the relational schema criteria, the conversion process criteria and tool implementation criteria.

1) Ontology criteria

- Ontology language – the ontology language supported by the method: OWL-DL, OWL-QL, OWL-RL, OWL-EL, RDF(S), DAML, etc.

2) Relational schema criteria

- Structure normalization – the relational schema normal form? : 3NF, BCNF, 5NF or 6NF. This criterion can be deduced from the conversion rules.
- Structure scope – the form of the predicate represented by the relvars of the generated relational: schema generic [G] (*RDF generic style* <predicate, subject, object>) or specific [S] (a specific predicate per class or association).
- Domains[2] – does the method convert the ontology data types and their constraints into domains (e.g. CREATE DOMAIN in PostgreSQL)?
- Primary keys – does the method generate [G] or calculate [C] the primary keys? A generated key is an artificial key defined independently of the set of axioms. A calculated key is a key deduced from the set of axioms.
- Secondary keys – does the method generate the secondary key from the set of axioms?
- Foreign keys – does the method convert the appropriate axioms into foreign keys?
- General constraints – does the method convert cardinalities into general constraints?
- Modification procedures – does the method define the procedures for modifying the data (insert, delete, and update triggers)?
- Supported target RDBMS – such as PostgreSQL, MySQL, Oracle, MSSQL, etc.

3) Conversion process criteria

- Axiom normalization – does the conversion process deal with composite axiom?
- Intermediate structure – the intermediate data structure used for the conversion of OWL into a relational schema: MOF (Meta-Object Facility) FOL (First order logic), RDF, Jena model, etc.
- Type conversion – does the conversion process specify or configure the conversion rules between ontology types and SQL types?
- Restriction conversion – does the conversion process convert the restrictions to general constraints? If yes: is explicit conversion [E] or metadata (implicit conversion [I]).
- Annotation conversion – does the conversion process convert the annotations to document the relational schema?

[1] http://www.obofoundry.org/ontology/iao.html

[2] A domain as defined by the type theory is a finite set of values and a type is a constrained domain that restricts the accepted values.

- Structural reversibility – does the conversion process make it possible to refer to the ontology construct?
- Structural reversibility algorithm – does the method describe the algorithm and propose an implementation of structural reversibility?
- Tuples reversibility – does the conversion process make it possible to import tuples stored in the DB in their full ontological expression?
- Tuples reversibility algorithm – does the method describe the algorithm and propose an implementation of tuples reversibility?

4) Implementation criteria

- Existence of an implementation – has the method been implemented?
- Tool availability – is the tool publicly available?

C. Literature Selection Process for the Evaluated Methods

The search process started in February 2018. The first intent of the search was to find a publicly available tool. To identify methods implemented or updated with current technologies, only papers from the year 2010 and up were retained. The search was conducted with Google Scholar and Engineering Village using the following keywords: “Ontology to relational schema”, “Ontology to relational database”, “Relation schema from ontology”. From 178 papers, 10 were selected and evaluated. The selection was based on the completeness of the paper. The completeness was defined regarding the number of criteria. A paper was selected if at least 15 (over 23) criteria can be evaluated. Moreover, authors of the papers were contacted to validate the evaluation or complete the missing values in the evaluation table and 6 out of 10 responded.

III. Results

This section presents the analysis of the results. The analysis is divided into two categories: general observation and, criteria observation. The Appendix I presents specific criteria results for each method.

A. General Observation

The OWL language is the most popular ontology language. However, the methods differ from one another according to the ontology constructs taken into account in the conversion process. The common ontology constructs used are: classes, objects properties, data properties, subclass axioms with simple restriction (some, exactly, min and max), functional properties characteristic as well as the domain and the range of properties. The methods also differ from one another in several other aspects: the conversion rules, the conversion steps, the quality of the relational schema generated, the availability of the tool, etc.

First, there seems to be a consensus on some conversions rules, especially the class conversion rule is the same for all methods, a class is transformed into a relation. But, properties and axioms are generated differently, i.e. a property can be transformed into an attribute or into a relation depending on the granularity of the conversion rules. In addition, several conversions are suggested to increase the integrity of the data, including the generation of secondary keys, general constraints, modification procedures, and so on. Furthermore, no tool supporting multiple ontologies with an advanced conversion process that was publicly available has been identified. It is thus very difficult to reproduce or share the results, let alone reuse it.

B. Criteria Observation

1) Ontology criteria: OWL defines four profiles: EL, RL, QL, and DL. The DL profile being the superset of the three others [26]. It is important to use DL profile to benefit from a higher degree of expressivity, and to cover more general modeling construct (i.e. universal and existential quantification, cardinality restrictions, functional properties, etc.). While some reasoning operations may be undecidable [26] the gain in expressivity is notable. On the other sides, optimized translation schema may be derived taking the restrictions of EL, RL, or QL into account.

DL is the most supported (7/10) and just one method supports specifically EL, QL and RL profiles, but no indication is given on the available optimizations, if any.

2) *Relational schema criteria:* The generated relational databases differ with respect to their structure and their constraints. On the one hand, all methods generate an ontology-specific relational schema except the method presented in [27]. The latter uses generic representations as a "triple store" where all the data are stored in a single large table (subject, predicate, object). This representation accepts any combination of "fact" but makes it difficult to verify data integrity and to process a large volume of data [13]. In addition, an object property is often converted to a join table, and a data property is always converted to an attribute. These conversions rarely take into account the value of the cardinality. On the other hand, regarding the schema constraint, primary key constraints are generated by most of the methods (7/10), and secondary keys are calculated only by few of them (3/10) using property characteristics (functional and inverse functional). In addition, only two methods generate general constraints, yet this generation is limited to an enumeration restriction (i.e. value enumeration inside a check constraint).

Furthermore, modification procedures are only generated by one method [27] and are limited to the insertion procedure. The generation is based on cardinality restrictions and on all the property characteristics (functional, inverse functional, transitive, symmetric, asymmetric, reflexive and, irreflexive).

Finally, the main RDBMS targeted are PostgreSQL, Oracle, MySQL, and MSSQL. Method [28] supports multiple RDBMS and method [29] uses an ontology database system called OntoDB.

3) Conversion process criteria: The conversion process is unique to each method. The methods differ in terms of the sequence of steps or conversion rules. A noteworthy point is that all methods only deal with simple axioms or do not give an explicit indication of the treatment of composite axioms.

Type conversion: few methods define conversion of ontology types to SQL types (4/10). In the case where a definition is presented, the conversion often depends on an internal configuration or a specific RDBMS. Thus, manual adaptation is needed to reuse the generated relational database in different RDBMS. In addition, some ontology-type does not have a direct mapping to SQL-type, a more advanced mapping mechanism is required.

Restriction conversion: for the majority of the methods (9/10) the participation applicable to an object property in an axiom are converted into referential keys. In addition, a few methods (2/10) generate general constraints (explicit conversion) and some of the methods (4/10) keep information about the participation in metadata tables (implicit conversion). The latter case implies that cardinalities are not verifiable by the RDBMS unless there are constraints or automatisms that take advantage of them. However, these constraints or automatisms are not generated by these methods. In this case, the RDBMS cannot guarantee intrinsically the integrity of the data as described by the ontology. Yet, the integrity of the database can be evaluated externally if the information is present in the metadata, and internally if explicit constraints were generated.

Individual conversion: individuals are converted into tuples by five methods (5/10) so the initial database schema can contain tuples. This is an interesting feature as the database can be ready for querying. In addition, to benefit from it tuples reversibility can be valuable.

Annotation conversion: annotations are rarely used by methods (2/10) to document the database or the automate some conversion rules. Thus, a complementary mechanism is needed to fetch information from the ontology and the relational database to take full advantage of semantics.

Structure reversibility: none of the methods defines the reversibility explicitly. But, according to the overall conversion process, structure reversibility can be easily defined by six of the methods. Only two of the six methods present algorithms for this use.

Tuples reversibility: none of the methods defines the reversibility explicitly. But, according to the overall conversion process, tuples reversibility can be easily defined by five of the methods. Three over five methods present algorithms for this use.

4) Implementation criteria: Most methods have an implementation (7/10) but unfortunately only one is publicly available [30] and its use is limited to one ontology. Furthermore, the tool evaluation is rarely detailed which made it difficult to have a clear view of the evaluation scale.

IV. Discussion

Ontologies are used as knowledge models to define axiomatically the application domain while the relational schema is used as a logical data model to store, modify, and retrieve in a secure way a large amount of data. In the context of data integration, the role of ontologies is twofold. First, an ontology presents a consensual knowledge model [31], thus more suitable for semantic interoperability and for defining a unique access point. Secondly, an ontology offers a formal definition of the database enabling automatic structure and data consistency verification (that can be done automatically by most of the reasoners). Both the ontological and the relational model must be used together to take advantage of the abstraction and expressiveness of ontologies as well as the operational functionalities of databases built using relational schema [16].

As illustrated here, more work is still needed to improve the overall database quality especially to consolidate the schema integrity. First, the relational schema must be normalized and general constraints must be defined based on the cardinality in the axiom. This implies better management of missing information, and participation of individuals according to the property. In addition, the normalization is even more important in the context of physical data warehousing (especially for temporal database and big data) or virtual data warehousing (mediation) where the data extracted from multiple sources are heterogeneous, highly fragmented and context dependent [17], [32]. A “high” normal form (like 5NF and 6NF) reduce uncontrolled redundancy and facilitate schema extension as every part of the schema represents one predicate [33]. Even more, the resulting structure (after normalization) is closer to the set-theory foundation of ontologies and databases, thus facilitating query formulation [34] as the query formulated using ontology entities can have a more direct mapping. Second, preserving axiom expressivity is important to guarantee lossless conversions. Some ontologies may be built with simple axioms definitions but in the biomedical field, this hypothesis is not guaranteed, and as mentioned earlier, multiple ontologies in the OBO foundry use this approach. Describing conversion rules according to axioms definition is, therefore, an important aspect to cover ontologies produced in a large number of domains. Third, the structure of relations in a database differs from the structure of ontological entities. "Ontological" views can be generated to provide a view of the ontology structure. Even more, through these views, modification procedures can be generated to provide better access and standardize data manipulation. Therefore, a method that allows the generation of views and modification procedures can be very beneficial for application development in the sense that data access and modification can be handled at the database level. Fourth, another interesting advantage of using ontologies is the capability to document entities using annotation. However, the methods do not use annotations to document the relational schema (actually documentation is rarely available). For example, ontology label annotation can be used to define different views with different languages increasing the schema accessibility. Finally, regarding ontology-SQL type compatibility, when the RDBMS allows it, domains must be created for ontology types to make it easier to change types during the life cycle of the database.

Maybe more importantly, knowledge and schema evolution are not mentioned in any paper. This subject may be out of the scope of the selected paper, but it is an important one to ensure a long-term solution. Sustainability is a challenge faced by every system, and explicit approaches to address this challenge are required.

Regarding structure reversibility and tuples reversibility, interesting work has been done in the reversibility aspect. That means that the methods are used in a more global context such as ontology data access, data acquisition, and data extraction.

Finally, what slowed down the adoption of ontologies is the lack of publicly accessible tools. Ontologies are by their very nature (shared understanding of a domain) public and shared resources. Thus, to encourage the reuse of ontologies or the use of common database schemas, tools must be publicly accessible and usable with multiple ontologies. Once this is the case, further work to assert that the method is usable at a certain scale will require several evaluations (tests) that are accepted by the community, easily used and reproducible.

V. Conclusion

Ontologies are definitely starting to be “popular” and they are now used in different forms. They are indeed a very promising approach to bring formal semantic to relational databases in order to enhance data interoperability. Nevertheless, in the context of using ontologies to generate a relational database, many issues remain to maintain the full expressivity of ontologies, among them: preserving property cardinalities and axiom expressivity, handling knowledge and schema evolution, handling missing information, and maintaining schema documentation. Complementary, relational approaches have much potential for bringing performance and power to ontology-based data operations. Finally, more work must be done to decreases the gap between ontologies and relation databases but the work presented is a step in this direction.

Acknowledgment

We thank the authors of the selected papers that revised the evaluation table, especially colleagues who took the time to reply and greatly enhance the scientific value of this work. We also thank master student Samuel Dussault for his help to fill the evaluation table. Finally, we would like to acknowledge the important support of the Unité de Soutien SRAP du Québec.

References

[1] P. P.-S. Chen, “The Entity-relationship Model—Toward a Unified View of Data,” ACM Trans Database Syst, vol. 1, no. 1, pp. 9–36, 1976.

[2] OMG, “Unified Modeling Language Specification Version 2.0,” Object Management Group, 2005. [Online]. Available: http://www.omg.org/spec/UML/2.0/. [Accessed: 31-Jan-2018].

[3] R. Elmasri and S. Navathe, Fundamentals of database systems, 6th ed. Boston: Addison-Wesley, 2011.

[4] D. Calvanese, G. D. Giacomo, D. Lembo, M. Lenzerini, R. Rosati, and M. Ruzzi, “Using OWL in data integration,” pp. 397–424, 2010.

[5] G. D. Giacomo, D. Lembo, M. Lenzerini, A. Poggi, and R. Rosati, “Using Ontologies for Semantic Data Integration,” in A Comprehensive Guide Through the Italian Database Research Over the Last 25 Years, Springer, Cham, 2018, pp. 187–202.

[6] T. Gruber, “Ontology,” Encyclopedia of Database Systems. Springer US, Boston, MA, pp. 1963–1965, 2009.

[7] A. Poggi, D. Lembo, D. Calvanese, G. De Giacomo, M. Lenzerini, and R. Rosati, “Linking Data to Ontologies,” J. Data Semant. X, vol. LNCS, no. 4900, pp. 133–173, 2008.

[8] B. Motik, P. F. Patel-Schneider, and B. Parsia, “OWL 2 Web Ontology Language Structural Specification and Functional-Style Syntax (Second Edition),” 2012. [Online]. Available: https://www.w3.org/TR/2012/REC-owl2-syntax-20121211/. [Accessed: 03-Apr-2016].

[9] A. Gali, C. X. Chen, K. T. Claypool, and R. Uceda-Sosa, “From Ontology to Relational Databases,” in Conceptual Modeling for Advanced Application Domains, S. Wang, K. Tanaka, S. Zhou, T.-W. Ling, J. Guan, D. Yang, F. Grandi, E. E. Mangina, I.-Y. Song, and H. C. Mayr, Eds. Springer Berlin Heidelberg, 2004, pp. 278–289.

[10] D. Krneta, V. Jovanovic, and Z. Marjanovic, “A direct approach to physical Data Vault design,” Comput. Sci. Inf. Syst., vol. 11, no. 2, pp. 569–599, 2014.

[11] J. Lee et al., “SAP HANA distributed in-memory database system: Transaction, session, and metadata management,” 2013, pp. 1165–1173.

[12] D. J. Abadi, A. Marcus, S. R. Madden, and K. Hollenbach, “SW-Store: a vertically partitioned DBMS for Semantic Web data management,” VLDB J., vol. 18, no. 2, pp. 385–406, 2009.

[13] A. A. Tzacheva, T. S. Toland, P. H. Poole, and D. J. Barnes, “Ontology Database System and Triggers,” in Advances in Intelligent Data Analysis XII, 2013, pp. 416–426.

[14] I. Astrova, N. Korda, and A. Kalja, “Storing OWL ontologies in SQL relational databases,” Int. J. Electr. Comput. Syst. Eng., vol. 1, no. 4, pp. 242–247, 2007.

[15] L. Al-Jadir, C. Parent, and S. Spaccapietra, “Reasoning with large ontologies stored in relational databases: The OntoMinD approach,” Data Knowl. Eng., vol. 69, no. 11, pp. 1158–1180, 2010.

[16] D.-E. Spanos, P. Stavrou, and N. Mitrou, “Bringing Relational Databases into the Semantic Web: A Survey,” Semant Web, vol. 3, no. 2, pp. 169–209, 2012.

[17] C. Khnaisser, L. Lavoie, H. Diab, and J.-F. Ethier, “Data Warehouse Design Methods Review: Trends, Challenges and Future Directions for the Healthcare Domain,” in New Trends in Databases and Information Systems, T. Morzy, P. Valduriez, and L. Bellatreche, Eds. Springer International Publishing, 2015, pp. 76–87.

[18] C. Pinkel et al., “RODI: A Benchmark for Automatic Mapping Generation in Relational-to-Ontology Data Integration,” in The Semantic Web. Latest Advances and New Domains, F. Gandon, M. Sabou, H. Sack, C. d’Amato, P. Cudré-Mauroux, and A. Zimmermann, Eds. Springer International Publishing, 2015, pp. 21–37.

[19] F. Baader, Ed., The description logic handbook: theory, implementation, and applications, 2. ed., paperback ed. Cambridge: Cambridge Univ. Press, 2010.

[20] E. F. Codd, “A Relational Model of Data for Large Shared Data Banks,” Commun. ACM, vol. 13, no. 6, pp. 377–387, 1970.

[21] H. Darwen and C. J. Date, “The Third Manifesto,” SIGMOD Rec, vol. 24, no. 1, pp. 39–49, 1995.

[22] H. Darwen, “How to Handle Missing Information without Using NULL,” Warwick University, 2005.

[23] C. J. Date and H. Darwen, Database Explorations: essays on the Third Manifesto and related topics. Trafford Publishing, 2010.

[24] OBO, “The Open Biological and Biomedical Ontology Foundry,” The OBO Foundry, 2018. [Online]. Available: http://www.obofoundry.org/. [Accessed: 06-May-2018].

[25] C. Curino, G. Orsi, E. Panigati, and L. Tanca, “Accessing and Documenting Relational Databases through OWL Ontologies,” in Flexible Query Answering Systems, 2009, pp. 431–442.

[26] B. Motik, B. Cuenca Grau, I. Horrocks, Z. Wu, A. Fokoue, and C. Lutz, “OWL 2 Web Ontology Language Profiles (Second Edition).” [Online]. Available: https://www.w3.org/TR/owl2-profiles/#OWL_2_RL. [Accessed: 10-May-2018].

[27] S. Achpal, V. Bannihatti Kumar, and K. Mahesh, “Modeling Ontology Semantic Constraints in Relational Database Management System,” presented at the IMECS International Multiconference of Engineers and Computer Scientists, Hong Kong, 2016.

[28] T. Podsiadły-Marczykowska, T. Gambin, and R. Zawiślak, “Rule-Based Algorithm Transforming OWL Ontology Into Relational Database,” in Beyond Databases, Architectures, and Structures, 2014, pp. 148–159.

[29] L. Bellatreche, Y. Ait-Ameur, and C. Chakroun, "A design methodology of ontology based database applications," Log. J. IGPL, vol. 19, no. 5, pp. 648–665, 2010.

[30] T. Hornung and W. May, "Experiences from a TBox Reasoning Application: Deriving a Relational Model by OWL Schema Analysis," in Proceedings of the 10th International Workshop on OWL: Experiences and Directions (OWLED 2013) co-located with 10th Extended Semantic Web Conference (ESWC 2013), Montpellier, France, May 26-27, 2013, 2013, vol. 1080.

[31] G. Pierra, "Context Representation in Domain Ontologies and Its Use for Semantic Integration of Data," in Journal on Data Semantics X, Springer, Berlin, Heidelberg, 2008, pp. 174–211.

[32] N. Golov and L. Rönnbäck, "Big Data normalization for massively parallel processing databases," Comput. Stand. Interfaces, vol. 54, Part 2, pp. 86–93, Nov. 2017.

[33] C. J. Date, Database Design & Relational Theory. Sebastopol, Calif.: O'Reilly Media, 2012.

[34] P. LePendu, D. Dou, Z. M. Ariola, and C. Wilson, Ontology-based Relational Databases. 2007.

[35] D. Dou, H. Qin, and P. Lependu, "OntoGrate : Towards Automatic Integration For Relational Databases And The Semantic Web Through An Ontology-Based Framework," Int. J. Semantic Comput., vol. 04, no. 01, pp. 123–151, 2010.

[36] D. d B. Saccol, T. d C. Andrade, and E. K. Piveta, "Mapping OWL ontologies to relational schemas," in 2011 IEEE International Conference on Information Reuse Integration, 2011, pp. 71–76.

[37] E. Vyšniauskas, L. Nemuraitė, and B. Paradauskas, "Preserving Semantics of Owl 2 Ontologies in Relational Databases Using Hybrid Approach," Inf. Technol. Control, vol. 41, no. 2, pp. 103–115, 2012.

[38] E. Jiménez-Ruiz et al., "BootOX: Practical Mapping of RDBs to OWL 2," in The Semantic Web - ISWC 2015, 2015, pp. 113–132.

[39] L. T. T. Ho, C. P. T. Tran, and Q. Hoang, "An Approach of Transforming Ontologies into Relational Databases," in Intelligent Information and Database Systems, 2015, pp. 149–158.

[40] H. Afzal, M. Waqas, and T. Naz, "OWLMap: Fully Automatic Mapping of Ontology into Relational Database Schema," Int. J. Adv. Comput. Sci. Appl. IJACSA, vol. 7, no. 11, 2016.

APPENDIX I

The table below (Table I) presents the evaluation result of the criteria and of the requirement for the selected paper respectively. The value "?" means that the information was not found in the article. The evaluated papers are:

A1. Dou et al. [35]
A2. Bellatreche et al. [29]
A3. Saccol et al. [36]
A4. Vyšniauskas et al. [37]
A5. Hornung and May [30]
A6. Podsiadły-Marczykowska et al [28]
A7. Jiménez-Ruiz et al. [38]
A8. Ho et al. [39]
A9. Afzal et al. [40]
A10. Achpal et al. [27]

TABLE I. CRITERIA EVALUATION TABLE

	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10
Ontology										
Ontology language	OWL ?	OWL ?	OWL ?	OWL ?	OWL DL	OWL DL	OWL QL,RL,EL	OWL ?	OWL DL	OWL DL
Schema										
Structure	? S	BCNF S	3NF S	BCNF S	BCNF S	3NF S	BCNF S	? S	? S	? G
Domains	No	No	No	No	No	No	No	No	No	No
Primary Keys	G	C	G	G	C	G	C	G	G	G
Secondary Key	No	?	No	Yes	?	?	Yes	?	?	Yes
Foreign keys	?	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Participation constraints	?	?	No	No	No	No	Yes	No	No	Yes
General constraints	?	?	Yes	Yes	?	Enum	Yes	Enum	No	Yes
Modification procedure	No	No	No	No	No	No	No	No	No	Yes
Target DBMS	?	OntoDB	PostgreSQL	?	?	Multi	?	MySQL	MSSQL	?
Process										
Axiom normalization	?	?	No	?	?	No	No	?	?	?
Intermediate Structure	FOL	MOF	No	OWL W3C	RDF	No	?	Jena	Jena	?
Type conversion	?	?	No	?	?	?	Yes	Yes	Yes	?
Restriction conversion	?	I	No	I	No	No	E*	I	I	E
Individual conversion	Yes	No	No	No	Yes	No	Yes	Yes	No	Yes
Annotation conversion	No	No	No	Yes	Yes	No	No	No	No	No
Structural reversibility	Yes Yes	Yes ?	No No	Yes ?	No No	? ?	No No	Yes No	Yes ?	? ?
Tuples reversibility	Yes Yes	Yes ?	No No	Yes ?	Yes Yes	? ?	Yes Yes	? ?	? ?	? ?
Tool										
Existence	Yes	Yes	Yes	?	No	Yes	Yes	Yes	Yes	?
Availability	No	?	?	?	Yes*	?	?	?	?	?

A5. Tool availability: The tool is a web application and works with Mondial ontology: http://www.semwebtech.org/rdf2sql/

A7. Restriction conversion: CHECK constraints are generated according to individual values

State Transition Testing Approach for Ad hoc Networks using Ant Colony Optimization

Ahmed Redha Mahlous, Anis Zarrad, Taghreed Alotaibi
Computer Science Department, Prince Sultan University
PSU, Riyadh, Saudi Arabia

***Abstract*—Nowadays, telecommunication software organizations are challenged to provide high-quality software to customers within their estimated time and budget in order to stay competitive within the market. Because quality is a defining aspect of the product, it is essential for a project manager to stay alert throughout the project lifecycle. Quality has a direct bearing on customer satisfaction, and if a company produces high-quality products, satisfied customers will rank it highly in customer satisfaction surveys. Additionally, dissatisfied customers are more vocal in their criticisms. Therefore, testing is an important step to produce more reliable systems. In this paper we address two important aspects of software testing for ad hoc network protocols. The first one is by integrating a high-level testing approach based on state transition on top of a network simulator in order to fill a perceived gap in existing network simulators. The second one is reducing testing effort by eliminating redundant test cases, in order to effectively improve the result accuracy of existing network simulators. In this paper, we implemented an automated state transition testing approach for wireless network routing protocols, using an improved Ant Colony Optimization (ACO) algorithm. The expected result is to provide maximum coverage in terms of states and transitions.**

Keywords—Component; ant colony; simulation; optimization; state transition; ad hoc routing protocol

I. INTRODUCTION

In the last few decades, competition in the software market has increased, and software developers are working on high quality products with limited time and budget. Quality is essential in every phase of the project lifecycle as it has a direct bearing on customer satisfaction. Customer dissatisfaction is harmful to a company’s reputation, which is why the testing phase is crucial for developing high-quality, reliable systems.

In the testing phase, developers focus on investigation and discovery, finding out whether their software works in line with customer requirements. Using the results from this phase makes it possible to reduce the number of errors within the software program because it's not possible to solve all the failures you might find during the testing phase [1].

Typically, network protocols are modeled using state-machine diagrams [2] that consist of a finite number of protocol states with or without connections between them. Connections between states are called events, and they facilitate the transition from one state to another. Network protocols are systems with large input and output parameters. For this reason it is necessary to find testing solutions that reduce parameter problems such as duplicates in the path, and at the same time increase the overall effectiveness of the testing. Our method of achieving this was by introducing a state transition testing approach, which is placed on the top of the network simulator in order to ensure an effective and optimal number of test cases.

Due to the complexity of networks, simulators play an important role in overcoming some of challenges that arise from implementing and testing network protocols. However, simulators also have their limitations [3], e.g. traffic generation, documentation, and scalability, especially with the current growth in the use of wireless devices. Furthermore, simulators do not provide an accurate portrayal of real life because they use the queuing theory and discrete events. For example if network congestion is high, estimation of the average occupancy becomes challenging due to high variance [3]. To this end, there is a need to integrate a new software testing approach in order to overcome potential weaknesses in the simulation environment.

Testing is an important phase in the software development lifecycle. The developed software needs to be tested thoroughly to ensure that it meets the needs and expectations of its users, and that it is free from errors. Before starting with the actual validation, testing activities must be planned properly in order to perform effective testing. Delivering a high quality product is crucial for maintaining customer satisfaction and reducing the risk of faults and the cost of repairing them. Quality contributes to the long-term revenue and profitability for companies, making it possible for them to charge and maintain higher prices for their products. Our main focus in this research was to integrate a high-level testing approach into a network simulator to guarantee effective testing. Traditionally, network research communities implement their proposed solutions in a network simulator and generate scenarios based on the network scale, node mobility, and traffic generation models. This approach can limit the scenario size and traffic model exceptions in the source code. Our focus is on the functionalities of network protocols at every point of the simulation.

This research adopts a quantitative methodology. First, related works from journals and conference proceedings will be reviewed. Then, data collection and statistical analysis of Ant Colony Optimization (ACO) will be performed for ad hoc networks. Finally, the proposed approach will be implemented and compared with existing approaches. Fig. 1 illustrates the stages of the methodology.

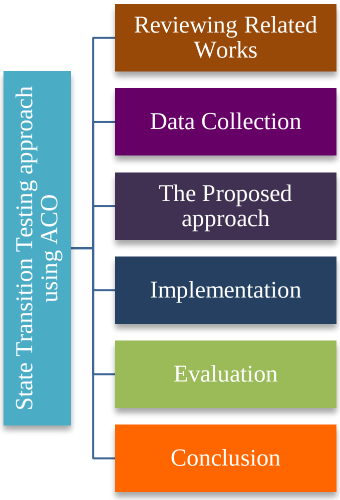


Fig. 1. Methodology.

This paper is organized as follows: Section II provides background and related works; Section III presents the proposed approach, Section IV presents the improvements in ACO and implementation details. Section V describes the results. Section VI provides a comparison between the original ACO and the improved one; and finally Section VII concludes the paper.

II. Background and Related Works

A. *Background*

In this section, we present a brief description of the protocols and algorithm used. Two protocols have been chosen and considered as a good choice for MANETS [6], DSDV and AODV.

1) Ad hoc On-Demand Distance Vector (AODV)

AODV is a routing protocol for ad hoc mobile networks with a large number of mobile nodes. The protocol's algorithm creates routes between nodes only when the source nodes request the routes. This protocol allows the network to be more flexible in order to give nodes the choice to enter and leave the network. The routes will be active unless there are no data packets being sent from the source to the destination. Once the source stops sending packets, the path will be time out and close. AODV additionally supports both unicast and multicast.

The protocol has different types of messages. It initiates the request when needed (“on demand”). Then, route discovery starts with “route request” and “route reply” messages. Finally, routes are maintained just as long as necessary. There are four types of messages used for communication among the nodes. Route Request (RREQ) and Route Reply (RREP) messages are used for route discovery. RREQ packets are broadcast by the source node to connect with the destination node, as the source node has no route entry to the destination node. For the RREP, it unicasts the message to the source node if the node is the destination or has a route to the destination. The other two types are Route Error (RERR) messages and Hello messages, which are used for route maintenance [7].

2) Destination-Sequenced Distance-Vector Routing (DSDV)

DSDV is a proactive routing protocol and a table-driven routing scheme for ad hoc mobile networks requiring each node to periodically broadcast routing updates. This is a table-driven algorithm based on modifications made to the Bellman-Ford routing mechanism. Each node in the network maintains a routing table that has entries for each of the destinations in the network and the number of hops required to reach each of them. Each entry has a sequence number associated with it that helps in identifying stale entries.

Each node periodically sends updates tagged throughout the network with a monotonically increasing even sequence number to advertise its location. New route broadcasts contain the address of the destination, the number of hops to reach the destination, the sequence number of the information received regarding the destination, as well as a new sequence number unique to the broadcast. The route labeled with the most recent sequence number is always used. When the neighbors of the transmitting node receive this update, they recognize that they are one hop away from the source node and include this information in their distance vectors. Every node stores the “next routing hop” for every reachable destination in their routing table. The route used is the one with the highest sequence number, i.e. the most recent one. When a neighbor B of A finds out that A is no longer reachable, it advertises the route to A with an infinite metric and a sequence number one greater than the latest sequence number for the route forcing any nodes with B on the path to A to reset their routing tables [8].

3) State transition testing (STT)

State Transition testing is a testing technique in which changes in input conditions causes state changes in the application under test. It is a process where the tester analyses the behaviour of an application under test for different input conditions in a sequence. The tester provides both positive and negative input test values and records the system’s behaviour. It is helpful where testing different system transitions is needed. The state transition testing model consists of four parts: states, transitions, events, and actions [9].

Moreover, state transition has state diagram, events and test cases as its output, and it is commonly used in black box testing. It uses model of the states for the component to occupy the transitions between those states, the events which cause those transitions, and the actions which may result from those transitions [9].

a) Ant colony optimization (ACO)

Ant Colony Optimization (ACO) is a metaheuristic for solving computational problems. Biologists noticed that blind ants can find the shortest path between food sources and their nests. They also found out that ants spread pheromones on their way, and so other ants can follow the previous pheromone trail by while traversing the paths. The higher the amount of pheromone present, the higher the probability that the ants will follow the trail. This indirect communication between the ants via pheromone trails allows them to find the shortest paths between their nests and food sources [10].

Metaheuristic algorithms are used to escape from local optima, control some basic heuristic: by constructing a heuristic starting from a null solution and adding elements to build a complete one, or a local search heuristic starting from a complete solution and iteratively modifying some of its elements in order to achieve a better one. The metaheuristic part permits the low-level heuristic to obtain solutions better than those it could have achieved alone, even if iterated [10].

B. Related Works

In this section, previous works related to the use of ant colony optimization for testing wireless networks are presented. We also present some of the most used network simulators by the network research community.

Authors in [11] proposed a formal model named “Evolving-Graph-Based Finite State Machine” (EGFSM) to describe the behaviors of protocols in a mobile ad hoc network for conformance testing. To enhance the description capacity for the dynamic behavior, the authors proposed a method to introduce the evolving graph theory to extend the Finite State Machine (FSM) model. They assumed that the topology of the protocol under testing is predictable. The test sequences generated from the proposed model can be adapted in test execution for specific network topologies. Finally, they presented a case study to validate the effectiveness of the proposed model and its generated test sequences.

The authors of [12] proposed a framework for testing wireless sensor networks and presented testing strategies for the same. Their paper addressed some of the characteristics of WSNs, such as node dependencies, the location of the mote, the lack of human interaction during runtime, and how these affect the testing process. The paper mentions two types of testing which must be applied on such applications. The first type is unit testing and it is applied on the motes of the networks. The second one is integration testing and it tests the network after the integration. This type of testing is not applicable for routing protocols since it is based on transitions and searching for paths, and needs intelligent algorithms to automate the testing process.

Authors in [4] proposed an ant colony optimization algorithm to generate an automatic state transition test sequence to offer a strong level of software coverage. The paper focused on providing full coverage and they mentioned that there is a previous work [5] that has applied ACO but with less coverage, while other papers used genetic algorithms to improve the quality of the testing, and neither of them considered full software coverage. They also used the visited status concept to ensure that each vertex is visited at least one. This concept is based on the value of the visited status parameter, for example if there is a connection between two vertices and the second one was not visited by any ants before then the ant will select this vertex. They used a tool called STTACO, a genetic algorithm, and their proposed approach, which based on ACO to apply it on real time case studies “enrolment system”. The proposed solution can be applied on wireless networks since it is a state machine, and the most important thing is that they provided full coverage.

The authors in [1] proposed ACO to generate a set of optimal paths in the control flow graph (CFG) and prioritize the paths. They also proposed an approach to generate a test data sequence within the domain and use it as the input of the generated paths. Their approach guarantees full software coverage with minimum redundancy, and their proposed algorithm prioritizes paths in two ways to decide which paths are to be tested first. The first way takes a CFG as an input and generates optimal and prioritized paths. The second way uses ACO for test data generation and uses it as the input of the generated paths. In the end, they applied it to a binary search program to generate paths as well as inputs. The benefits of their work are: providing full path coverage through the CFG (node, branch, loop, condition), using ACO to generate paths as well as test data, prioritizing the paths to ensure effective testing, and removing redundancy. The authors used a control flow graph as their input, whereas in this research, a state transition table is used as input.

Authors in [13] described a state machine testing approach for cyber-attacks and malicious activities. A map component is implemented between the system-under-test and the learner component. The learner operates abstract inputs and outputs while the mapper transforms the abstract inputs into concrete ones which are accepted by the system. In the opposite direction, it transforms the concrete outputs into abstract ones for the learner. Simulation results show improved outputs from a security viewpoint.

Authors in [24] presented the method of inference and analysis of formal models of botnet command and control (C&C) protocols. Their contribution was to establish a novel state-machine approach for reducing the number of inputs in a realistic high-latency network environment. A Mealy machine modeling approach [2] was used. Optimization of the L* algorithm was implemented to made the inferring procedure more effective.

Although simulators have a major role in evaluating routing protocols, they also have their limitations. Starting with NS2, which is the first developed simulator, the simulator has some issues related to interoperability and coupling between models. There is also a lack of memory management, debugging of split language objects. NS3 was developed to overcome the problems with NS2, but there is still a need for improvements to the animator tool for wireless scenarios, user friendless, ease of use, as well as good tutorials and wider community support.

There are other simulators such as OPNET [14] and QuaNet [15], but they are not useful for students and researchers because they are commercial network simulators and there are no educational licenses for them. Another limitation related to OPNET is that because it is proprietary software, customization options are limited.

Authors in [16] presented a performance comparison between NS2, NS3, OMNeT++, and GloMoSiM using the AODV routing protocol. They stated that in order to evaluate routing protocols using simulators, three variables must be considered: memory usage, computation time, and CPU utilization [16]. Table I illustrates their comparison.

A similar work was presented in [17] where the authors compared various network simulators such as OPNET [14], NS2 [18], NS3 [19] QualNet [16], OMNeT++ [20], J-Sim, and Backplane, and tested their suitability when used for simulation of critical infrastructure. Other research works [21] and [22] have presented comparative study to compare various simulators. However, they do not give a comparative study, and instead only provide a description of each simulator independently. On the other hand, many researches have been conducted on testing routing protocols using high-order testing approaches such conformance testing [23] and black box testing [24]. Authors in [24] proposed a framework for WSN testing to apply distributed unit testing concepts in the development process.

To the best of our knowledge, no previous work has been done to integrate high order testing techniques on top of a simulator environment for correctness and test case selection. Random test cases can easily hide protocol defects because there are no specific procedures to be followed in the testing stage. Thus, random test cases even though will not cover all the cases. In practice, network simulator software is commonly used in testing and protocol assessment. The trustworthiness of results produced from simulation models must be investigated because simulators are less accurate compared to real-life scenarios, as they use queuing theory and discrete events. Simulator output results are often taken as evidence without further verification. In this work, a new layout-level verification tool is added in order to find critical test cases and test implemented protocols.

TABLE I. PERFORMANCE COMPARISON

	NS2	NS3	OMNeT++	GloMoSiM
Memory usage	Highest amount of memory	Lowest amount of memory	Average	Average
CPU usage	Higher	Higher	Lowest	Lowest
Computation time	Highest computation time	Lowest	Low	Low

III. PROPSOED APPROACH

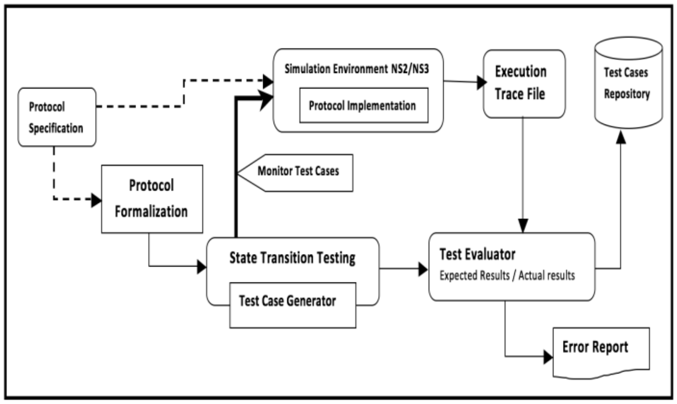


Fig. 2. The proposed approach.

The proposed approach enables testers to be exposed to the design of the protocol and reduces the cost investment.

As shown in Fig. 2, the state transition testing approach was placed on top of the network simulator. The reasons for such decision are to ensure effective test cases along with an optimal number of test cases.

The main components are:

a) Simulation environment: refers the simulation software. The proposed approach can be adapted to any environment, such as NS2, NS3, OMNeT++, etc. In this work, NS2 is used. The trace file entity refers to the actual output generated from NS2. It captures events occurring in the modelled network. The trace data is in ASCII code and is organized into 12 fields.

b) Monitor: is in charge of injecting test cases generated by the state transition testing in the simulation environment. A trace file is generated and passes through the test evaluator module.

c) Protocol Formalization: implementing the state machine formalization model for a specific routing protocol. **Protocol Specification**: represents the protocol's task with a comprehensive description of the intended purpose.

d) State Transition Testing (STT): is the main component for learning about the protocol being tested. This learning is important to create test cases for the simulation environment module. The testing action has two roles: (1) running state transition testing to achieve the highest coverage with minimum redundancy, and (2) transferring generated test cases to the simulation environment module. A complex role is assigned to STT to define test cases based on system knowledge and communication scheme. This is necessary to ensure effective testing with reduced cost. An improvement of the ACO algorithm is implemented in section 4 to enhance the test case quality section and the system performance.

e) Test Evaluator: is used in deciding whether to pass or fail a test case based on protocol formalization and actual/expected results. The expected results are the conditional criteria that show the output that should be generated from the test case. In our case, the expected results when all states and transitions are covered. The test passes if the actual result matches the expected result based on the following specifications:

- All states have been reached at least once.
- All transitions have been executed at least once.
- All feasible paths have been executed.

The actual results are retrieved from the trace file that represents the behaviour observed when a protocol is tested. All tests with a 'pass' are stored in the test case repository for future use in regression testing. Tests with a 'fail' need additional investigation.

A. Apply STT for AODV and DSDV

One of the components in our proposed approach is protocol formalization, which is implementing the state machine formalization model for a specific routing protocol. In this work, state transition testing is applied on wireless network routing protocols AODV and DSDV. Fig. 3 illustrates the AODV protocol using a state diagram. The protocol messages are illustrated as states and transitions. AODV first starts by initiating the request on demand. Then, route discovery is initiated and once it has found one it will request a route and get route reply if a valid route found. After route maintenance, two different types of messages will be sent: Hello and Error.

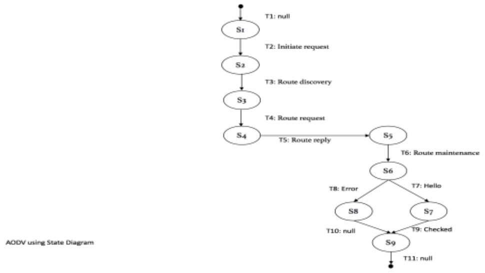


Fig. 3. AODV using STT.

Table II presents the states and events of AODV, which happened in order to move from one state to the other. Each event is described in Table III.

For instance to move from S1 to S2, Event X which corresponds to Initiate Request as shown in Table III should happen first. The same thing applies to the transition from one state to another for all states. Each transition should be done once the corresponding event has occurred.

TABLE II. STATE TRANSITION TABLE FOR AODV

	S1	S2	S3	S4	S5	S6	S7	S8	S9
S1		Event X							
S2			Event Y						
S3				Event J					
S4					Event I				
S5						Event K			
S6							Event Z	Event F	
S7									Event C
S8									Event G
S9									

TABLE III. DESCRIPTION OF AODV EVENTS

Event	Description
Event X	Initiate Request
Event Y	Route Discovery
Event J	Route Request
Event I	Route Reply
Event K	Route Maintenance
Event Z	Hello
Event F	Error
Event C	Checked
Event G	Null

State transition testing was also applied on DSDV as shown in Fig. 4. The protocol messages are illustrated as states and transitions. It first starts by creating a routing table for each node and monitoring the tables for two types of events. The first type occurs when the update is triggered and the second type is a periodic update. For the triggered update, this means that a new change has been detected. This first type of update packet is sent for a major change, and contains all the routing information available at a node. In this case the required action is applying a full dump. The second type is sent for a minor change in the routing table, e.g. new nodes added or link breakage. This type of update packet contains only the information that has changed since the last full dump was sent out by the node. So, the action is applying incremental.

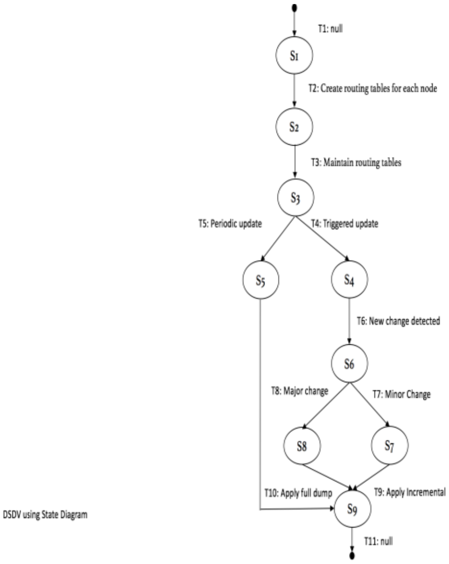


Fig. 4. DSDV using STT.

Table IV presents the states and events of DSDV, which occurred in order to move from one state to the other. Each event is described in Table V.

Table V presents a description of the DSDV events, which describes the normal functioning of the protocol. It comprises the creation of the routing protocol table and its maintenance, detecting updates and applying full dump.

TABLE IV. STATE TRANSITION TABLE FOR DSDV

	S1	S2	S3	S4	S5	S6	S7	S8	S9
S1		Event X							
S2			Event Y						
S3				Event J	Event I				
S4						Event H			
S5									Event G
S6							Event K	Event Z	
S7									Event F
S8									Event C
S9									

TABLE V. DESCRIPTION OF DSDV EVENTS

Event	**Description**
Event X	Create routing tables for each node
Event Y	Maintain routing tables
Event J	Triggered update
Event I	Periodic update
Event H	New change detected
Event K	Minor Change
Event Z	Major change
Event F	Apply Incremental
Event C	Apply full dump
Event G	Null

IV. PROPOSED APPROACH

In this section the description of the proposed algorithm, which is an improved version of the ACO algorithm is presented.

A. The proposed improvement to the ACO Algorithm

In this research work, two improvements to the Ant Colony Optimisation (ACO) algorithm are introduced in order to cover critical states with an optimal number of transitions in the state transition diagram of the protocol under test (PUT), as demonstrated in the following pseudo-code.

```
Algorithm: Ant Colony Optimization

Input: 2D State Transition Table represents the State Transition Diagram.
Output: Number of paths covered, number of visited states, number of transitions covered.

Algorithm for ant p:
Initialize all parameters
    set evaporation factor to 0.1
    set $\alpha$ to 1
    set $\beta$ to 3
    set evaporation factor to 0.1epresentumber of transitions in the s
    Set count: count= Cyclomatic complexity.
    Set key: key1=end _node, it is a variable which store the value of end node.
While (count>0)
 Evaluation at vertex ‘i’
Initialize: i = start.
 Update the track: Update the visited status for the current vertex ‘i’
 i.e. if (Vs[i] = =0)       then Vs[i] =1
Evaluate Feasible Set: This is to determine F (p) for the current vertex ‘i’.
Evaluate the probability from the current vertex ‘i’ to all feasible sets s in the F (p).
The probability is calculated based on the following formula:
    $P = \frac{(\tau_{i,j}^{\alpha})(\eta_{i,j}^{\beta})}{\sum_{1}^{k}((\tau_{i,j}^{\alpha})(\eta_{i,j}^{\beta}))}$
for every k belonging to a feasible set F (p).

Move to next vertex: Using the rule below to move to the next vertex
R1: Select paths (i->j) with maximum probability (Pij).
R2: If two or more paths (from i j and i k) have equal probability like (Pij = Pik ) then select path according to below rule:
    R2.1. Compare each entry in the feasible set with the end_ node i.e.
If (j== end _node) then select ‘k’ as the next node otherwise follow R2.2
    R2.2. If Vs [j] =Vs [k] then select randomly
     Update track: update the covered transition for Tij, if Tij=0 then Tij=1
```

```
Start= next_node
Update the parameter Update Pheromone: Pheromone value updated for transition (i->j) according to the following rule:
If (j= = end_node)
    Go step 2
else
    τ   Go
At the end of each transition and once the ant reaches the destination vertex j the pheromone will be "evaporated" according to:
            $τij = τij - 0.1$
 If (start! = end_node)
go to step 2.3.
else count =count-1.
Go to step 2
End //End of algorithm
```

- The first is the optimization of the ant number in the start state by introducing the Cyclomatic Complexity (CC) concept. The Cyclomatic Complexity (CC) is defined as "CC= Number of closed areas in the diagram +1".

The two advantages of using Cyclomatic Complexity are that it improves the running performance time and optimizes the minimum number of ants. Simulation results show that the running time is largely reduced, compared to the traditional approach where the number of ants is computed based on the ant factory index.

- The second improvement that this research has introduced is the selection factor to critical states. The main idea is to first generate test cases for paths covering a maximum number of critical states. This way, the proposed algorithm reduces the testing effort and guarantees testing of the system's main functionalities.

- The factor P between two states *i*, and *j* is computed as follows:

$$P = \frac{\left(\tau_{i,j}^{\alpha}\right)\left(\eta_{i,j}^{\beta}\right)}{\Sigma_1^k(\left(\tau_{i,j}^{\alpha}\right)\left(\eta_{i,j}^{\beta}\right)} \qquad (1)$$

Where:

- $\tau_{i,j}$ is the pheromone value of the transition between state i and j.

- $\eta_{i,j}$ is the state priority (high = 3, normal =2, and low =1). In this case, the proposed idea will not face the scenario where there is an equal selection factor from a state which has two feasible paths, because of the existence of a pheromone value in the expression and that will vary from one transition to the other, and also because of the consideration of the evaporation factor (which is explained later). The priorities are determined based on the criticality of the transition.

- ∑ The summation of all possible transitions from i to all the neighbouring vertices.

- α= 1 and β=3 (give more influence on the priority states)

This means that if there are two feasible transitions with a high and low priority each, the high priority transition will be selected as shown in the selection factor formula (1). The low priority transition will be selected in the next iteration since it has not been visited. In the case of two transitions, which have been visited, the transition with the higher priority will be selected again. Each value will be discussed in more details:

a) Visited status set: Vs shows information about all the states, which are already traversed by the ant p. For any state 'i':

- Vs (i) =0 shows that vertex 'i' is not visited yet by the ant p.

- Whereas Vs (i) =1 indicates that state 'i' is already visited by the ant p.

b) Feasible set: The procedure evaluates the entire possible transition from the current vertex i to the all the neighboring vertices with the help of a state transition diagram. For example, if there is a vertex i connected to j and k, this means that there are two possible feasible transitions from vertex i§. Otherwise, the algorithm will terminate if there is no feasible path from the current vertex i.

c) Pheromone value**:** The pheromone value will be incremented by 1 in each transition using the following formula: $\tau ij = \tau ij + 1$, and by the end of each transition the pheromone will be evaporated by this rule:

When the ant moves from i to j the pheromone value will be increased at this moment using this formula: $\tau ij = \tau ij + 1$, but when the ant reaches to j the pheromone value will be evaporated using $\tau ij = \tau ij - 0.1$. The pheromone value will also be evaporated for previous events that the ant passes through. For example, if there is a transition from S1 to S2 and from S2 to S3 the evaporation is calculated as:

S1 $\longrightarrow$ S2

$\tau ij = 2$ at S2 the pheromone will be evaporated

$\tau ij = 2 - 0.1 = 1.9$

S2 $\longrightarrow$ S3

$\tau ij = 1.9 + 1 = 2.9$, the pheromone will be evaporated

τ ij $= 2.9 - 0.1 = 1.8$

The pheromone will also be evaporated for S1 and S2

τ ij $= 1.9 - 0.1 = 1.8$

d) The evaporation factor: The evaporation will be calculated to avoid conflict in event selection when priorities are equal.

V. Results

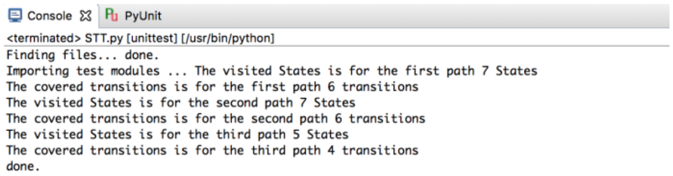
As stated in Section III-A, one of the steps in our proposed approach is protocol formalization, which is implementing the state machine formalization model to a specific routing protocol. In this research, state transition testing on the wireless network routing protocol DSDV is applied.

The state transition diagram for DSDV contains three paths (3), nine states (9), and ten transitions in total (10). Table VI shows the number of states, transitions, and paths of the protocol under test.

TABLE VI. Number of States, Transitions, and Paths in DSDV

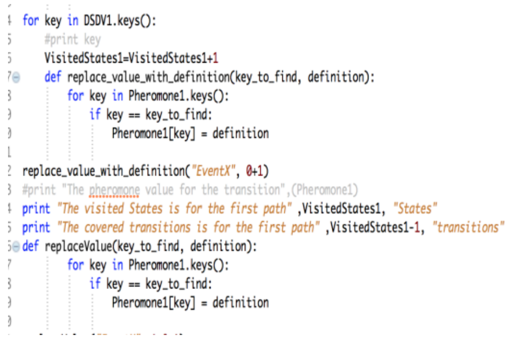
State transition components	Values
States	9 states
Transitions	10 transitions
Paths	3 paths

As stated in the test criteria, Fig. 5 shows the output after implementing the proposed algorithm in Python. The output shows the number of visited states for each path and the number of covered transitions. Fig. 6 illustrates how the automation testing is done and how the pheromone value is updated.



```
<terminated> STT.py [unittest] [/usr/bin/python]
Finding files... done.
Importing test modules ... The visited States is for the first path 7 States
The covered transitions is for the first path 6 transitions
The visited States is for the second path 7 States
The covered transitions is for the second path 6 transitions
The visited States is for the third path 5 States
The covered transitions is for the third path 4 transitions
done.
```

Fig. 5. The output after applying the proposed algorithm.



```
for key in DSDV1.keys():
    #print key
    VisitedStates1=VisitedStates1+1
    def replace_value_with_definition(key_to_find, definition):
        for key in Pheromone1.keys():
            if key == key_to_find:
                Pheromone1[key] = definition

replace_value_with_definition("EventX", 0+1)
#print "The pheromone value for the transition",(Pheromone1)
print "The visited States is for the first path" ,VisitedStates1, "States"
print "The covered transitions is for the first path" ,VisitedStates1-1, "transitions"
def replaceValue(key_to_find, definition):
        for key in Pheromone1.keys():
            if key == key_to_find:
                Pheromone1[key] = definition
```

Fig. 6. Part of automation process.

Table VIII shows the number of visited states, covered transitions, and paths.

TABLE VII. The Coverage after Applying the Proposed

Variables	Coverage
States	19
Transitions	16
Paths	3

Fig. 7 illustrates the coverage on the three levels: states, transitions, and paths as described in Table VII.

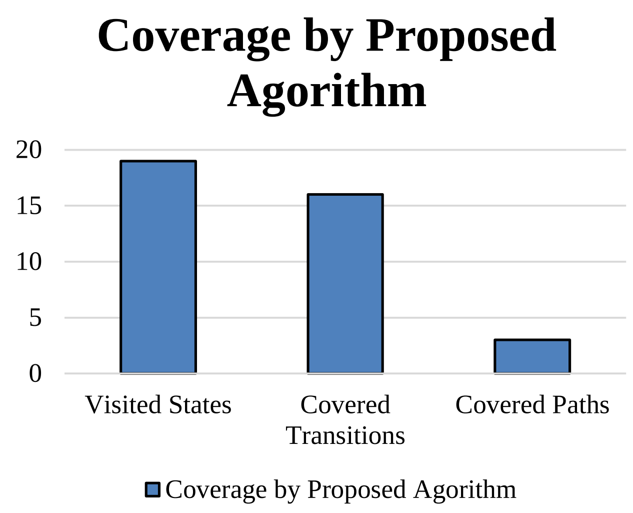


Fig. 7. The coverage of DSDV using the proposed algorithm.

This research also applies the normal ACO algorithm on the DSVV protocol in order to compare results from both approaches. Table IX shows the coverage of states, transitions, and paths.

TABLE VIII. The Coverage after Applying the Normal ACO

Variables	Coverage
States	19
Transitions	16
Paths	3

The number of visited states, covered transitions, and paths are the same as in the proposed approach.

This paper presented the protocol under test, DSDV, and the number of states, transitions, and paths that the protocol contains. It then demonstrated the results after applying both the proposed approach and the traditional testing approach in order to compare the two.

VI. Discussion and Evaluation

A. Discussion

This research paper addresses the problems in existing state-based approaches, such as infeasible paths and the length of test sequences being too long, making the testing process too complex. The proposed algorithm based on ACO finds good paths through the graph. It not only detects feasible paths but also generates minimal non-redundant test cases by providing all definition-use paths, compared to existing state based approaches. This eliminates redundant test cases and saves time consumption.

As mentioned previously, the test completion criteria have been chosen to be as follows:

- All states have been reached at least once
- All transitions have been executed at least once
- All feasible paths have been executed

The number of visited states in each path has been visited. Note that for some states the value of the visited states is more than one, meaning that the ants have visited the state more than one time. For example, {S1, S2, S3, S9} have been visited three times, and similarly {S4, S6} have been visited twice. Other states have only been visited once. Thus, all the states have been ensured to be visited at least once and that all the transitions and feasible paths have been executed. Fig. 8 shows how all the mentioned criteria have been met. The line with "bold dashes" denotes the most critical path, then the "dashes", and finally the "dots" cover the least critical path.

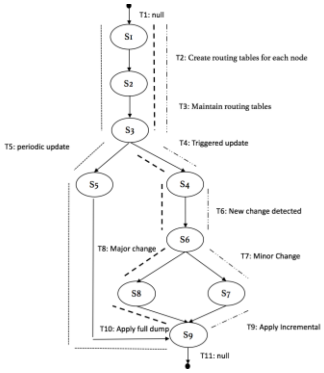


Fig. 8. The coverage of the proposed algorithm.

Furthermore, the normal ACO algorithm has been applied on the protocol in order to compare it with the output from the proposed approach. As demonstrated, the results of the proposed approach are the same as the normal ACO approach in terms of coverage due to the limited number of paths, transitions, and states that the DSDV protocol contains.

B. *Evaluation*

This section presents a comparison of the results between the normal ACO and the proposed approach. Table IX shows the results of both approaches.

TABLE IX. COMPARISON BETWEEN THE NORMAL ACO AND THE PROPOSED APPROACH

	The normal ACO	**The proposed approach**
States	19	19
Transitions	16	16
Paths	19	3

As illustrated in Table IX there is a remarkable difference in terms of the number of paths and coverage parameters. The system's performance was enhanced compared to the normal ACO. Thus, the proposed algorithm ensures that minimal number of paths is used which enhances the performance of the algorithm.

VII. CONCLUSION AND FUTURE WORK

The proposed approach is an improvement to the ACO algorithm. The modification was applied on the selection factor and included the priorities of each event to be calculated with the pheromone value, in order to help the ants in making their decisions. The use of α and β ensures to give the priority states more influence.

As a result, the approach reduced the number of generated test cases and provided full state and transition coverage. The test completion criteria were: all states reached at least once, all transitions executed at least once, and all feasible paths being executed.

Moreover, normal ACO testing has been applied on the protocol in order to perform a comparison between the results. The results showed that the number of visited states, covered transitions, and paths are the same as the results from the proposed algorithm.

As a future work, two ideas are to be proposed: the first is to apply the proposed algorithm on another complex protocol to see the differences in results between the automated testing and the traditional one. The second is to allow users to interact with the software by developing a graphical user interface, which adds animations to the output.

REFERENCES

[1] S. Biswas, "Applying Ant Colony Optimization in Software Testing to Generate Prioritized Optimal Path and Test Data," 2nd Int'l Conf. Electr. Eng. Inf. Commun. Technol. 2015, no. May, pp. 21–23, 2015.

[2] S. Seth and M. A. Venkatesulu, TCP/IP Architecture, Design and Implementation in Linux, PDF. Wiley-IEEE Press, 2009.

[3] J. Heidemann, K. Mills, and S. Kumar, "Expanding confidence in network simulations," IEEE Netw., vol. 15, no. 5, pp. 58–63, 2001.

[4] P. R. Srivastava and K. Baby, "Automated Software Testing Using Metahurestic Technique Based on an Ant Colony Optimization," Electron. Syst. Des. (ISED), 2010 Int. Symp., 2010.

[5] "Software Testing - Quick Guide." .

[6] A. Arya and J. Singh, "Comparative Study of AODV , DSDV and DSR Routing Protocols in Wireless Sensor Network Using NS-2 Simulator," vol. 5, no. 4, pp. 5053–5056, 2014.

[7] A. Khandakar, "Step by step procedural comparison of DSR, AODV and DSDV routing protocol," Int. Proc. Comput. Sci. …, vol. 40, no. Iccet, pp. 36–40, 2012.

[8] B. Jagdale, "Analysis and Comparison of Distance Vector,DSDV and AODV Protocol of MANET," Int. J. Distrib. Parallel Syst., vol. 3, no. 2, pp. 121–131, 2012.

[9] R. Mavinakere, "State transition testing -," pp. 1–18.

[10] T. S. Marco Dorigo, Mauro Birattari, "Ant Colony Optimization . A Computational Intelligence Technique," IEEE Comput. Intell. Mag., vol. 1, no. 4, pp. 28–39, 2006.

[11] T. Shu, M. Gu, and J. Xia, "An Evolving-Graph-Based Finite State Machine Model for Protocol Conformance Testing in MANETs," vol. 10, no. 10, pp. 359–368, 2015.

[12] O. Banias and D. I. Curiac, "Wireless Sensor Network software testing framework," Comput. Cybern. Tech. Informatics (ICCC-CONTI), 2010 Int. Jt. Conf., pp. 517–521, 2010.

[13] J. Bieniasz, P. Sapiecha, M. Smolarczyk, and K. Szczypiorski, "Towards model-based anomaly detection in network communication protocols," 2016.

[14] OPNET, "OPNET Technologies – Network Simulator | Riverbed." [Online]. Available: https://www.riverbed.com/gb/products/steelcentral/opnet.html?redirect=opnet. [Accessed: 18-Dec-2016].

[15] "Qualnet - Packet Trace | SCALABLE Networks." [Online]. Available: http://web.scalable-networks.com/qualnet-network-simulator. [Accessed: 18-Dec-2016].

[16] A. ur Rehman Khan, S. M. Bilal, and M. Othman, "A Performance Comparison of Network Simulators for Wireless Networks," 2012 IEEE Int. Conf. Control Syst. Comput. Eng., pp. 34–38, 2012.

[17] M. Balouchestani, K. Raahemifar, and S. Krishnan, "Increasing the reliability of wireless sensor network with a new testing approach based on compressed sensing theory," 2011 Eighth Int. Conf. Wirel. Opt. Commun. Networks, pp. 1–4, 2011.

[18] K. Fall and K. Varadhan, "The network simulator (ns-2)," URL http//www. isi. edu/nsnam/ns, 2007.

[19] nsnam.org, "Documentation « ns-3," 2011. [Online]. Available: https://www.nsnam.org/documentation/. [Accessed: 21-Dec-2016].

[20] J. Heidemann and U. S. C. Isi, "OMNeT++ Discrete Event Simulator," Audio, no. March, pp. 1–9, 2002.

[21] E. Weingärtner, H. Vom Lehn, and K. Wehrle, "A performance comparison of recent network simulators," in IEEE International Conference on Communications, 2009.

[22] Lessmann, P. Janacik, L. Lachev, and D. Orfanus, "Comparative Study of Wireless Network Simulators," Seventh Int. Conf. Netw. (icn 2008), pp. 517–523, 2008.

[23] R. Alena, D. Evenson, and M. Rundquistt, "Analysis and Testing of Mobile Wireless Networks."

[24] S. Ji, Q. Pei, Y. Zeng, C. Yang, and S. P. Bu, "An automated black-box testing approach for WSN security protocols," in Proceedings - 2011 7th International Conference on Computational Intelligence and Security, CIS 2011, 2011, pp. 693–697.

Verifiable Search Over Updatable Encrypted Data in Cloud Computing

Selasi Kwame Ocansey[1], Changda Wang[1*], Wolali Ametepe[1], Qinbao Xu[1], Yu Zeng[1]
[1]School of Computer Science and Communication Engineering
Jiangsu University, Zhenjiang, P.R. China

***Abstract*—With all the benefits from cloud computing, there are negative influences for the data trust and integrity since clients lose control over the outsourced data in clouds. We propose a verification scheme that supports keywords based search among the encrypted data which is updatable. During the verification process the outsourced cloud data are protected from being inferred by the cloud server. Additionally, if the cloud server returns wrong or incomplete search results the clients will be able to detect such failures. A novel concept in our scheme is the ability of clients to update their outsourced data and to ensure the data's correctness. With our scheme, the data's update efficiency is high and the client's computational cost is low, which makes our scheme very suitable for resource constrained devices.**

Keywords—Cloud computing; verification; outsourced data; update; correctness; search results

I. Introduction

Cloud storage is one of the most significant cloud computing techniques which offer elastic storage services in a "Pay-Per-Use" mode. A lot more individuals and companies are outsourcing their data to storage providers such as iCloud, MediaMax, Dropdox and Strongspace to reduce storage cost and management. However, practical cloud storage usage [1] is still faced with privacy and security risks as a consequence of cloud users losing control over their data to the cloud. Encrypting data before outsourcing is the most effective way in guaranteeing confidentiality. Outsourced encrypted data also encounter challenges especially in terms of its searchability which severely has an effect on the retrievability of data. In addressing this concern, a cryptographic primitive known as Searchable Symmetric Encryption (SSE) is proposed. SSE permits a Cloud Storage provider (CSP) to return keyword based queries on the encrypted data without any data information as well as the keywords being learnt. Most SSE schemes [2]-[9] require data owners to build searchable indexes at the setup phase so as to make subsequent keyword queries executable in an efficient way.

Due to the mistrust of cloud server (*CS*), it is important to ensure that the contents of the outsourced database will not be tampered with as well as the operations performed on the database are correct. Additionally, a proof should be provided by the *CS* to protect the outsourced data's modification by unauthorized users. Furthermore, the malicious *CS* should not be able to update (add or delete) the data. Finally, the clients should be able to verify the returned search results.

A. Our Contribution

The focus of our paper is to ensure that outsourced data with updatable functionality is authentic. Our main contributions are as follows:

1) We propose a new verifiable scheme where cloud clients can verify every outsourced data's block. Thus, the client with the help from additional block information can verify the authenticity of the stored data in the database.

2) Our proposed scheme also supports data updating. Since our scheme is dynamic every data block can be updated by the cloud client without the data being revealed by the updating process. In furtherance, clients can also verify how many times data in block position p in the database has been updated.

B. Our Paper's Organization

The paper's remaining sections are organized as follows: In Section 2, related works are explained and the related preliminaries are given in Section 3. The system model is presented in Section 4 and Section 5 introduces our proposed scheme. Evaluation of our scheme's security and performance analysis is elaborated in Section 6. We conclude our paper in Section 7.

II. Related Works

Data outsourcing has the advantage of shifting cloud clients' data management burden to CSPs which are honest-but-curious (HBC) so as to either save bandwidth resources or its computation. This problem from malicious CSPs paves way for security threats. In guaranteeing the privacy and integrity of data, various cryptographic protocols and primitives have attempted to fight this challenge. Cloud clients also have the ability to guarantee the correctness of search results and also detect fatal search operation. Lots of research has been done on outsourced database verification and determining the search result's correctness. Most of these works are based on techniques which are fully homomorphic encrypted such as [10], [11]. These schemes however lacked practicality. Benabbas et al. [12] proposed a verifiable database scheme with update and retrieval queries based on composite order bilinear groups. A vector commitment was used by Catalano et al. [13] to generate a verifiable database with efficient update. The vector commitment is used as the scheme's response proof. Another scheme which is used to authenticate outsourced database's query results using signature with Merkles hash tree was proposed by Ma et al.

[14]. However, schemes based on Merkles hash tree require much information in order to verify results. Zheng et al. [15] proposed a verifiable scheme by utilizing attribute based encryption but this scheme is not feasible and practical in large datasets. After this, Sun et al. [16], [17] proposed a verifiable conjunctive keywords search schemes foe static and dynamic database but the flaw with these schemes is their need for a secure channel form.

III. Preliminaries

This section introduces some preliminaries that our scheme uses. We also provide our verifiable scheme's algorithms and finally describe the polynomials used.

A. Hash Function

A hash function is a compressive primitive algorithm that accepts arbitrary length inputs (block of data) and outputs a fixed size bit string. Hash function inputs are typically called messages and the outputs as message digests. A cryptographic hash function should be able to stand against the known cryptanalytic attacks. A cryptographic hash function has the following properties:

1) It is computationally infeasible to find two different messages x_1 and x_2 that share the same hash h, such that $h(k, x_1) = h(k, x_2)$, where k is the hash key.

2) Given a hash h value y it should be computationally infeasible to find any message x such that $y = h(x)$

3) Given an input x_1 , it should be computationally infeasible to find a different input x_2 such that $h(x_1) = h(x_2)$.

3.2 Verifiable outsourced database algorithm

A verifiable outsourced database is made up of the algorithms below:

a) Setup: Given security parameters, this algorithm outputs secret key SK for the cloud's client and for the database the public key PK.

b) Initialization: This enables clients to perform pre-computation on cloud data and generates a plaintext encryption algorithm as well as updating operations verification algorithm.

c) Query: Index i is inputted after being generated by an algorithm after which the CS returns the encrypted data, the information about verification, the updating times counter and the proof.

d) Verify: Using SK the encrypted data and the number of updating times are verified by the client.

e) Update: After updating the encrypted data, the client verifies the CS generated verification information and then updates the proof.

B. Polynomial Computation of Our Scheme

Client wants to compute a high degree polynomial's value at some point a malicious powerful *CS*'s help. It is assumed ∂_w is an encoded input and ∂_k is an encoded output, the polynomial $\mathcal{P}$ is an encrypted function. The polynomial

$$p(w) = a_0 + a_1 w + a_2 w^2 + \cdots + a_n w^n$$

Where $a_i \in \mathbb{Z}_q, 0 \leq i \leq n$ is a high degree polynomial that will be outsourced to the *CS*. The function on the value w will be computed by the client. A transformed polynomial $\mathcal{P}(\partial_w)$ is constructed for the secure outsourcing and verification.

1) Client-Side Computation:

Client selects randomly $d \in \mathbb{Z}_q, e \in \mathbb{Z}_q, R \in \mathbb{Z}_q, \ell_0 \in \mathbb{Z}_q, \ell_1 \in \mathbb{Z}_q$ and $f \in \mathbb{Z}_q$ and computes:

$$z_0 = e + a_0$$

The transformed polynomial's coefficient

$$\mathcal{P}(\partial_w) = z_0 + z_1\partial_w + z_2\partial_w^2 + \cdots + z_n\partial_w^n$$

is then generated as

$$z_i = a_i d^i - f^i$$

where 1≤ $i \leq n$. $\mathcal{P}(\partial_w)$ will be outsourced to the *CS*.

Client then computes

$$\tau_0 = g^{\ell_0 + Rz_0}$$

$$\tau_1 = g^{\ell_0 + \ell_1^i + Rz_i}$$

where i =1, 2, ……n

$\mathcal{P}(\partial_w)$ and τ_i will be sent to the *CS*.

If the client wants to compute the value of $p(w)$ at the point w, the client computes $\partial_w = \frac{w}{d}$ and sends ∂_w to the *CS*. The client then computes

$$\mathcal{Q} = \prod_{i=0}^{n} t_1^{\partial_w^i} = g^{\ell_0 \frac{1-(\ell_{1\partial_w})^{n+1}}{1-\ell_{1\partial_w}}}$$

2) Cloud Server-Side Computation:

CS returns

$$\partial_k = \mathcal{P}(\partial_w) \text{ and}$$

$$T = \prod_{i=0}^{n} t_1^{\partial_w^i}$$

The CS then sends (∂_k, T) to the client.

Client computes

$$\mathcal{Q} = \prod_{i=0}^{n} g^{\ell_0 \ell_1^i \partial_w^i} = g^{\ell_0 \frac{1-(\ell_{1\partial_w})^{n+1}}{1-\ell_{1\partial_w}}}$$

and verifies whether $T = \mathcal{Q} g^{R\partial_k}$ holds . If the equation holds, the client can get the final result by computing

$$k = \partial_k - \ k_1 \text{mod } q$$

where

$$k_1 = \sum_{i=1}^{n} f^i \partial_w^i - e$$

IV. System Model

Our system model considers a scenario where in order to guarantee outsourced data's authenticity, data will be encrypted before uploading to the *CS*. The cloud's client can query for their outsourced data anytime and in any location. Due to the *CS* being malicious and untrusted the returned query results needs to be verified for its correctness. Also, outsourced data can be updated anytime by clients and a proof information for the verification generated.

A. Security Model

In our scheme the Cloud Clients are fully trusted since they are the owners of the outsourced data. The *CS* is assumed to be malicious, thus, they may strictly follow predefined protocols but may infer on outsourced data. The *CS* will honestly provide the query results to meet the requirements of the security and should also obtain only encrypted data from the clients.

B. Design Objectives

Our proposed scheme has three (3) main objectives as follows:

a) Privacy preserving: During the updating process, the data's plaintext will not be revealed since it's encrypted before outsourcing to the cloud. Information about stored data should not be obtained by the *CS*.

b) Low computational ***cost***: the cost of computation during the verification phase and the updating process should be low.

c) Verification: Clients should be able to verify the returned result's correctness from a query on stored cloud data. The client can also verify the correctness of the number of the times a stored data is updated.

V. Our Proposed Scheme

We introduce our verifiable scheme in details. It is made up of the following subsections: overview of our proposed scheme and our system's initialization.

A. The Model's Overview

The proposed scheme is based on the verifiable polynomial scheme which will be explained in the next subsection. Our scheme considers encrypted outsourced database (ciphertext) which are in the form $(i, w, \beta, \gamma, proof)$, where i is the index, w is the ciphertext form of the data $\mathcal{W}$, β counts the number of times i that has been updated, γ is the verification information of w which helps generate the $proof$.

When i is queried by the client, the *CS* will return the $(i, w, \beta, \gamma, proof)$ tuple. This protocol's security will guarantee to the client the correctness of w and β. The client will also verify whether w, γ and $proof$ are correct. In updating the index i, the client updates β by setting $\beta' = \beta + 1$. The verification information of $\gamma'$is then verified by the client. If γ' is valid, the updated data $w'$is stored with the new proof $proof'$ on the tuple $(i, w', \beta', \gamma', proof')$. Since data may be updated frequently the existing proof will become non-functional and non-usable by the *CS*.

B. The System's Initialization

The algorithms used in this section have been defined in the subsection 2.3 (Polynomial computation of our scheme) of the paper.

Initialization. Two polynomials $p_1(w)$ and $p_2(w)$ are generated by the clients. $p_1(w)$ is a high degree polynomial

$$p_1(w) = a_0 + a_1 w + a_2 w^2 + \cdots + a_n w^n$$

and $p_2(w)$ is a simple polynomial, so that the client can efficiently compute. In protecting the original data in the database the client selects $d \in \mathbb{Z}_q$ and hides the w which is the original data as $\partial_w = \frac{w}{d}\, d$. The client randomly selects $e \in \mathbb{Z}_q, R \in \mathbb{Z}_q, \ell_0 \in \mathbb{Z}_q, \ell_1 \in \mathbb{Z}_q$ and $f \in \mathbb{Z}_q$, after that $p_1(w)$ into $\mathcal{P}(w)$.

$\mathcal{P}(w) = z_0 + z_1 w + z_2 w^2 + \cdots + z_n w^n$ where

$$z_0 = e + a_0$$

$$z_1 = a_1 d - f^1$$

$$z_2 = a_1 d^2 - f^2$$

$$z_n = a_n d^n - f^n$$

The client then computes $\tau = (\tau_0, \tau_1, \ldots., \tau_n)$ as the public key

$$\tau_0 = g^{\ell_0 + R z_0}$$

$$\tau_1 = g^{\ell_0 \ell_1 + R z_1}$$

$$\tau_2 = g^{\ell_0 \ell_1^2 + R z_2}$$

$$\tau_n = g^{\ell_0 \ell_1^n + R z_n}$$

The secret key in this system is

$$SK = (d, g, e, R, \ell_1, f, p_1(w), p_2(w))$$

where g is a generator of $\mathbb{Z}_q$ and the public key $PK = (\mathcal{P}(w), \tau)$

Ciphertext Generation. For each data in the database, the client masks the original data w as $\partial_w = \frac{w}{d}$, after the masking client uploads to the *CS*. When *CS* receives the masked data ∂_w for the first time, it records the data's position index and sets the counter $\beta = 1$, and then computes

$$\partial_k = \mathcal{P}(w)$$

$$T = \prod_{i=0}^{n} t_j^{\partial_w^j}$$

The *CS* sends ∂_w, ∂_k and T to the client. Client computes $\mathbb{Z}$ with ∂_w.

$$\mathbb{Z} = g^{\ell_0 \frac{1-(\ell_1 \partial_w)^{n+1}}{1-\ell_1 \partial_w}}$$

The equation below is verified by the client to ascertain if it holds $T = \mathbb{Z} g^{R \partial_k}$, Else, client outputs $\perp$. Otherwise, client computes $\gamma = p_2(\partial_k)$, and generates the proof $\mathcal{H} = h(pos||g||\beta)$ where $h(\cdot)$is a non-collision hash function and

pos is the position of the index. The proof $\mathcal{H}$ is sent to the *CS* by the client. $\mathcal{C}$ is the ciphertext that is stored in the outsourced database

$$\mathcal{C} = (i, \partial_w, \partial_k, \beta, T, \mathcal{H})$$

Query. The data in the position i is queried by the client. The *CS* returns to the client the ciphertext $\mathcal{C}_i = (i, \partial_{w_i}, \partial_{k_i}, \beta_i, T_i, \mathcal{H}_i)$.

Verify. Client computes $\mathbb{Z}_i$ with ∂_{w_i},

$$\mathbb{Z}_i = \prod_{j=0}^{n} g^{\ell_0 \ell_1^j \partial_i^j} = g^{\ell_0 \frac{1-(\ell_1 \partial_i)^{n+1}}{1-\ell_1 \partial_i}}$$

And then verifies whether the following equation holds

$$T_i = \mathbb{Z}_i g^{R\partial_{k_i}}$$

If it does not hold, client outputs $\perp$. Otherwise, clients compute $\gamma_i = p_2(\partial_{k_i})$ and then verifies

$$h(i||g^{\gamma_i}||\beta) \overset{?}{=} \mathcal{H}_i$$

If this equation does not hold, client outputs $\perp$.

Update. If the data in position i is to be updated to w_i', the client masks w_i' as $\partial_{w_i'}$ and sends it to the CS.

Cloud server computes

$$\partial_{k_i'} = \mathcal{P}(\partial_{w_i'})$$

$$T_i' = \prod_{j=0}^{n} t_j^{\partial_{w_i'}^j}$$

$\partial_{k_i'}$ and T_i' are returned to the client.

Client then computes $\mathbb{Z}_i'$ and $\partial_{w_i'}$ after which it's verified whether the following equations hold

$$T_i' = \mathbb{Z}_i' g^{R\partial_{k_i}}$$

If it does not hold, client outputs $\perp$. Else, client computes $\gamma_i' = p_2\left(\partial_{k_i'}\right)$ and generates a new proof

$$\mathcal{H}_i' = h(i\left|\left|g^{\gamma_i'}\right|\right|\beta + 1)$$

and then sends the proof to the *CS*. The *CS* then updates the $\mathcal{H}_i'$ (proof) and sets β. The new ciphertext is

$$\mathcal{C}_i' = (i, \partial_{w_i'}, \partial_{k_i'}, \beta_i', T_i', \mathcal{H}_i')$$

Where $\beta_i' = \beta_i + 1$.

VI. Security Evaluation

This section evaluates our scheme's security with theorems and proofs. We also evaluate our scheme's performance based on the cost of query, verification and update.

A. Security Analysis

Theorem 1. The proposed scheme secures the ciphertext from integrity.

Proof of Theorem 1. A new ciphertext has to be generated by the challenger if the ∂_{w_i} (stored data) is changed to $\partial_{w_i'}$,thus, $\mathcal{C}_i' = (i, \partial_{w_i'}, \partial_{k_i'}, \beta_i', T_i', \mathcal{H}_i')$. The challenger's advantage is defined as:

$$Adv_c = Pr\{\mathcal{C}_i' \ is \ valid\}$$

If the $\mathcal{C}_i'$ (new ciphertext) is valid, the probability is

$$Pr_{\mathcal{C}_i'=1} = Pr\{V\left(\partial_{w_i'}, \partial_{k_i'}, \mathcal{H}_i'\right) = 1\}$$

$$= Pr\{V\left(\partial_{w_i'}, \partial_{k_i'}\right) = 1 \wedge (\mathcal{H}_i') = 1\}$$

$$= Pr\{V\left(\partial_{w_i'}, \partial_{k_i'}\right) = 1\} Pr\{V(\mathcal{H}_i') = 1\}$$

where $V(\cdot)$ is the verification function; $V(\cdot) = 1$ implies that data verification has been passed.

The challenger can generate the valid pair $(\partial_{w_i'}, \partial_{k_i'})$. Thus, $Pr\{V\left(\partial_{w_i'}, \partial_{k_i'}\right) = 1\}$

However, γ_i is generated by ∂_{k_i}

$$\gamma_i = p_2(\partial_{k_i})$$

$p_2(w)$ is the secret key and cannot be acquired by the challenger. If ∂_k is changed to $\partial_{k_i'}$, it is difficult to generate γ_i which equals to $\partial_{k_i'}$. Thus, it is hard for the challenger to forge the proof $\mathcal{H}_i'$.

$Pr\{\mathcal{H}_i' = h(i||g^{(\gamma)}||\beta)\}$ is negligible, thus, $Pr\{V(\mathcal{H}_i') = 1\}$ is negligible.

Hence, $Pr_{\mathcal{C}_i'=1} = P\,r\left\{V\left(\partial_{w_i'}, \partial_{k_i'}\right) = 1\right\} P\,r\{V(\mathcal{H}_i') = 1\}$ is negligible. Thus making

Adv_c negligible. The stored data ∂_{w_i} is therefore unforgeable.

Theorem 2. The update's counter cannot be falsified.

Proof of Theorem 2. If β is changed to β_i', the challenger has to generate $\mathcal{H}_i'$ (new proof) to pass the verification. The proof $\mathcal{H}_i' = h(i||g^{(\gamma)}||\beta_i')$ is generated by γ_i. Nonetheless, the challenger can obtain the index i's position and the updates counter β from the ciphertext. Though β_i' is generated by ∂_{w_i} , which can be obtained from the ciphertext, the client's secret key is the polynomial $p_2(w)$. Hence, the challenger cannot obtain β_i' from ∂_{k_i}.

With a given hash it is also impossible to generate a message. Given k, $Pr\{w|h(w) = h(k)\}$ is negligible. Thus, making it difficult to generate a new proof $\mathcal{H}_i' = h(i||g^{(\gamma)}||\beta_i')$.

Theorem 3. Data in a position cannot be substituted with another from a different position.

Proof of Theorem 3. When queries are made for a data in position i, the CS may replace the data in the position i by the data in position j. The ciphertext $\mathcal{C}_i = (i, \partial_{w_i}, \partial_{k_i}, \beta_i, T_i, \mathcal{H}_i)$ will be changed into $\mathcal{C}_i' = (i, \partial_{w_j}, \partial_{k_j}, \beta_j, T_j, \mathcal{H}_j)$.

The $(\partial_{w_i}, \partial_{k_i})$ will pass the verification for the new ciphertext $\mathcal{C}_i'$. The client will then obtain $\beta_i = \wp_2(\partial_{k_j})$. When the client queries the index i, the proof is

$$\mathcal{H}_i' = h\left(i || g^{(\gamma)} || \beta_j\right)$$

$$i = h(i || g)$$

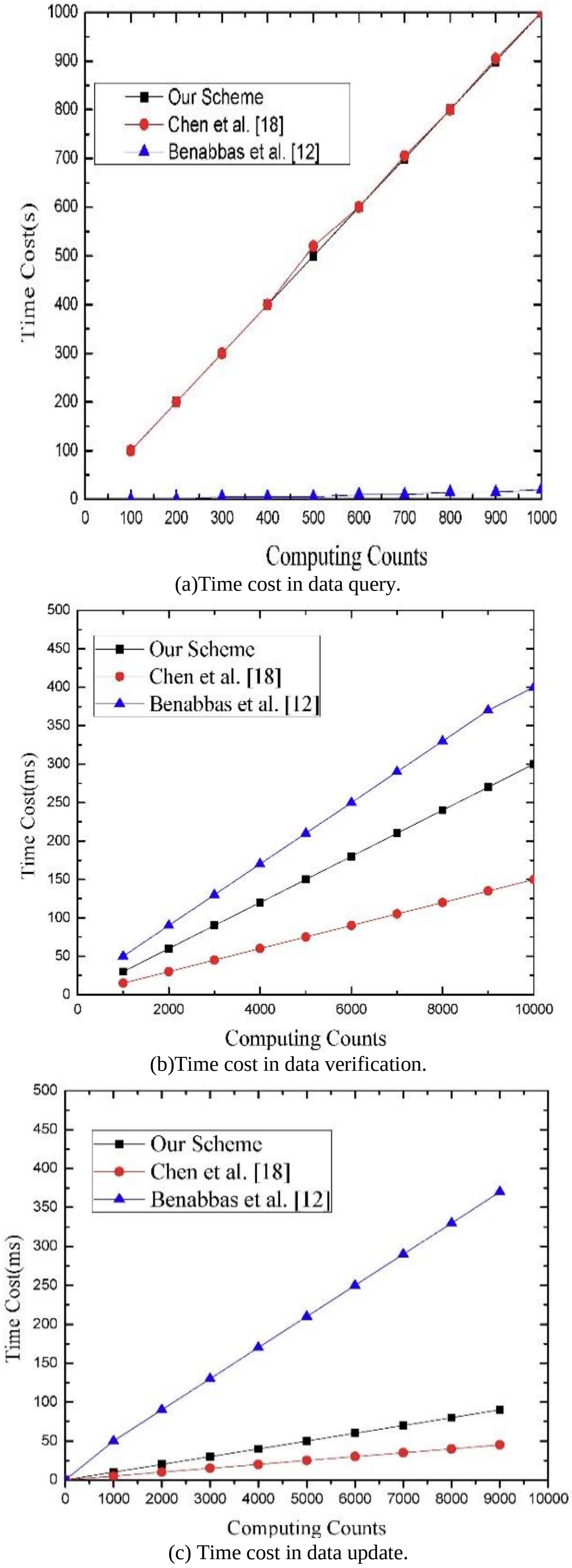
However, when
$$\mathcal{H}_j' = h\left(j || g^{(\gamma)} || \beta_j\right)$$

the CS cannot obtain $g^{(\gamma)}$, where the client's secret key is used to generate γ and g is also a secret. Due to the non-collision hash function property it is impossible for $\mathcal{H}_j'$ to generate $\mathcal{H}_i'$. Hence, the conclusion that data in position i cannot be substituted by data in position j.

B. Performance Evaluation

We provide a detailed experimental evaluation of our proposed scheme. Our experiments are based on the Pairing based cryptography (PBC) [19] library. We performed 20 runs for each test and the average was taken. All experiments were performed on a computer with 16GB RAM, Intel i7-4600 2.7 GHz CPU with Linux Operating system. This specification of the computer will help to measure precisely both the cloud server and the client server's overhead precisely. We provide in our experiment the time cost simulation for our scheme when n=50 and the related schemes [12], [18]. The time cost in the query, verification and update phases are shown in Fig. 1(a), 1(b) and 1(c), respectively.

The *CS* on receiving query request from its client performs a search over the indexes and computes on the queried data. Fig. 1(a) shows our scheme's computational overhead and that of [12], [18] is significant and linearly grows with the computing counts. We however argue that the query's computational overhead is mainly performed by the *CS* rather than the resource constrained client. In a real world scenario, the result indicates that the time is acceptable. Data verification's efficiency comparison among the three schemes is shown in Fig. 1(b). It shows that our scheme can achieve data verification with nearly the computation overhead over the scheme [18]. However, when compared with [12] our scheme shows an increase in computational overhead. Our claim is that our scheme can achieve both verifiability and updating functionalities. Notably, in our scheme the data verification's computational cost can be reduced by the client using the secret key. Simulation results shown in Fig. 1(c) show our scheme being more efficient than that of [12] in the update algorithms by the clients. All these make our scheme most suitable for real world applications.



(a)Time cost in data query.

(b)Time cost in data verification.

(c) Time cost in data update.

Fig. 1. Efficiency comparison for data query, verification and update.

VII. Conclusion and Future Work

In this era of Big Data, end users are faced with many challenges. In dealing with these huge amounts of data, end users need high storage capacities and powerful devices which can perform complex computations. The insurgence of cloud computing has provided solutions to these problems. Cloud computing provides services which includes providing clients with huge storage space and also performing powerful computational operations on these stored data. Pertaining to this paper, we construct a new verifiable scheme which supports updated data on the cloud. During the updating process the original data will be protected from malicious adversaries which includes the CSPs. Our scheme's computational cost is low during the verification, update and query phases.

One disadvantage of the proposed scheme is that when clients continuously insert data into the same database index the number of the level in hierarchical commitment increases. The storage and computational overheads of the *CS* will linearly increase thereby reducing the verifiable scheme's efficiency. Future works should thence try to solve the problem of how to construct a verifiable scheme which is efficient and supports updatable operations regardless of the type of insertion.

Acknowledgments

This work was supported by the National Science Foundation of China under grant 61172269 and the Jiangsu Provincial Science and Technology Project under grant BA2015161.

References

[1] Armbrust M., Fox A., Griffith R., Joseph A. D., Katz R., Konwinski A., Lee G.,. Patterson D, Rabkin A., Stoica I., and Zaharia M. A view of cloud computing. Commun. ACM, 53(4):50–58, 2010.

[2] Cash D., Jaeger J., Jarecki S., Jutla C., Krawczyk H., Rosu M.-C., and Steiner M. Dynamic searchable encryption in very large databases: Data structures and implementation. In Network and Distributed System Security Symposium, NDSS'14, 2014.

[3] Cash D., Jarecki S., Jutla C., Krawczyk H., Rosu M.-C, and Steiner M. Highly-scalable searchable symmetric encryption with support for Boolean queries. In Advances in Cryptology - CRYPTO'13, pages 353–373. 2013.

[4] Curtmola R., Garay J., Kamara S., and Ostrovsky R. Searchable symmetric encryption: Improved definitions and efficient constructions. In Proceedings of the 2006 ACM Conference on Computer and Communications Security, CCS'06, pages 79–88, 2006.

[5] Jarecki, C. Jutla, H. Krawczyk, M. Rosu, and M. Steiner. Outsourced symmetric private information retrieval. In Proceedings of the 2013 ACM conference on Computer and Communications Security, CCS'13, pages 875–888, 2013.

[6] Kamara S. and Papamanthou C. Parallel and dynamic searchable symmetric encryption. In Financial Cryptography and Data Security, pages 258–274. 2013.

[7] Kamara S., Papamanthou C., and Roeder T. Dynamic searchable symmetric encryption. In Proceedings of the 2012 ACM Conference on Computer and Communications Security, CCS'12, pages 965–976, 2012

[8] Naveed M., Prabhakaran M., and Gunter C. A. Dynamic searchable encryption via blind storage. In IEEE Symposium on Security and Privacy, SP'14, 2014.

[9] Stefanov E., Papamanthou C., and Shi E. Practical dynamic searchable encryption with small leakage. In Network and Distributed System Security Symposium, NDSS'14, 2014.

[10] Backes M., Fiore D. and Reischuk R.M. Verifiable delegation of computation on outsourced data. In: 2013 ACM SIGSAC Conference on Computer and Communications Security, CCS'13, Berlin, Germany, November 4–8, 2013, pp. 863–874, 2013.

[11] Boneh D. and Waters B. Conjunctive, subset, and range queries on encrypted data. In: Theory of Cryptography, Proceedings of the 4th Theory of Cryptography Conference, TCC 2007, pp. 535–554. 2007

[12] Benabbas S., Gennaro R. and Vahlis Y. Verifiable delegation of computation over large datasets. In: Advances in Cryptology -CRYPTO 2011—Proceedings of the 31st Annual Cryptology Conference, pp. 111–131, 2011

[13] Catalano D. and Fiore D. Vector commitments and their applications. In: Public-Key Cryptography—PKC 2013—Proceedings of the 16th International Conference on Practice and Theory in Public-Key Cryptography, pp. 55–72, 2013

[14] Ma D., Deng R.H., Pang H. and Zhou J. Authenticating query results in data publishing. In: Information and Communications Security, Proceedings of the 7th Inter-national Conference, ICICS 2005, pp. 376–388, 2005

[15] Zheng Q., Shouhuai X. and Ateniese G. VABKS: verifiable attribute-based keyword search over outsourced encrypted data. In: Proceedings 33th IEEE international conference on computer communications, 522-530, 2014

[16] Sun W.H., Wang B., Cao N., Li M., Lou W.J., Hou Y.T. and H. Li. Verifiable privacy-preserving multi-keyword text search in the cloud supporting similarity-based ranking. IEEE Trans Parallel Distrib Syst 25(11):3025–3035, 2014

[17] Sun W.H., Liu X.F., Lou W.J., Hou Y.T. and Li H. Catch you if you lie to me: efficient verifiable conjunctive keywords keyword search over large dynamic encrypted cloud data. In: Proc. 34th IEEE international conference on computer communications, 2015

[18] Chen X., Li J., Huang X., Ma J.and Lou W. New publicly verifiable databases with efficient updates, IEEE Trans. Dependable Secure Comput. 12 (5), 546–556, 2015

[19] Lynn B. The pairing-based cryptography library, http://crypto.stanford.edu/pbc/, 2006

An Effective Virtual Reality based Remedy for Acrophobia

Maria Abdullah
Department of Computer Systems Engineering
Dawood University of Engineering & Technology
Karachi, Pakistan

Zubair Ahmed Shaikh
Department of Computer Sciences
Mohammad Ali Jinnah University,
Karachi, Pakistan

***Abstract*—Virtual reality (VR) Exposure Therapy with sophisticated technology has been used in the Psychological treatment. The goal is to design a virtual environment using HCI (HMD) device with an interactive and immersive realistic 3D graphic scenes for exposure therapy of acrophobia that allows patient to sense height and gets used to the fearful feelings .The degree of fear is then used to evaluate the effectiveness of the system before and after therapy with the help of comparison. One may feel a little uneasy and perhaps accelerated heart rate, excessive sweating and shortness of breath, etc. are some of the most common physical symptoms of anxiety upon exposure to height. This extreme or irrational fear of height is called "Acrophobia". The HMI based simulation is used which used the body sensation elucidation as physical symptoms of anxiety upon exposure to height to predict the results. The test reveals that anxiety level decreases from 16% at first level exposure and 8% at last level exposure. It is concluded from the results that VR exposure therapy is more effective than real exposure therapy and also the post test for VR exposure therapy were significantly better than post real exposure results. This system provides cost effective solution for rehabilitation environment.**

Keywords*—HMI; virtual reality (VR); HMD; acrophobia; VR exposure therapy; cognitive behavioral therapy; 3D VR environment

I. Introduction

Virtual reality (VR) innovation has been utilized as a part of the mental treatment of acrophobia and has come to command the treatment of various uneasiness issues. It is currently realized that virtual reality introduction treatment VR exposure therapy regimens are profoundly powerful for acrophobia treatment. Acrophobia, or the dread of statures, is described by a silly dread of tallness circumstances bringing about the shirking of such circumstances or the encounter of such circumstances with stamped trouble. This dread for some is persisted with trouble, though for others, the dread is intense to the point that stature circumstances must be maintained a strategic distance from by and large. Acrophobia can be characterized as a nonsensical and exciting apprehension of stature. Numerous specialists have demonstrated victory in diminishing acrophobia utilizing virtual reality. Virtual reality is an incredible innovation, which advances persistently and which must be continually improved. Virtual the truth is a valuable device to treat fears, yet there are additionally different conditions which could be tackled with the assistance of this innovation. We have numerous opportunities, we simply need to continue endeavoring to discover arrangements and accomplish a definitive objective [1].

In VR exposure therapy patients are presented to virtual uneasiness inciting conditions rather than genuine on edge circumstances. VR exposure therapy depends on the presumption that individuals feel 'display' in the virtual condition. Nearness is defined as a mental state or subjective discernment in which, despite the fact that part or the majority of a person's present affair is produced by or potentially filtered through human-influenced innovation. Those taking after focuses highlighted those hugeness of this examination venture and in addition demonstrate the sources from claiming inspiration to those fruition from claiming this study: 1) A lot of people patients would not dealt with since they can't manage the interview fees of a psychotherapist, something like that it might be an expense compelling result. 2) Those tolerant can't make general sessions from claiming treatment due to going by issues or the long run issues, yet toward those recommended framework they could profit at home additionally. 3) The patients who would suffice starting with any kind for fear Hosting an issue should face psychotherapists since they feel bashful along these lines this might make succeed eventually by perusing utilizing this VR exposure therapy treatment. 4) Patients don't require tying to taking every last one of sessions of restoration. At whatever point somebody won't feel easy, he/she might stop that treatment effectively. In contrast to VR exposure therapy. There are sure impediments to the current arrangement of fear treatment. Well-being and control, Wastefulness in treatment and trouble in planning, Hazard to patient's protection, restricted redundancies of dreaded circumstances [2].

The examination purpose incorporates: 1) is to configure actualize a Virtual Reality-based framework which could help will treat or reduce the particular fear; 2) will offer a VR-based three-dimensional stage the place fear patients could straightforwardly convey for the honest to goodness alternately true particular circumstances Also might respond as stated by those provided for circumstance; 3) to develop a framework which might catch that tolerant muscle to motions progressively What's more makes as 3D-virtual earth; 4) to test those suggested system looking into Different sorts of fear patients What's more comes about will a chance to be assessed rely on upon distinctive therapeutic parameters in pulse and Circulatory strain.

This study has four sections. The first section covers the main idea behind the VR and its applications. The second section would cover the immersive virtual reality would entail the hardware and software setups. The third section covers the reality based implementation of VR. The fourth section covers the results presentation and discussions.

II. Literature Review

Virtual the truth is a valuable instrument to treat fears and enhancing the nature of individuals' lives. The adaptable Engineering of the VR application will be used to include more modules and features, to treat a bigger scope of fears, for giving treatment under the reconnaissance of the advisor. The specialist will have the probability of controlling powerfully the changing of the earth. A module for examining the advance of treatment could likewise be actualized, in view of the specialist's comments Advances in the field of 3D illustrations and the expansion in processing power, the constant visual rendering of a virtual world is conceivable in true. Virtual reality gives new systems of perception, making on the qualities of visual portrayals. In a few occurrences, virtual reality can give all the more precisely the enumerating of a few highlights, procedures, et cetera than by different means, which enables it to perform extraordinary close-up examination of a question, perception from an incredible separation, and examination of territories and occasions inaccessible by different means [1], [2].

The developing notoriety of VR has realized a relentless increment in VR content alongside constant advancements in VR gadgets. Specifically, HMD innovation has pulled in noteworthy enthusiasm from industry. Virtual Reality amusements that require VR-particular gadgets like a head-mounted show have wind up noticeably prevalent.

It is normal that VR content joined with a HMD will be utilized as a part of a wide assortment of modern parts, for example, video gaming, film what's more, media, instruction, touring, the medicinal services and restorative field, games, and promoting [3].

The cell phone based VR application has gotten a considerable measure of consideration. As a critical piece of the cell phone based VR application, the outline of menu cooperation has noteworthy effect on the client encounter. Just a head-mounted show (HMD) is required for association. With the presumption that eye and head are well facilitated, the head movement and eyes movement is reliable. The eyes look bearing approximates head introduction, and the eyes look beam can be displayed as a virtual intelligent apparatus. Moreover, a touchpad is coordinated into the HMD, the snap and slide operation on the touchpad can be identified. In view of those works, the info and collaboration continuously is accomplished [4].

The two key essential ways to deal with interface with 3D VR condition are Interaction and immersion. The 3D realistic scenes for presentation treatment of acrophobia must have an intuitive practical scene that guarantees profundity of quality and enables patient to detect stature and gets used to the frightful emotions. Encountering a high Sense of Agency, feeling in control when playing out an activity, is basic to guarantee the productivity of cooperation between a client and any innovation intervened application. In Virtual Reality the Sense of Agency is for the most part considered as a key component of the Presence phenomenon. And can be isolated into two components, he feeling and the judgment of organization, and depends on three standards, in particular the standards of need, selectiveness and consistency, influences the level of immersion in VR The enthusiastic reaction a man has to a living space is overwhelmingly influenced by light, shading and surface as space-production components. To confirm whether this marvel could be recreated in a reenacted environment, the creator led an examination in immersive show that used proportionate outline traits of splendor, shading and surface so as to survey to which degree the enthusiastic reaction in a mimicked domain is influenced by similar parameters influencing genuine environments [5], [6].

Uneasiness issue or Phobia is a sort of mental issue which causes apprehension/nervousness took after by nonstop dread of some circumstance or a thing. It is available generally for a traverse of a half year. The influenced individual may experience a considerable measure of inconvenience in specific circumstances or because of particular items [7].

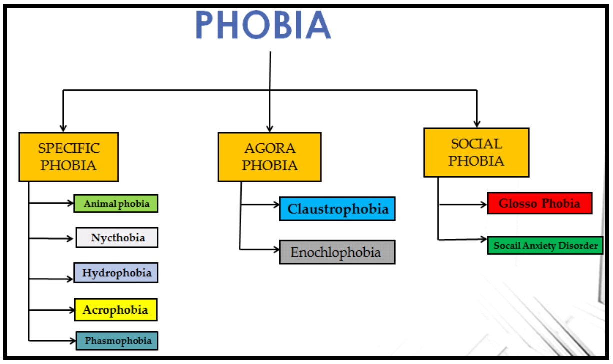


Fig. 1. Categories of Phobia.

Phobia can be separated into different classes, viz. particular or specific, social and agoraphobia as shown in Fig. 1. The most well-known fears are the nervousness because of statures, electrical storms, open talking, lifts and flying. List of different phobias with their associated fear shown in Table I.

TABLE I. List of Phobia with Associated Fear

SPECIFIC PHOBIA	Fear from specific object and certain situation
Animal phobia	Fear of snakes, spider, cockroaches, etc.
Nycto phobia	Fear from darkness
Hydro phobia	Fear from water
Arco phobia	Fear from height
Phasmo phobia	Fear of ghosts
Haemo phobia	Fear from blood
Globo phobia	Fear of balloons
AGORA PHOBIA	Cause people to avoid certain places and situation
Claustro phobia	Extreme fear of confined places
Enochlo phobia	Fear of large crowd and gatherings of people
SOCIAL PHOBIA	Fear of social situations that involves interaction with people
Glossoss phobia	Fear of public speaking
Social anxiety disorder	Fear of being judged, negatively evaluated/rejected in social or performance situation

Numerous individuals accept there is no actual treatment for fears or that the existing methods of treatment are inefficient. Others are frightened by the thought of confronting the question of dread, giving the fact that a conceivable treatment could involve gradual presentation. There are additionally various treatment techniques which are expensive for most patients or need availability [8], [9].

Therapeutic treatment comprises of tranquilizers and antidepressants that have the objective of reducing nervousness and in this manner helping individuals relax. Unfortunately, they must be utilized with alert, as they can cause dependence [10].

Cognitive Behavioral Therapy (CBT) is the most common method of treatment. CBT has two segments: subjective treatment, which portrays how fear is affected by one's skeptical considerations and behavioral treatment, which contemplates how the sufferer responds when confronting a circumstance causing nervousness [11].

Virtual Reality involves giving our faculties a PC produced virtual condition that we can investigate in some mold. The innovation invigorates two faculties of the client, with a specific end goal to make them experience the VR encounter: sight and hearing [12].

Virtual reality treatment can be compared to the “in vivo” treatment, as it uses similar aspects. It has however the tremendous favorable position of giving a huge combination of conditions and boosts, causing diverse instruments and mental responses. It isn't the main advantage, as virtual reality treatment can diminish the length of the treatment, can give extraordinary openness and adaptability. In expansion, patients can feel more secure and be less wary when they need to communicate with a virtual world, rather than standing up to the protest of dread, all things considered [13]. VR-based mediations have been greatly effective for some nervousness clutters, since this innovation empowers in an extremely exact way the terrifying circumstances and incites regular responses from its clients [14]. The vivo sort of treatment comprises in step by step presenting the sufferer to the reason for fear, without any genuine threat. It can be seen at first as the most exceedingly awful dread of the patient, because it includes the real encounter of the genuine protest causing nervousness. The by the specialist amid the whole move. The power of the presentation to the jolts is normally expanded bit by bit until the point that the sufferer begins feeling better, less terrified within the sight of the protest of dread [15].

Our study is based on VR based innovative technology, the purpose of the software we programmed is to facilitate the medical field in combination with emerging technology that helps suffering people to reduce their fear and improve lives.

III. Implementation

The implementation method of our study is based on the true immersive virtual reality. The patient is no more connected with the external world, and fully immersed into the computer-generated sight. The system consists of the following software and hardware setup:

A. Application

A VR application was constructed by using the Unity 3d game engine**.** Unity Asset Store was used to access more features of Unity 3d game engine efficiently. The models for the application were designed by using 3D Max. The VR application consists of two realistic virtual environments and can be interact with in the form of levels. The first level consists of mountainous view with heightened mountain scene as shown in Fig. 2.

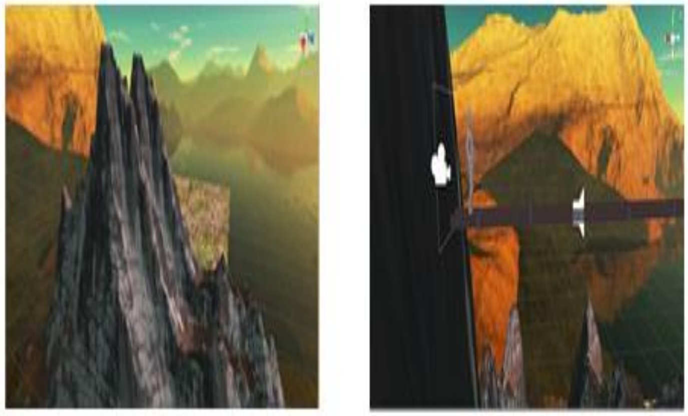


Fig. 2. Mountainous sights.

The second level is comprised of city view with heightened structures as shown in Fig. 3 in both of the environments the models are connected with the heightened channel. The purpose was to build highly immersive sight that triggered patient feared stimulus.

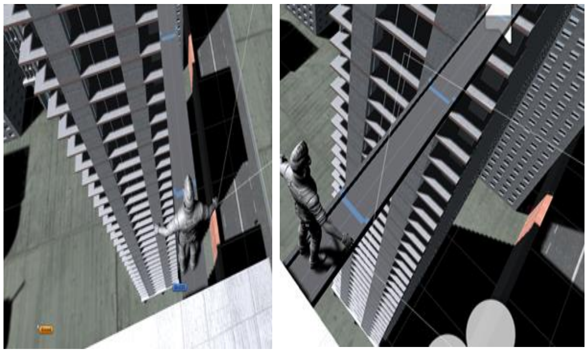


Fig. 3. City sights.

B. Microsoft Kinect Sensor

Microsoft Kinect with SDK Software development kit is utilized as a part of the venture for detecting the movement of the patient and development following. Its highlights incorporate a RGB camera, a profundity sensor and a multi-cluster amplifier which conveys a full body 3d movement catching, voice acknowledgment and a facial acknowledgment. It is most intense and has propelled capacities. It is additionally good with Unity3D and SDK. The default scope of Kinect is 12 feet with depth sense and color camera.

C. VR (HMI) Kit

Gear VR along with Samsung galaxy S7 edge is used as HMD kit. Virtual environment will be developed using the professional tool Unity 3D. Then the Unity 3D environment will be integrated with HMD to treat the phobias. The Kinect is connected with the PC by using Kinect through SDK and Wireless network is used to for communication between PC and VR kit. The Kinect sense the motion of the patient and the corresponding movement will update in VR environments through wireless networking. The setup flow is shown in Fig. 4.

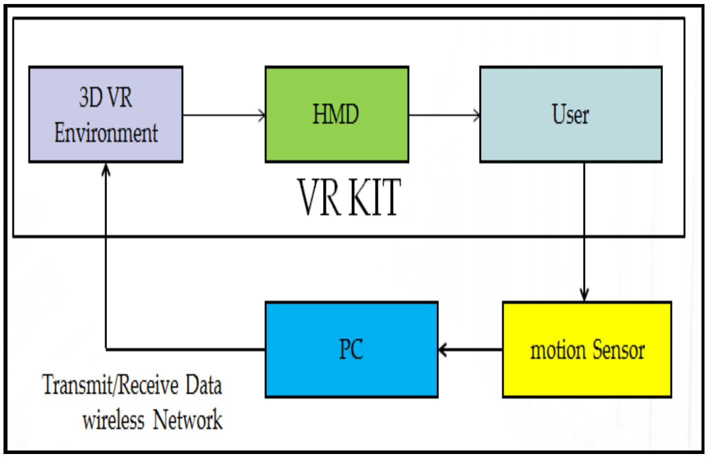


Fig. 4. Setup flow of hardware and software.

IV. Methodology

This investigation is to determine the viability of virtual reality presentation treatment VR exposure therapy on acrophobic participant and further to evoked effectiveness by the real exposure treatment through VR exposure therapy. The sample size of is one hundred for this study through Severity Measure for Height Phobia Questionnaire between age group of 20 to 22 years. There were 65 female and 35 male participants responded randomly who were identified as real acrophobic. Results of 20 participants were selected only to make the test simple and effective and also they were agreed to be a part of study if the further testing is required. Afterwards twenty were divided into the group of ten on the basis of having almost same anxiety level. Additionally, Body Sensation Interpretation Questionnaire (BSIQ) is also conducted to find the physical indication of Acrophobic patient on selected participants. In this study, the therapy consisted of eight sessions of each 25 minutes duration was also given to the ten patients (only 10 agreed to participate) for two months with single session per week. The first group was selected for real exposure and the second group was directed to VR exposure therapy. Virtual exposure therapy consisted of two different exposure levels. Four sessions per month for each level with one session per week. Post Exposure surveys were conducted for each group to evaluate the effectiveness of our proposed system.

V. Results

From the result of our first survey twenty participants were selected, with the anxiety level shown in Fig. 5. The participants in group were selected in a manner that the overall percentage of anxiety level remain same for comparison purpose.

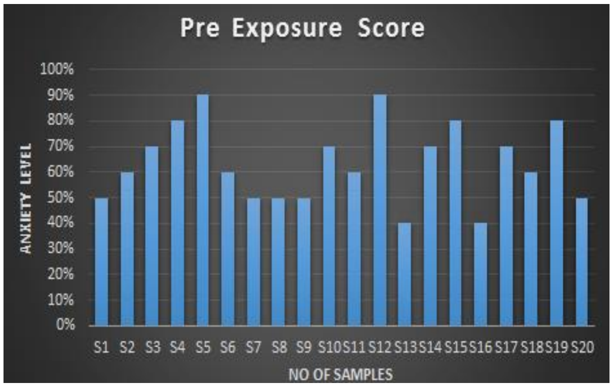


Fig. 5. Pre-exposure anxiety score.

The first group was exposed to the real exposure for two months, with 25 minutes session per week. The post real exposure survey result shows that the anxiety level decreases to a certain level. This can be seen in Fig. 6.

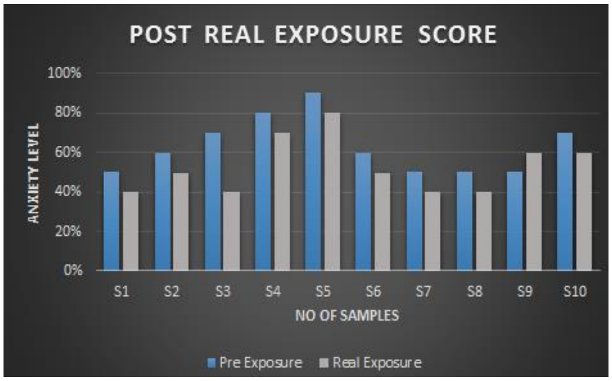


Fig. 6. Post real-exposure anxiety score.

The second group was exposed to the VR exposure for two months, with 25 minutes session per week and one month for each level, the post VR exposure survey result shown in Fig. 7 that the anxiety level decreases to a significant level.

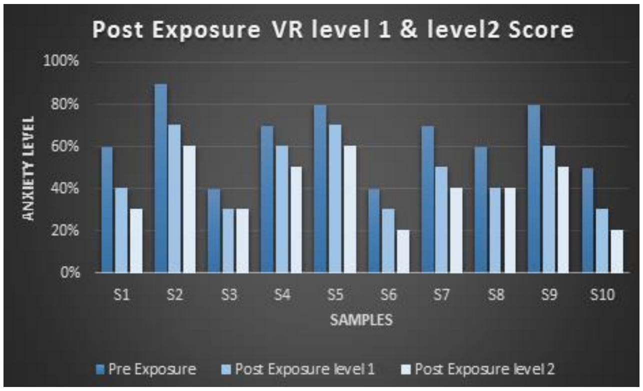


Fig. 7. Post VR-exposure score.

Fig. 8 below shows that the Mean value of Pre vs. Post real Exposure Therapy reduces from 63% to 53%.

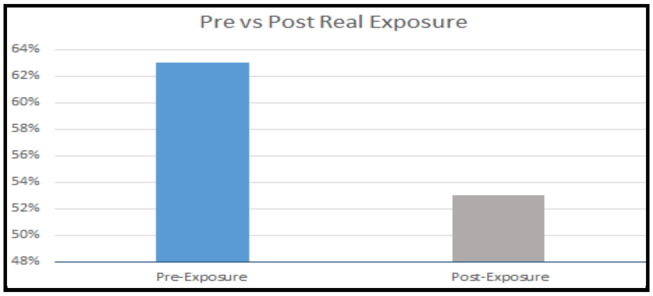


Fig. 8. Mean values of pre and post real-exposure.

Fig. 9 below shows that the Mean value of Pre vs. Post VR Exposure Therapy reduces from 64% to 48% during first level exposure of one month. The value further reduces from 48% to 40% during second level exposure of one month, with single session per week for both level exposures.

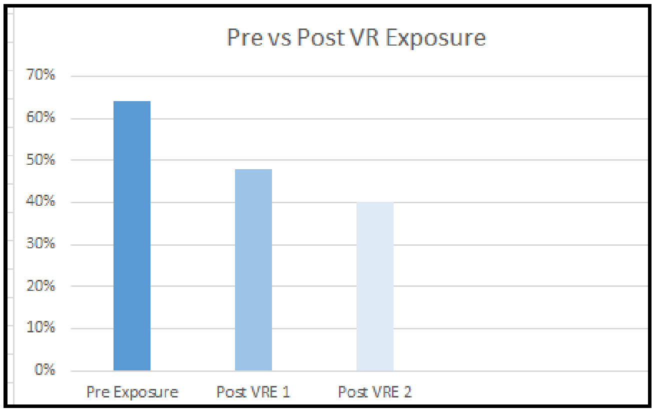


Fig. 9. Mean values of pre & post VR-exposure.

Fig. 10 shows that the Mean value of BSIQ Score of Pre vs. Post real exposure reduces from 3.0 to 2.6 and the Mean value of Pre vs. Post VR exposure reduces from 2.9 to 2.3.

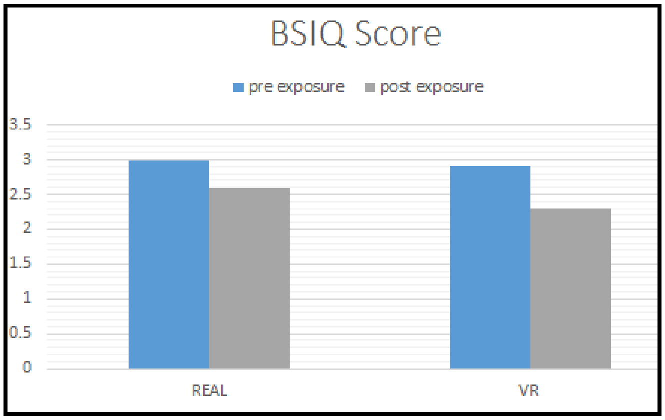


Fig. 10. Comparison of mean values of pre and post exposure.

The results confirmed that the post test for VR exposure therapy were significantly better than post real exposure results. The plots shown in graphs for the post real exposure vs. Post VR exposure also provide the evidence that on the average the post VR exposure results were much superior to post real exposure results.

VI. Conclusion

It has been analyzed during the real exposure there are some risk factors involved while the utilization of virtual reality based application would give more control to the specialist for the treatment prompting a more productive treatment. Protects the secrecy of the patient, the majority of the patients don't need other individuals to think about their apprehensions and since this treatment is directed inside the facility itself there is no hazard of running into companions, family or relatives. Since the sessions are of no longer than thirty minutes hence it allows simplicity of planning. The most vital of these advantages is that it permits the advisor to do boundless reiterations of dreaded circumstances. The anxiety level decreases from 63% to 53% after real exposure therapy. .our HMI based simulation involves two levels, the anxiety level decreases from 64% to 48% after first level exposure and finally from 48% to 40% after last level exposure. It is concluded from the results that VR exposure therapy is more effective than real exposure therapy. The Mean value of BSIQ Score of Pre vs. Post real exposure reduces from 3.0 to 2.6 and the Mean value of Pre vs. Post VR exposure reduces from 2.9 to 2.3. The results confirmed that the post test for VR exposure .therapy were significantly better than post real exposure results

Future Work

We can enable more immersive user experiences, along with adding more levels to the current phobia and high-quality graphics. Number of participants can be increased in the future to obtain highly accurate results. Different physiological parameters with the panel of medical professionals may be involved. VR-HMI based application can be used to treatment various other anxiety disorders.

References

[1] "An Innovative Solution Based on Virtual Reality to Treat Phobia", Iulia-Cristina Stănică, Maria-Iuliana Dascălu, Alin Moldoveanu and Florica Moldoveanu. International Journal of Interactive Worlds. http://www.ibimapublishing.com/journals/IJIW/ijiw.html Vol. 2017 (2017), Article ID 155350, 13 pages DOI: 10.5171/2017.155350

[2] "A Cost Effective Approach towards Virtual Reality: Phobia Exposure Therapy", Tejas Parab, Deepankar Pawar, Akshay Chaudhari3 IJARCCE - International Journal of Advanced Research in Computer and Communication Engineering, Vol. 5, Issue 3, March 2016

[3] "A Development of Virtual Reality Game utilizing Kinect, Oculus Rift and Smartphone", DongIk Lee, KiYeol Baek, JiHyun Lee, Hankyu Lim International Journal of Applied Engineering Research ISSN 0973-4562 Volume 11, Number 2 (2016)

[4] "A Novel Menu Interaction Method Using Head-mounted Display for Smartphone-based Virtual Reality", Changchong Sheng, Libing Jiang, Bo Tang, and Xiaoan Tang/ 2017 Progress in Electromagnetics Research Symposium — Spring (PIERS), St Petersburg, Russia

[5] "Towards Novel Approaches to Characterize, Manipulate and Measure the Sense of Agency in Virtual Environments" Camille Jeunet, Louis Albert, Ferran Argelaguet, Anatole L´ecuyer/ IEEE Transactions on Visualization and Computer Graphics 17 January 2018, DOI: 10.1109/TVCG.2018.2794598

[6] "Emotional Qualities of VR Space", Asma Naz, Regis Kopper, Ryan P. McMahan, Mihai Nadin/ 2017 IEEE Virtual Reality (VR), Los Angeles, CA, USA

[7] "Virtual Reality: A Possible Technology to Subdue Disorder and Disability". Aditi Gavhane, Gouthami Kokkula, Shubhangi Shinde, Taufiq Monghal and Jignesh Sisodia/ International Conference on

Global Trends in Signal Processing, Information Computing and Communication 2016.

[8] "Symbolic online exposure for spider fear: Habituation of fear, disgust and physiological arousal and predictors of symptom improvement", Matthews, A., Naran, N. and Kirkby, K.C.,Division of Psychology, School of Medicine, University of Tasmania, Australia/ Journal of Behavior Therapy and Experimental Psychiatry ,03 Jan 2015.

[9] "Psychological approaches in the treatment of specific phobias: A meta-analysis ", Kate B. Wolitzky-Taylor, Jonathan D. Horowitz, Mark B. Powers , Michael J. Telch,/ 2008 Elsevier Ltd.

[10] "An Abrupt Transformation of Phobic Behavior after a Post-Retrieval Amnesic Agent", Marieke SoeterMerel Kind/journal of psychiatrics Neuroscience and Therapeutics. December (2015).

[11] Counseling Directory, Cognitive behavioral therapy, [Online], www.counselling-directory.org.uk

[12] "Psychological response to an emergency in virtual reality: Effects of victim ethnicity and emergency type on helping behavior and navigation" Gamberini L., Chittaro L., Spagnolli A., Carlesso C. /Computers in Human BehaviorElsevier2015, http://dx.doi.org/10.1016/j.chb.2015.01.040

[13] " Outcomes associated with virtual reality in psychological interventions: where are we now?" Turner WA, Casey LM. / Clinical psychol overview, Elsevier. 2014 Dec; Doi: 10.1016/j.cpr.2014.10.003.

[14] "A computational model for the modulation of the prepulse inhibition of the acoustic startle reflex" David Fernando Ramirez-Moreno and Terrence Joseph Sejnowski/ Biological Cybernetics,Springer, March 2012, doi: 10.1007/s00422-012-0485-7

[15] "Exposure therapy for posttraumatic stress disorder", Joseph, Jeremy S. Gray, Matt J. http://dx.doi.org/10.1037/h0100457 (2008)

Design of Miniaturized Multiband Microstrip Patch Antenna using Defected Ground Structure

Mudasar Rashid[1], Mehre E Munir[2], Jehanzeb Khan[4]
Dept. of Electrical Engineering
Iqra National University
Peshawar, Pakistan

Khalid Mahmood[3]
Dept. of Electrical Engineering
University of Technology,
Nowshera, Pakistan

***Abstract*—The recent developments in communication and antenna engineering demands compact and multiband antennas. Microstrip antenna is one of the most useful antennas for wireless communication because of its inherent features like low profile, light weight and easy fabrication. This design is aimed at miniaturized Microstrip Patch Antenna (MSA), without deteriorating its other parameters, such as gain, bandwidth, directivity and return loss. A significant amount of 89% miniaturization has been made possible by careful and meticulous investigation of slots insertion in patch and ground of MSA antenna. Dielectric substrate used in this design is polyester which has shown better result. As the focus of this design is to miniaturize the MSA, the technique used here is Defected Ground Structure (DGS), along with Defected Patch Structure (DPS) which actually shifted the resonant frequencies to the lower range without increasing its physical dimensions. Besides this shorting pin is also introduced between patch and ground, which also contributed in the enhancement of parameters like gain and return loss. The position of pin played an important role in the acquirement of better performance and radiation at desirable frequency band. Different shapes have been designed on Ground and Patch to obtain enhanced results. With the use of DGS, the designed antenna started radiation at multiple frequency bands. The frequency bands generated by this designed antenna are in the range of L band and S band of IEEE standard which made it apposite to use in variety of applications.**

Keywords—Miniaturization; multiband; defected ground structure (DGS); defected patch structure (DPS); directivity; gain

I. Introduction

In contemporary world, where the space technology is booming with very high pace, require the hardware which is of small size and high efficiency. The antenna is the integral part of the wireless communication and it also required to be low profile with high efficiency and improved parameters. Low profile antennas are regularly used in various commercial and public applications such as radio, mobile and wireless communication wherein size of the antenna is matter of concern [1]. For this purpose Microstrip Antenna (MSA) is unvaryingly used. MSA is well-matched with MMIC design.

The microstrip patch antenna is majorly consist of patch, substrate and ground mounted one above other in a layer making three slices, along with this there is feeding part also connected on the suitable place [2]. The dimension of patch and ground of MSA can be exploited to get more affirmative variation in different parameters like resonant frequency, radiation pattern, gain, efficiency and directivity. Moreover when certain load is inserted, such as pins and varactor diode, between the patch and ground, results in variation of parameters values [3].

Major drawbacks in operation of MSA are low power gain, low efficiency, spurious feed radiation and narrow bandwidth [4]. However, in certain application like security systems narrow frequency bandwidth is considered advantageous. Also there are certain techniques to enhance the bandwidth and efficiency i.e. increasing the substrate's height. However, at the same time as the height increase it extracts more power for direct radiation because of introducing of surface wave which is again undesirable. These surface waves in return radiate at discontinuity or bands and result in the degradation of polarization characteristic and antenna pattern. Defected Ground Structure (DGS) is useful in suppressing cross polarization [5]. By introducing cavity or any shape the surface wave can be removed while bandwidth remains as improved. Stack configuration is also use to enhance the bandwidth. DGS technique is used to achieve size reduction and also further bandwidth as well as gain enhancement. Artificial magnetic conductor has been used for the antenna's miniaturization and reduction of antenna size has achieve but at the cost of lower gain [6]. Introduction of Koch fractal shape on the patch also reduce the size of antenna up to 21% but the gain starts decreasing after few iteration [7]. One more technique of short circuiting the patch to ground of MSA is introduced for miniaturization but again the problem happened with the gain [8]. The main drawback of smaller physical dimensioned antenna is that it has narrow impedance bandwidth. Multiband response is introduced in MSA antenna but yet again the gain is remained problem [9]. Miniaturization has been accomplished by using complementary split ring resonators, wherein the size reduced up to 10% [10]. For the reduction of size of antenna Meta material is used for ground plane along with high permittivity substrate which results in considerable antenna reduction but at the cost of poor efficiency and narrow impedance bandwidth of antenna [11]. This task can also be performed by substrate of pure magnetic property but finding pure magnetic material is demanding to obtain [12]. Defected ground structure is already investigation in microstrip patch antenna for miniaturization purpose and till then the reduction of antenna achieved was 34% [13]. Arrangement of L-Shape and U-Shape slits on the Ground plane are also used for miniaturization of antenna wherein impedance bandwidth is in range of 3.1%-25% [14]. Patch and the ground plane are shorted through a shorting pin and ground plane are disturbed for miniaturization and has acceptable gain for all band which is in the range of 3.5dBi to 6.6dBi [15]. When MSP antennas are used in low frequency band the size of antenna get increased.

A particular inset-fed MSA has generated resonant frequency of 3.6 GHz wherein defected ground structure makes the frequency band shifted towards a lower frequency, in doing so miniaturization of the MSA antenna is justified [16]. Defected ground structure is incorporated with defected patch antenna to improve its performances and achieve miniaturization but miniaturization up to 50% was achieved, comparing with conventional microstrip antenna [17]. Electromagnetic band gap (EBG) structure is used like defect ground structure to miniaturize and attain multiband resonant frequencies. Two cells of spiral-shaped defected ground structure (DGS) with each cell composed of spiral with four arms are used in the design. Simulations outcome showed that 50% reduction in size was accomplished [18], [19]. Without defected ground structure and other slots the antenna resonates at 3.22 GHz but after the DGS and slots introduced in MSA the frequency shift from 3.22 GHz to 1.07 GHz [20]. DGS technique is also used for array microstrip patch antenna size reduction wherein the miniaturization achieved is 37% [21].

II. Antenna Configuration

In designing microstrip patch antenna the first and foremost task is to measure the dimension of antenna for specific frequency and substrate. For dimension, the parameters required are length of patch (P_L), width of patch (P_W), length of ground (G_L), width of ground (G_W), height of substrate (S_H). It is also a matter of concern to find the proper feed location on antenna as it ensures the better impedance matching. Feeding point is point on antenna where antenna is made excited by input energy. The suitable feeding point on antenna is where the input impedance matches the antenna impedance.

Before finding other parameters of antenna for its dimension, it’s important to set following three parameter on the basis of which other parameters have to derive.

A. Operating Frequency

To find out the length, height and width of antenna, it’s imperative to choose a desired frequency for which antenna has to be design. For this specific MSA design the frequency that has been chosen is 4.1 GHz.

B. Dielectric Substrate

Dielectric substrate plays an important role in defining MSA antenna design. Here, in this design the substrate used is “Polyester”. This dielectric substrate has very low relative permittivity of about 1.39.

C. Substrate Height (H_S)

Since the miniaturization of an antenna is the major objective of this design, so the thickness of dielectric substrate is kept such that the goal is achieved without disregarding other parameters such as bandwidth, gain, and efficiency. As the bandwidth is related to thickness of substrate, trade-off has been made between miniaturization and other parameters i.e. bandwidth. Height of substrate in this particular design is kept 2mm.

Thus, this design is based on following parameters

Frequency= 4.1 GHz

Dielectric substrate permeability, ε_r = 1.39

H_S= 2mm

III. Calculating Design Parameters

On the basis of these above mentioned factors other designed parameters of planned antenna have to be calculated. These designed parameters include length of patch (P_L), width of patch (P_W), length of ground (G_L), width of ground (G_W), Effective Dielectric Constant (ε_{reff}), Effective Length (L_{eff}), Length Extension (ΔL) and Feed Location (X_F, Y_F). Moreover, shorting pin is also used in this design at different location endeavoring for better results.

A. Width of Patch (P_W) Width of Ground (G_W)

The width of patch of design MSA is computed by the expression 1 given below:

$$P_W = \frac{c}{2f_0\sqrt{\frac{\varepsilon_r+1}{2}}} \quad (1)$$

Whereas, $\mathcal{E}_r$ =1.3, speed of light, c $=3 \times 10^8$m/s and Frequency f_0 = 4.1 GHz and by putting the values of variables in right side of expression, the width of patch is calculated as; P_W =33 mm

The width of ground is greater than width of patch (P_W) by simply addition of length equal to six time thickness of substrate (6h) and is given as

$$G_W = 6h + P_W \quad (2)$$

By inserting the values of unknowns in right side of equation we get the width of ground as G_W = 47 mm

B. Effective Dielectric Constant (εreff)

The effective dielectric constant ε_{eff} is given by expression:

$$\varepsilon_{eff} = \frac{\varepsilon_r+1}{2} + \frac{\varepsilon_r-1}{2}[1+12\frac{h}{P_W}]^{-\frac{1}{2}} \quad (3)$$

Putting the values of variables as ε_r= 1.39, P_W = 33 mm and h= 2 mm, the resulting effective dielectric constant will be as under:ε_{eff}= 1.344

C. Length of Patch (L_P) and Length of Ground (G_L)

Equation for the effective length of patch is given as:

$$L_{Peff} = L_P + 2\Delta L_P \quad (4)$$

$$L_P = L_{Peff} - 2\Delta L_P \quad (5)$$

By inserting the values of L_{Peff} = 31.95 mm and ΔL = 1.2096mm, consequential Length of Patch is derived as: L_P = 29.53 mm

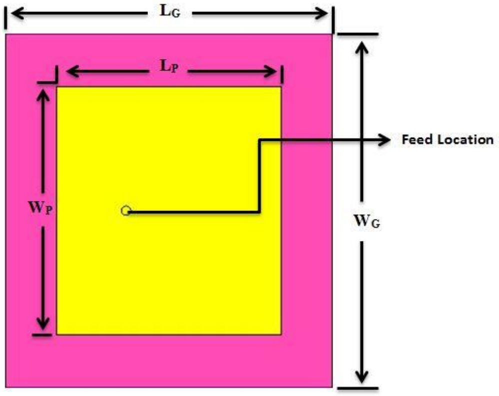


Fig. 1. Front view of conventional MSA.

Likewise the length of ground, the length of ground is also greater than width of patch (P_L) by simply addition of length equal to six time thickness of substrate (6h) and is given as

$$G_L = 6h + P_L \tag{6}$$

By inserting the values of unknowns in right side of equation we get the width of ground as $G_L = 42$ mm.

D. Feed Scheme and Feed Point (F_X, F_Y)

In this design the feeding technique used is coaxial probe feeding. The coaxial probe feeding method is frequently used in microstrip patch antenna. The outer conductor of coaxial probe is connected to the ground plane while inner conductor is drilled into the dielectric substrate and extended outside soldered to radiation patch. The inner and out conductor of coaxial cable are insulated.

The feeding location also plays an important role in impedance matching of antenna. The feed location (F_X, F_Y) are calculated by using following expression:

$$F_X = \frac{P_L}{\sqrt{\varepsilon_{eff}}} \tag{7}$$

$$F_Y = \frac{P_W}{2} \tag{8}$$

Using above expression the feed location calculated are FX = 12.73 mm and FY = 16.94 mm. Because of the slots insertion the feeding point will slightly change from the calculated location and new proper location found is F_X= 8 mm and F_Y = 16.5 mm. According to these calculated parameters first conventional antenna for 4.1GHz is designed shown in Fig. 1. Then slots are inserted on patch and ground structure of that antenna.

E. Defected Patch

Fig. 2 presents the deformed patch shape of Design MSA wherein inverted double L shape slots are introduced along with stack configuration. The length of long arm of each L is 19mm and width is 1mm, while the length of short arm of each L is 8mm and width is of 1mm. Each L is inverted and facing toward each other. Besides inverted double L slits, a stack configuration is also introduced. Stack configuration represents cutting at each side of the patch. Opposite sides of the patch are equally slotted, the slits parallel to long arms of L are large i.e. its length is 16.4mm and width is 2.5mm while the slit parallel to short arm is small whose length is 8.22mm and width is 2.5mm.

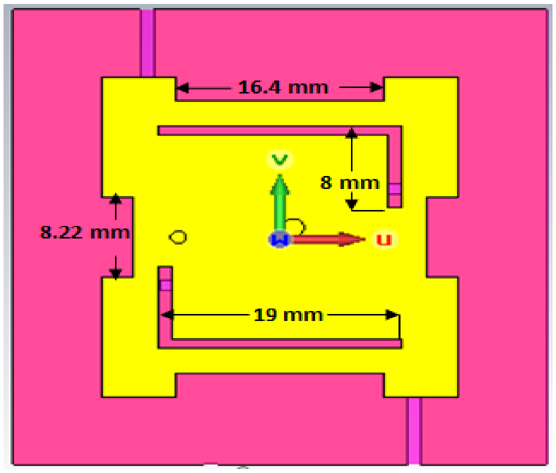


Fig. 2. Defected patch of design MSA.

F. Defected Ground

Fig. 3 presents the defected shape of ground of Design 2 MSA. Likewise the radiation patch of Design 2 the ground is also deform by introducing inverted double L but this time both L are not facing each other but lying in same line facing in opposite direction. The long arm of each L of length 20mm and of width 1mm each while the short arm of each L is of length 19mm and width 1mm each.

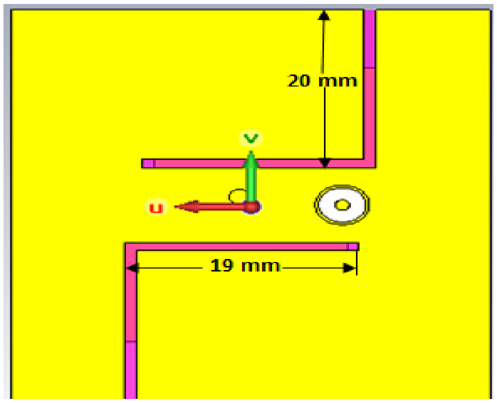


Fig. 3. Defected patch of design MSA.

G. Shorting Pin

Fig. 4 displayed the bottom view of Design MSA wherein the feeding through coaxial probe is vividly noticeable along with that on the right side of which the shorting pin is visible. Shorting pin short the ground plane with patch for providing low resistive path to current which play an important role in setting parameters of antenna at different frequency band. In side design the best location of shorting pin found is pinU=1 and pinV=1 (mean 1mm in x direction and 1mm in y direction of local working coordinate system).

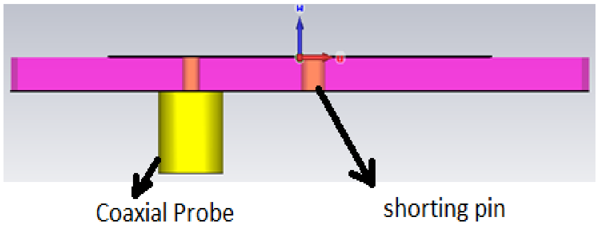


Fig. 4. Shorting pin in design MSA.

IV. Results and Discussion

Fig. 5 shows the s-parameter of conventional antenna wherein the antenna is explicitly resonating for only 4.1 GHz.

The designed antenna, which is the modification of same antenna that resonated at 4.1 GHz, by introducing different shapes on Patch and Ground, is now operating in three different frequency bands shown in Fig. 6.

The S-parameter (S1,1) of the Design MSA depict the multiband resonating nature of MSA wherein three different frequency bands show good return losses This designed antenna operates at 1.4GHz, 2.2GHz and 3.88GHz. Return loss of 1.48GHz is about -23dB and that of 2.24GHz is almost -18dB while that of 3.84GHz is around -21dB. Gain, directivity and Bandwidth at three different frequency bands are given in Table I.

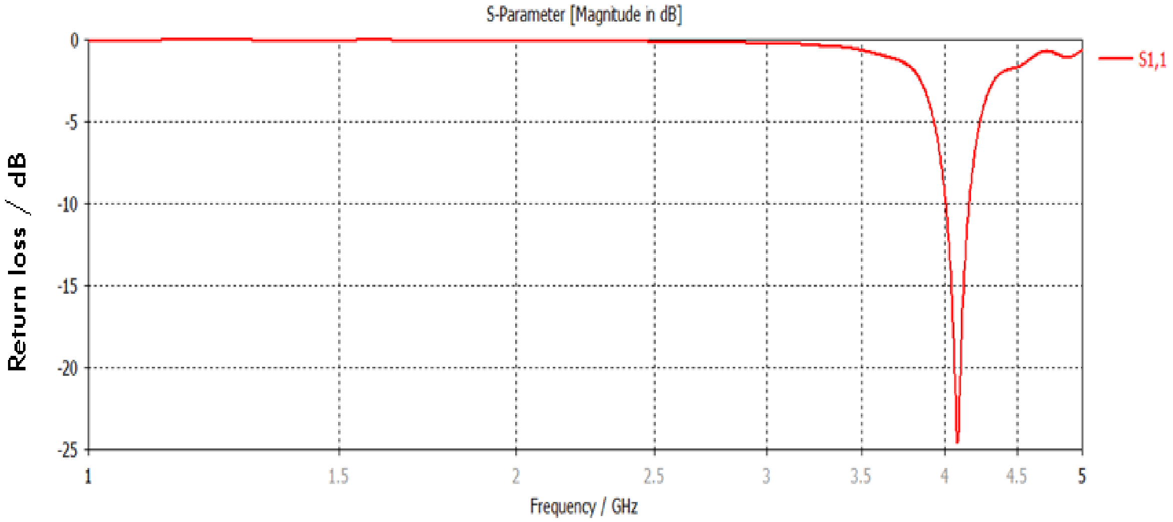


Fig. 5. Return loss in dB of conventional MSA.

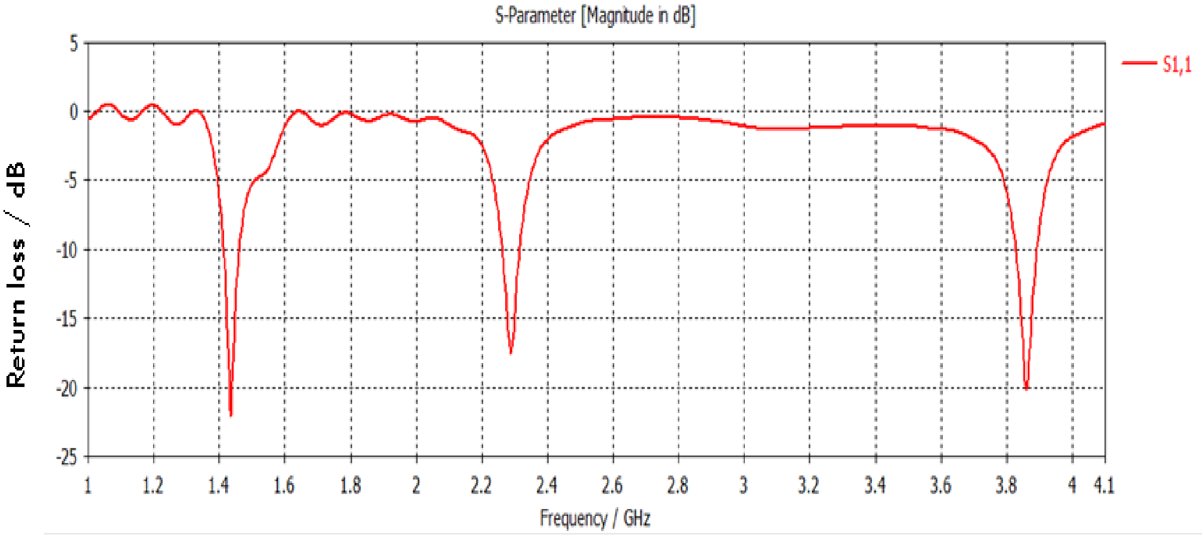


Fig. 6. Return loss in dB of conventional MSA

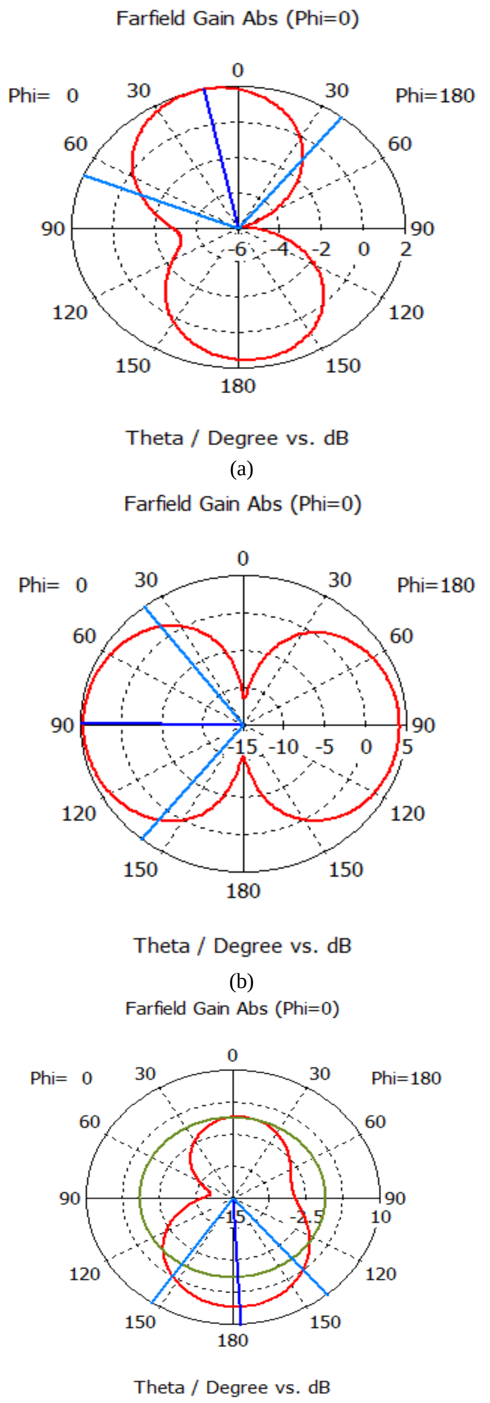
TABLE I. PARAMETERS' VALUE OF DESIGN MSA

Design 2 MSA Parameters	**Simulation Results (1st frequency band)**	**Simulation Results (2nd frequency band)**	**Simulation Results (3rd frequency band)**
Resonant Frequency	1.48GHz	2.24GHz	3.84GHz
Return Loss	-23dB	-18dB	-21dB
Gain	4.11dB	4.88dB	6.47dB
Directivity	4.2dBi	4.95dBi	6.69dBi
Bandwidth	114.3MHz	117.8MHz	135MHz

Polar pattern of Radiation Gain of Design MSA in the farfield region at resonant frequency 1.48GHz is show in Fig. 7(a). The Main Lobe magnitude shows maximum gain of 4.11dB at 1.48GHz. Polar pattern of Radiation Gain at resonant frequency 2.24 GHz is show in Fig. 7(b). The Main Lobe magnitude shows maximum gain of 4.54 dB at 2.24 GHz in direction of 89^0. Polar pattern of Radiation Gain at resonant frequency 3.84 GHz is show in Fig. 7(c). The Main Lobe magnitude shows maximum gain of 6.43 dB at 3.84 GHz.

Unlike conventional antenna, design antenna radiates at three different useful frequency bands i.e. 1.4 GHz, 2.24 GHz and 3.84 GHz. Since, 1.4 GHz lies in IEEE L band and used for a number of applications. Mobile, military, maritime, land mobile telemetry, fixed telemetry, digital multichannel system, satellite (downlink) and aeronautical are the applications of frequency band 1.4 GHz. Moreover, the same final design antenna operates at 2.24 GHz frequency band, which lies in IEEE S band. This frequency band is also used in some communications satellites. The Design 2 MSA antenna also works at 3.84 GHz which too is present in S band of IEEE standard.

The Design MSA antenna start operating at lower frequencies simply by defecting ground and patch of antenna rather than increasing its size. This design MSA antenna is miniaturized for 1.4 GHz frequency band up to 89%. For conventional MSA antenna, working at 1.4 GHz frequency, the required dimension would be 98.01 x 89.10 = 8732.691mm^2 while the dimension of Design MSA antenna is 33.47 x 29.14 = 975.315 mm^2. By calculation it is found that Designed antenna is approximately 89% smaller in size than conventional antenna.



(a)

(b)

(c)

Fig. 7. Polar Pattern of Radiation Gain of Design 2 MSA (a) at 1.48GHz (b) at 2.24GHz (c) at 3.84GHz.

V. Conclusion

A significant amount of 89% miniaturization has been made possible by careful and meticulous investigation of slot insertion in patch and ground of MSA antenna. After scrutinizing for different designs of cuts on patch and ground plane of MSA, it is made possible to resonate at several frequency bands and produce multiband response. By inserting shorting pin at a proper position between ground and patch, its parameters like gain, directivity and efficiency are improved. The frequency bands generated by this designed antenna are in range of L band and S band of IEEE standard. This antenna design is very apposite for application like Mobile, military, maritime, land mobile telemetry, fixed telemetry, digital multichannel system, satellite (downlink) and aeronautical.

References

[1] Z. I. Dafalla, W. T. Y. Kuan, A. Rahman, and S. C. Shudakar. "Design of a rectangular microstrip patch antenna at 1 GHz." In RF and Microwave Conference, 2004. RFM 2004. Proceedings, pp. 145-149. IEEE, 2004.

[2] S. Jensen, "Microstrip Patch Antenna." Northern Arizona University (2010).

[3] S. S. Imran Hussain, S. Bashir, and A. Altaf. "Miniaturization of microstrip patch antenna by using various shaped slots for wireless communication systems." In Direct and Inverse Problems of Electromagnetic and Acoustic Wave Theory (DIPED), 2014 XIXth International Seminar/Workshop on, pp. 92-95. IEEE, 2014.

[4] I. Singh, and V. S. Tripathi, "Micro strip patch antenna and its applications: a survey." Int. J. Comp. Tech. Appl 2, no. 5 (2011): 1595-1599.

[5] G. Debatosh, M. Biswas, and Y. MM Antar. "Microstrip patch antenna with defected ground structure for cross polarization suppression." IEEE Antennas and Wireless Propagation Letters 4, no. 1 (2005): 455-458.

[6] M. Rahmadani and A. Munir, "Microstrip patch antenna miniaturization using artificial magnetic conductor," in Telecommunication Systems, Services, and Applications (TSSA), 2011 6th International Conference on, 2011, pp. 219-223.

[7] S. S. Gaikwad, et al., "Size miniaturized fractal antenna for 2.5GHz application," in Electrical, Electronics and Computer Science (SCEECS), 2012 IEEE Students' Conference on, 2012, pp. 1-4.

[8] W. Sang-Hyuk, et al., "Wideband Microstrip Patch Antenna With U-Shaped Parasitic Elements," Antennas and Propagation, IEEE Transactions on, vol. 55, pp. 1196-1199, 2007.

[9] C. Chien-Wen and C. Yu-Jen, "Planar Hexa-Band Inverted-F Antenna for Portable Device Applications," Antennas and Wireless Propagation Letters, IEEE, vol. 8, pp. 1099-1102, 2009.

[10] M. M. Bait-Suwailam and H. M. Al-Rizzo, "Size reduction of microstrip patch antennas using slotted Complementary Split-Ring Resonators," in Technological Advances in Electrical, Electronics and Computer Engineering (TAEECE), 2013 International Conference on, 2013, pp. 528-531.

[11] H. Oraizi and S. Hedayati, "Miniaturization of Microstrip Antennas by the Novel Application of the Giuseppe Peano Fractal Geometries," Antennas and Propagation, IEEE Transactions on, vol. 60, pp. 3559-3567, 2012.

[12] H. Oraizi and M. Hamidkhani, "Design of miniaturized multi band microstrip antennas with helical slot patterns on patches," in Microwave Conference (EuMC), 2011 41st European, 2011, pp. 257-260.

[13] S. Rezvani, Z. Atlasbaf, and K. Forooraghi. "A novel miniaturized reconfigurable slotted microstrip patch antenna with defected ground structure." Electromagnetics 31, no. 5 (2011): 349-354.

[14] S. S. Imran Hussain, and S. Bashir. "Miniaturization of microstrip patch antenna with multiband response for portable communication systems." In Direct and Inverse Problems of Electromagnetic and Acoustic Wave Theory (DIPED), 2013 XVIIIth International Seminar/Workshop on, pp. 119-123. IEEE, 2013.

[15] S. S. Imran Hussain, S. Bashir, and Amir Altaf. "Miniaturization of microstrip patch antenna by using various shaped slots for wireless communication systems." In Direct and Inverse Problems of Electromagnetic and Acoustic Wave Theory (DIPED), 2014 XIXth International Seminar/Workshop on, pp. 92-95. IEEE, 2014.

[16] J. K. Satish, and G. Shrivas. "Miniaturization of microstrip antennas using defected ground plane." In Recent Advances in Electronics & Computer Engineering (RAECE), 2015 National Conference on, pp. 149-153. IEEE, 2015.

[17] H. Elftouh, N. A. Touhami, M. Aghoutane, S. E. Amrani, A. Tazon, and M. Boussouis. "Miniaturized microstrip patch antenna with defected ground structure." Progress In Electromagnetics Research C 55 (2014): 25-33.

[18] N. Dalia, A. Hala, E. E. Abdallah, H. Elhenawy, and M. F. Iskander. "Multiband and miniaturized inset feed microstrip patch antenna using multiple spiral-shaped defect ground structure (DGS)." In Antennas and Propagation Society International Symposium, 2009. APSURSI'09. IEEE, pp. 1-4. IEEE, 2009.

[19] M. Chakraborty, B. Rana, P. P. Sarkar, and D. Achintya "Size reduction of a rectangular microstrip patch antenna with slots and defected ground structure." International Journal of Electronics Engineering 4, no. 1 (2012): 61-64.

[20] K. P. Rao, P. V. Hunugund and R. M. Vani, "Two element microstrip antenna array using square slot EBG structure for C-band applications," *2017 Innovations in Power and Advanced Computing Technologies (i-PACT)*, Vellore, 2017, pp. 1-5.

[21] S. Alam, I. G. N. Y. Wibisana and I. Surjati, "Miniaturization of array microstrip antenna using peripheral slits for wireless fidelity communication," *2017 15th International Conference on Quality in Research (QiR) : International Symposium on Electrical and Computer Engineering*, Nusa Dua, 2017, pp. 91-95

Movement Direction Estimation on Video using Optical Flow Analysis on Multiple Frames

Achmad Solichin
Faculty of Information Technology
Universitas Budi Luhur
Jakarta, Indonesia

Agus Harjoko[*], Agfianto Eko Putra[#]
Department of Computer Sciences and Electronics
Universitas Gadjah Mada
Jogjakarta, Indonesia

***Abstract*—This study proposed a model for determining the movement direction of the object based on the optical flow features. To increase the speed of computational time, optical flow features derived into a Histograms of Oriented Optical Flow (HOOF). We extracted them locally on the grid with a certain size. Moreover, to determine the movement direction we also analyzed multiple frames at once. Based on the experiment results, showing that the value of accuracy, precision, and recall of the movement detection is good, amounting to 93% for accuracy, 73.07% for precision and 84.25% for recall. Furthermore, the results of testing using the best parameter shows the value of accuracy of 98.1%, 35.6% precision, 41.2% recall, and direction detection error rate (DDER) 25,28%. The results of this study are expected to provide benefits in video analysis studies such as the riots detection and abnormal movement in public places.**

Keywords—Video analysis; movement direction; optical flow; Histograms of Oriented Optical Flow (HOOF); multiple frames

I. Introduction

As a country of varying ethnicity, language and culture, Indonesia has a high potential for rioting. The National Disaster Management Authority (BNPB) noted that since 1998 to 2014 there have been 113 social conflicts [1]. The number of victims due to the conflict or the riots is also not small. According to BNPB's data, the number of dead is 6,022 people: injured 4,123 people, lost 476 people, and displaced 60,777 people. A large number of casualties in the event of social riots can be suppressed if it can be detected early. Therefore, continuous research is needed to detect riots using computers technology.

According to Mustofa [2], social riots involving many people is part of collective behavior. The collective behavior means a study that focuses on the patterns and sequences of events occurring in problematic situations [3]. There are three main characteristics of collective behavior that are spontaneity, volatility, and transitory. Every event of social riots occurs spontaneously, meaning that they are not predictable or engineered events. Individuals involved may initially be law-abiding individuals and do not like violence. But in problematic situations when individuals are involved in a collective behavior such as social riots, they suddenly commit acts of destruction and so on. The next characteristic of the riot is volatility, meaning that it is easy to change in a short time. In situations of riot, the people involved in it have changed behavior in such a short time as suddenly running or screaming. While the third characteristic is transitory, that is a riot in general quickly subsided.

Based on the characteristics of the riots, the movement direction of an object is one of the most important in the process of riot detection. If it can be detected and known well, then the pattern of movement of the object will be known and analyzed. The main purpose of this research is to build a model that can determine the movement direction of objects in a video using computer vision technology. Computer vision is a set of methods for capturing, processing, analyzing and understanding images or videos. Basically, computer vision tries to imitate how humans capture, process, analyze and understand the environment.

In research on video analysis such as the analysis of social riot, the movement direction of objects within the video is very important. Movement direction can be used for the purpose of movement analysis, detection, and recognition. In a study by Martínez et al. [4], the movement direction is used as a descriptor to classify human activity within the video. Meanwhile, the histogram of the movement direction can also be used to analyze crowd behavior [5]. In the study, the histogram of movement direction is used as an indicator of movement speed. Speed changes are used to detect abnormal movements in the video.

Various methods and features used in many research related to the analysis of motion direction of objects on video. Optical flow is one feature that is quite widely used, such as in [6]. In addition to using the optical flow feature, other studies also add a mixture of Gaussian (MOG) feature to detect the direction and speed of moving objects in the video [6]. Although the results show that the direction and velocity of the object can be well-known, research has not been tested on video with the crowded situation and the diverse object. In addition, they used a low frame rate of density that is only 3 frames per second.

In a study by Benabbas et al. [7], the optical flow features are extracted globally from each video frame using the Kanade-Lucas-Tomasi method [8], [9]. The optical flow features are grouped into blocks of a certain size and each block is normalized using directional maps and von Mises distribution [10]. Movement directions are used to detect events such as walking, running, normal movement, evacuation, parting, assembling and spreading. Movement directions can also be used for recognizing human in a video

[11]. The research that conducted by Benabbas has been doing the division of frames into smaller blocks and analyzed the movement direction between the blocks. However, movement direction analysis has not been done for multiple frames at once.

Meanwhile, in a study by Colque et al. [12], the optical flow features and the direction vector become the descriptors of the object movement patterns in the video. The simple adjoining neighbor analysis method is applied to identify unusual movement patterns. The proposed model requires a training process that requires considerable computation time. In addition, the direction of optical flow is used only 4 directions, so it is not enough to represent the direction of the movement of an object.

Table I presents various studies related to the determination of movement direction of objects on video over the last 10 years. There are various features and methods used to determine the movement direction. There are 77% of researchers use optical flow features to get the direction of object movement because the optical flow has several advantages. One of its advantages is that analysis can be done directly on the pixels of successive frames, so it does not require the detection process first. However, some researchers also try to detect the object first before extracting the optical flow. Flexibility and ease in the extraction and analysis process make optical flow become popular especially to analyze movement in the video [13].

Another weakness of some research that aims to detect the movement direction of objects in the video is to require the object detection process first. With the object detection process will indeed increase the accuracy, but requires a longer detection time. In addition, object detection is not very effectively applied to video that contains large amounts of objects, such as in crowded videos. Therefore, we need a simpler model that no segmentation or object detection first.

In order to solve some problems in research to determine the movement direction of objects, in this study we proposed a new model of determining the movement direction of objects. In the previous studies, the movement direction of objects requires the process of segmentation and object detection in advance, so it requires considerable computing time, especially if applied to the crowded video. In this research, we developed a model for determining the movement direction of objects in the crowded video without involving segmentation and object detection process, using optical flow features that are derived to HOOF, dividing the frame into a certain size of grids, and the HOOF accumulation analysis for multiple frames. This research produces a model of determining the movement direction of objects on video that faster than previous methods.

TABLE I. STUDIES RELATED TO THE MOVEMENT DIRECTION OF AN OBJECT ON VIDEO

#	Paper	Feature/ Method	Direction	Description
1	[14], [15]	Optical flow, von Mises distribution, and K-means	Circular directions	The extraction of motion pattern used to recognize object behavior on video.
2	[16]	Optical flow, HOOF	8 directions	The calculation of object direction is done on the silhouette of the object; the object is divided into 8 region
3	[17]	Optical flow, HOOF	12 directions	Histogram of direction used for segmentation of the crowded object on video.
4	[4]	Optical flow, HOOF	32 directions	Movement direction used as a descriptor to classify the movement pattern.
5	[18]	Optical flow and Mixture of Gaussian (MoG)	Based on optical flow directions	Video input only 3 fps have not been tested on higher video density.
6	[19]	Edge and KLT tracker	Based on optical flow directions	Objects are segmented based on the edge features and the KLT method, the movement direction is determined based on optical flow features.
7	[20]	Optical flow, Kalman filter	8 directions	Object directions were detected using the proposed method.
8	[12], [21], [22]	Optical flow, Simple Nearest Neighbor Search	4 directions	HOFM descriptor was implemented to detect abnormal events on a video.
9	[23]	Optical flow, Grid-based HOOF	12 directions	HOOF features were extracted locally on a grid and used to determine the movement direction of the object on every grid.
10	[24]	Optical flow, SLAM	-	The video shot motion analysis based on LK optical flow method to solve the problem of film video shot motion.

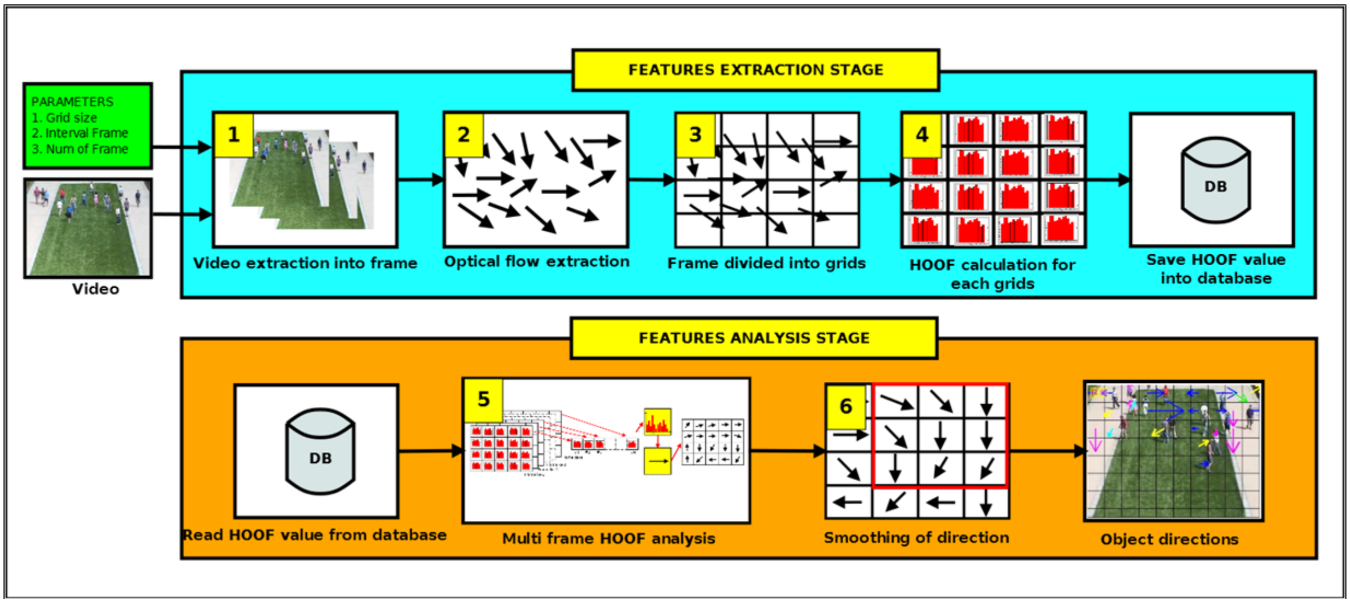


Fig. 1. The proposed model.

II. Proposed Method

This study aims to develop a model to determine the movement direction of objects in a crowd video using optical flow features that have computation time faster than previous methods while maintaining its accuracy value. To produce a model with faster computation time, then in this study:

- Derive the optical flow features into the Histogram of Oriented Optical Flow (HOOF), so it will reduce the dimensions of the features. To reduce it, we used simple statistical methods that are expected not to require a large computation time.

- Eliminate the process of segmentation and object detection because the process requires large computational time. The video constraint used in this study is a crowd video consisting of many human objects, so the process of segmentation and object detection is not effective to implement.

- Divide the frame into grids. It will keep a good accuracy value, and also accelerates the computing process because the accumulation of HOOF values is done only on the grid.

- Determination of the movement direction is not only done by analyzing HOOF on a single frame only but also done on multiple frames at once by accumulating the value of HOOF. It aims to maintain a good accuracy value.

We develop a video analysis model to determine the movement direction of objects consisting of two main processes: feature extraction and feature analysis. Fig. 1 presents the complete proposed model. This model is the main contribution of this research. Generally, the model has video input with 320x240 size, frame rate 25-30 fps and AVI (audio video interleave) format. The model also accepts several parameters that affect the process at each stage. The main parameters of the proposed model are as follows:

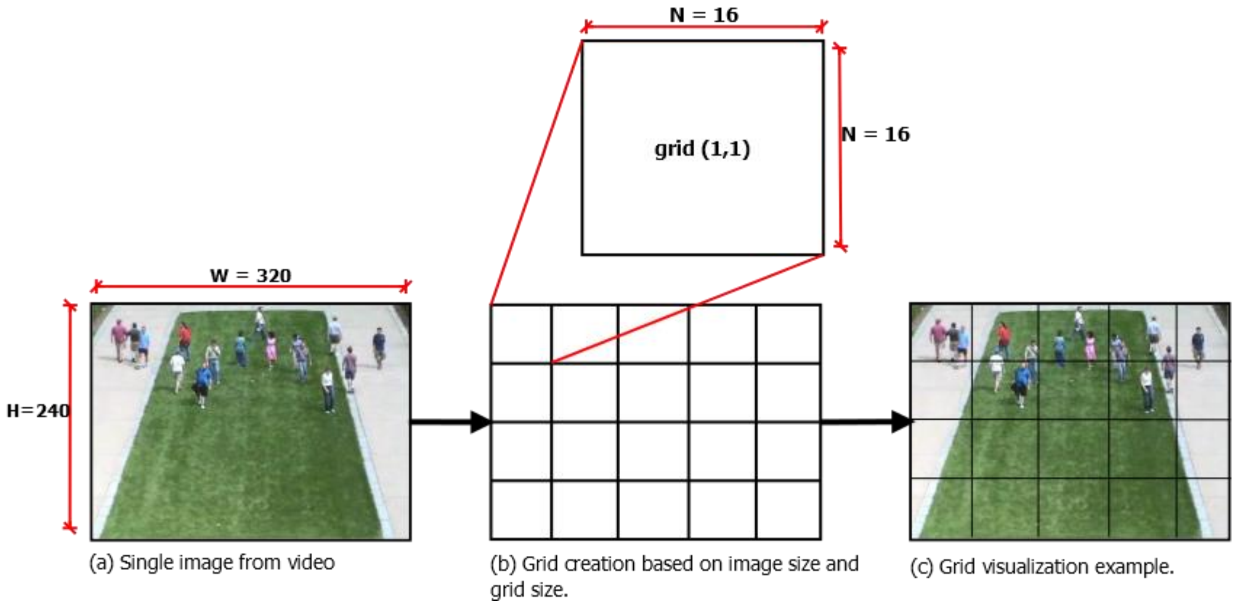
- The size of the grid, denoted by N. The value of N determines the size of the grid that is used to divide the area of a frame in the HOOF feature extraction process.

- Interval frame, denoted with Iframe. Value Iframe affect HOOF features analysis process that has been stored in the database. The value of 1 means interval frame analysis performed for each frame and the interval frame of 2 means that the next frame will be ignored.

- The number of frames (P). It determines the number of frames that are included in the analysis process.

The feature extraction process begins by dividing the video into frames and extracting optical flow features for each frame. The process of feature extraction is a process to take the necessary features of an object. In this study, optical flow values were obtained using the classic Horn-Schunck [25] method by applying the convolution kernel [26], [27]. In this research, the optical flow features extraction process is performed for each frame of the input video. The optical flow value of each frame is stored in a variable for processing at a later stage.

The optical flow value reflects the movement of each point of the video frame. The greater value of the optical flow (r), the more significant the movement occurs. Under these conditions, then in this study applied a threshold value to eliminate the less significant movement of each video frame. Based on the results of program testing, in this study used the threshold value of 0.1. If the optical flow value (r) is less than the threshold, then the optical flow value is changed to 0.

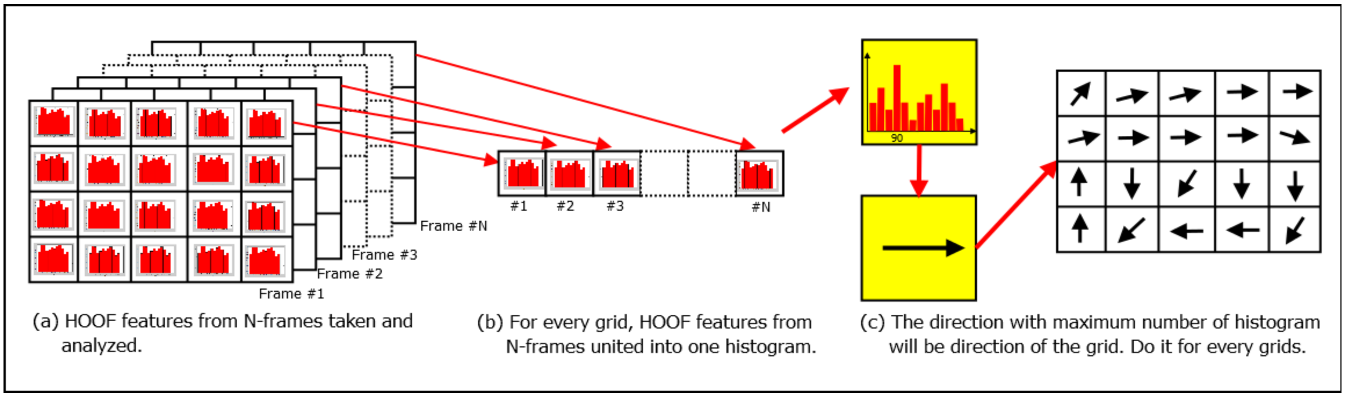


(a) Single image from video

(b) Grid creation based on image size and grid size.

(c) Grid visualization example.

Fig. 2. The frame is divided into a number of grids.



(a) HOOF features from N-frames taken and analyzed.

(b) For every grid, HOOF features from N-frames united into one histogram.

(c) The direction with maximum number of histogram will be direction of the grid. Do it for every grids.

Fig. 3. Multiple frames of HOOF feature analysis process for determining the movement direction.

The frame is divided into a number of grids with a certain size (see Fig. 2). The division of frames into grids is based on model parameters as described in the previous sections. Grid size determination is highly dependent on the video dataset used. The grid size that is too large or small can cause movement direction information to be lost or distorted by another movement. In this study, we tested several grid sizes, namely 8x8, 16x16, and 32x32 pixels. With an N x N grid on an image of width W and height H, a grid of m x n will be obtained. The value of m is obtained by dividing the width of the image (W) by the grid size (N). The value of n is obtained by dividing the image height (H) with the grid size (N). If the division yields a fractional value, then the result is rounded up. The box coordinates on Grid (m, n) are (x, y, w, h). The value of x is obtained by equation $x = \big((m-1)*N\big)+1$ and y with equation $y = \big((n-1)*N\big)+1$. For w and h values are grid size (N).

Optical flow extraction produces values in complex numbers. Optical flow value consists of two components, the horizontal (u) and vertical (v). The direction of the optical flow (denoted by θ) at a point (x, y) can be calculated based on the values u and v using (1). The value of θ is an angle between 0 and 360 degrees.

$$\theta_{x,y}=tan^{-1}\frac{v_{x,y}}{u_{x,y}} \qquad (1)$$

After the optical flow direction of each point is obtained, then it is normalized into 12 directions of movement. The normalized direction is denoted by θ_b and has a possible value of θ_b = {0, 30, 60, 90, 120, 150, 180, 210, 240, 270, 300, 330}. The calculation of optical flow normalization at point (x, y) is done with (2). The twelve directions of movement are codified into numbers 1, 2, 3, ..., 12. We also add 13 as a code meaning no movement direction. The absence of a movement direction needs to be codified as it allows at a point no movement at all.

To calculate the accumulated HOOF values for each grid, we use the equation $hoof(b)=\sum\theta_b(x,y); (x,y)\in W$ with hoof (b) is HOOF value of direction code b, which b ϵ {1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13}, θ_b (x,y) is an angle of b at point (x, y) such that point (x, y) is all point that located at area W. The process of histogram calculation of the optical flow direction is done by summing all the points that have a certain optical flow angle direction. The HOOF value calculation process is performed for each grid. Thus the HOOF value on the grid (m, n) is the histogram obtained based on the optical flow direction of the grid (m, n) only.

$$\theta_b = \begin{cases} 0 & if \quad 0\leq\theta_{x,y}<15 \text{ or } 345\leq\theta_{x,y}<360 \\ 30 & if \quad 15\leq\theta_{x,y}<45 \\ 60 & if \quad 45\leq\theta_{x,y}<75 \\ 90 & if \quad 75\leq\theta_{x,y}<105 \\ 120 & if \quad 105\leq\theta_{x,y}<135 \\ 150 & if \quad 135\leq\theta_{x,y}<165 \\ 180 & if \quad 165\leq\theta_{x,y}<195 \\ 210 & if \quad 195\leq\theta_{x,y}<225 \\ 240 & if \quad 225\leq\theta_{x,y}<255 \\ 270 & if \quad 255\leq\theta_{x,y}<285 \\ 300 & if \quad 285\leq\theta_{x,y}<315 \\ 330 & if \quad 315\leq\theta_{x,y}<345 \end{cases} \qquad (2)$$

The next process is analyzing the HOOF features to get the movement direction. It performed against a HOOF features database that coming up from the feature extraction process at an earlier stage. The analysis conducted on several frames at a time on the same grid position as shown in Fig. 3 Step (a) and (b). The number of frames being analyzed is determined based on the parameters of the model (denoted by P). For example, if you want to analyze the movement of the grid (1,1), the entire HOOF value of the grid (1,1) for each frame that is observed is taken and analyzed. HOOF value of the entire grid coupled into a single histogram. After the combined histogram is obtained, then the direction of movement is determined based on the highest histogram values. The process combines the histogram is done for each grid such that the entire grid of a series of frames can be determined the movement direction of the object as shown in Fig. 3 Step (c).

We also applied thresholding process to determine a grid will be analyzed its HOOF value or not. For grids that contain movement, it will be processed, while that does not contain movement does not need to be processed. Thus the computation process becomes faster. Based on the results of experiments that have been done, the best method is the max-min method. The threshold value (T) is obtained by adding the largest HOOF value with the smallest HOOF value, then divided by value 2.

The movement direction in a grid is obtained by accumulating the entire HOOF value of the entire grid analyzed. The HOOF grid values of the selected frames will be accumulated and recalculated to obtain the HOOF grid values that are accumulated from multiple frames at once. In other words, the analysis is not only done for a single frame but done for many frames (multi-frame). The process of accumulating HOOF values between grids is done using (3).

$$hoof_b(m,n) = \sum_{frame=1}^{P} hoof_b(m,n) \qquad (3)$$

with hoof$_b$ (m, n) is the histogram of b of some frames on the grid (m, n). P notation is the number of frames analyzed.

III. Results and Discussions

In this section, we explained the experiment results and discussed them.

A. The Experiment Scenario

The experiment on the whole system produced aims to determine the ability of the system of movement direction analysis on the video. It was performed using UMN public dataset [28]. In general, a series of tests are performed to find out:

- The value of success rate (SR) of the proposed model in order to determine the best parameter. The success rate of the detection process of a tested frame is calculated based on the number of grids detected correctly compared to the number of all grids detected by the system. Furthermore, the success rate of the overall tested data is averaged so as to produce the success value of the whole model. The evaluation of the proposed model is validated visually by the expert.
- The values of the accuracy, precision, and recall of the movement direction model using the best parameter. The experiment aims to determine the performance of the proposed model. Testing of accuracy, precision and recall value is performed using the best parameters based on test results. In addition to the accuracy, precision and recall value, it is proposed a measure of the performance of the classification model of the movement direction of the object on the video. It calls the direction detection error rate (DDER). The value of DDER indicates the error rate of the system classifies the movement direction. The error rate is based on how far the direction is detected by the system with the actual direction. For example, if the real direction of the object movement is the 3 o'clock direction, while the system detects it as the clock direction 9, then it has the highest error rate because the direction is opposite. While if the clock direction 3, classified as the direction of the clock 3 also then the error rate is 0 (smallest).
- The comparison of the speed of the HOOF features extraction process from the proposed model using a method that uses segmentation and object detection.

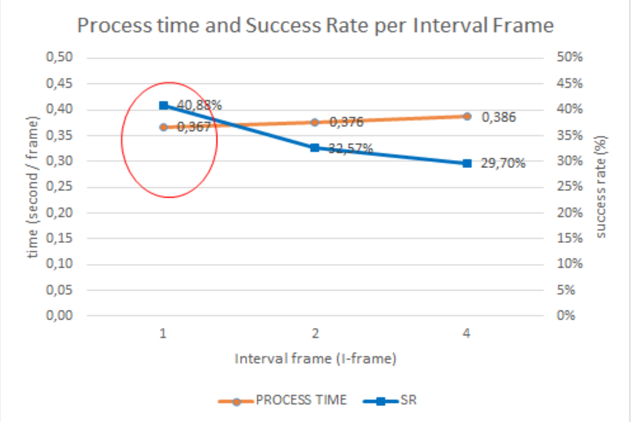


Fig. 4. Process time and success rate (SR) per interval frame.

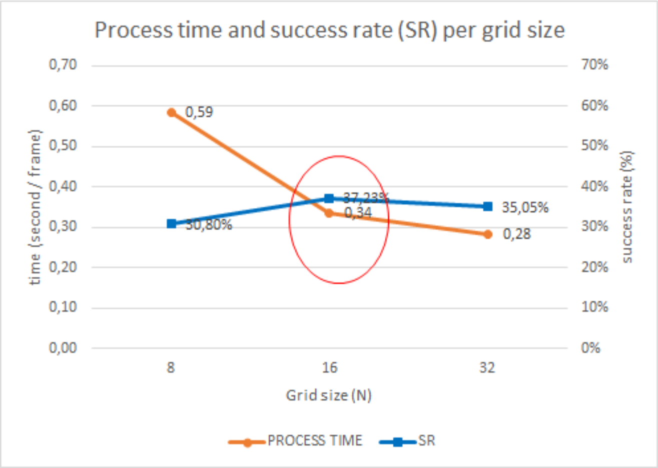


Fig. 5. Process time and success rate (SR) per grid size.

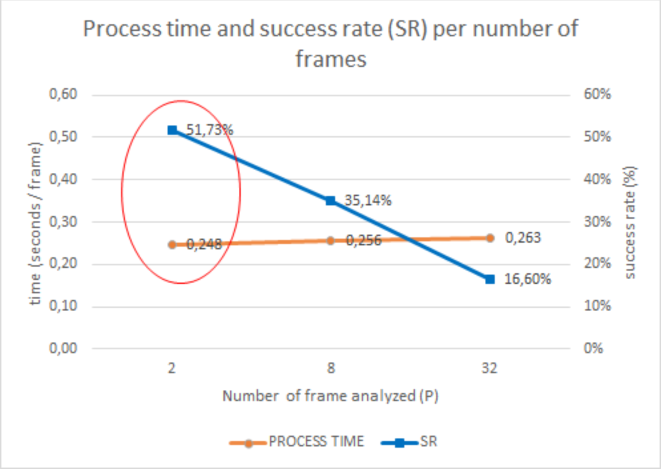


Fig. 6. Process tima and success rate (SR) per number of frames.

B. *The Testing of Success Rate (SR) of the Movement Direction Detection*

The success rate test of the proposed model is performed for each combination of the grid size (N), frame interval (I_{frame}) and the number of frames analyzed (P). It turns out that the resulting SR value has a very wide variation. This happens because the value of SR depends on the value of the parameters used. The diversity of the success rate generated can be seen statistically by calculating the standard deviation value, the mean and the smallest and largest value. The standard deviation from the test result of 23.67% indicates that the data distribution is very spread and varied. The smallest and largest value range is also very wide, i.e., between 0% and 100%.

Based on the test results can also be seen that the average success rate of the proposed model in detecting the movement direction of objects in the video is 34.36%. The result is still below 50% and still needs to be improved in the future. Based on the test results, it can also be concluded that the best success rate is achieved on the test parameter with grid size 16, frame interval 1, and the number of frames 2. To strengthen the conclusion, the best parameter analysis based on two measures is the success rate (SR) and the process speed for each test parameter separately.

Fig. 4 presents the graph of process speed and success rate for each grid size (N). From the graph, it is seen that the best success rate value is achieved on grid size 16, while for the process speed is achieved by the grid size 32. As explained in the previous section that the grid size greatly affects the success rate of the process of determining the movement direction of the object and the optimal value depends on the size of the object in the video. While the speed of the process will be smaller as the size of the grid is used. From the graph, the optimal cutting point is reached at grid size 16, so the best grid size parameter is 16x16.

Meanwhile, the process speed and the success rate of movement direction detection for each frame interval is shown in Fig. 5. Based on the graph, it is seen that the process time increases the value of the larger frame interval. The best process speed is achieved at the frame interval parameter of 1, with a velocity of 0.367 seconds / frame. In line with that, when viewed from the value of success rate, then the greatest success rate value is also achieved with the parameter of the frame interval of 1. The best success rate value is 40.88%. Thus it can be concluded that the best frame interval parameters recommended for use are 1. Next in Fig. 6 presents the process time and success rate for each parameter of the number of frames analyzed (P value). Based on the graph, it is clear that the larger the number of frames that are analyzed then from the side of the process time greater and from the side of success rate decreases. Therefore, it can be concluded that the optimal frame number parameter is worth 2 frames.

C. *The Testing of the Accuracy, Precision, and Recall*

The testing of accuracy, precision, and recall of the proposed model is done using the best test parameters are grid size 16, frame interval 1, and a number of frames analyzed 2. The accuracy, precision and recall value is calculated by confusion matrix method. The movement direction for each grid is classified into 13 class labels consisting of sets {0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12}. The class 0 label indicates that there is no movement on a grid.

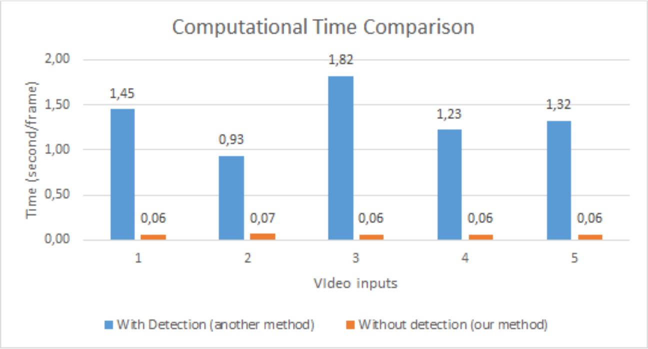


Fig. 7. Computational time comparison of the optical flow and HOOF extraction process.

Based on the test results, the accuracy of the model is quite high. The average accuracy for the entire class is 98.1%. It shows that the system is able to classify the movement direction correctly. Thus the system performance is in a good category. The results appear slightly different at the precision value. The precision value describes the number of correctly classified directions divided by the total grid having direction.

The average precision value of the whole class is only 35.6%. This indicates that the precision of the proposed model still needs improvement. Similarly, the value of recall that shows the level of system effectiveness in recognizing each class label is still quite low, that is equal to 41.2%.

In addition to the accuracy, precision, and recall of the proposed model, this study also calculated the DDER (direction detection error rate). The DDER value indicates the level of movement detection error from the proposed model. The higher the DDER value, the greater the error detection rate and the lower the proposed model's performance. Based on model performance test results, then we calculated DDER value. The resulting DDER value is 25.28%. Based on these values can be concluded that when the system detects the movement direction, the error rate in determining the direction of movement is 25.28%.

D. The Proposed Model Speed Comparison

In this research, we proposed a method of determining the movement direction of objects in a crowd video that does not involve the process of segmentation and object detection first, using the optical flow features that are derived into HOOF, dividing the frame into a number of grids, and HOOF accumulation analysis for multiple frames at once. The main purpose of this research is to produce a method or model of movement direction detection that has faster computation time than the previous method while maintaining its accuracy value.

To know the performance of the resulting model, especially in terms of time compared to other methods that require the process of segmentation and object detection, testing is done by comparing the proposed model with the previous method. Nevertheless, researchers have difficulty in finding similar methods that have the same research objective in determining the movement direction of objects. Therefore, in this study, we only compare computational time for the process of extracting HOOF features. The computation time of the process of determining the direction of movement of the object is not comparable because of the variety of techniques used by other researchers with unequal output results.

Fig. 7 presents computational time comparisons in the extraction process of optical flow and HOOF features. The comparison results show that for the computational time of optical flow features, eliminating the segmentation and object detection process saves computation time of 22 times more efficient. Meanwhile, for the computational time of feature extraction HOOF method without segmentation is slower than the method with segmentation, which is about 42% slower. It can be understood because by doing the detection and segmentation of the object first, the number of areas that the HOOF value is calculated will be less or limited only in the segmented area only.

IV. Conclusion

Based on the results of testing and discussion in the previous section some conclusions can be drawn as follows:

1) The video analysis model for determining the movement direction of objects on a video based on optical flow features that derived into Histogram of Oriented Optical Flow (HOOF) proved to be used to detect the movement direction of objects in the video.

2) The test also shows that the success rate of the model is influenced by three main parameters: grid size, frame interval and a number of frames analyzed. The success rate of the model is better on testing with 16x16 grid size, frame interval of 1, and a number of frames analyzed by 2 frames.

3) The testing of the model using the best parameter resulting in an average accuracy of 98.1%, 35.6% precision, 41.2% recall and a direction detection error rate of 25.28%.

4) The proposed model of this study has computation time of HOOF feature extraction which is faster than the similar method that requires segmentation and object detection first, with an average computation time of optical flow extraction of 0.06 seconds / frame and HOOF extraction of 0,29 seconds / frame. The results are 22 times faster than another method that implements segmentation and detection first.

In the future, our approach that proposed in this study will be developed with better optical flow estimation methods such as Deep Neural Network method [29], promising better computational time. Our method also needs to be implemented with another video dataset.

Acknowledgment

We would like to thank Budi Luhur University and Gadjah Mada University for supporting this research. This research was also supported by the Ministry of Research, Technology and Higher Education, Indonesia through contract number 0426/K3/KM/2017.

References

[1] BNPB, “Data dan Informasi Bencana Indonesia,” Badan Nasional Penanggulangan Bencana, 2015. [Online]. Available: http://dibi.bnpb.go.id. [Accessed: 07-Feb-2018].

[2] M. Mustofa, “Memahami kerusuhan sosial, suatu kendala menuju masyarakat madani,” J. Kriminologi Indones., vol. 1, no. I, pp. 10–19, 2000.

[3] K. Lang and G. Lang, “Collective Behavior,” International Encyclopaedia of the Social Sciences. Crowell Collier and Macmillan, Inc, 1968.

[4] F. Martínez, A. Manzanera, and E. Romero, “A Motion Descriptor Based on Statistics of Optical Flow Orientations for Action Classification in Video-Surveillance,” Commun. Comput. Inf. Sci., vol. 346, no. 2, pp. 267–274, 2012.

[5] H. M. Dee and A. Caplier, “Crowd Behaviour Analysis Using Histograms of Motion Direction,” in 2010 IEEE 17th International Conference on Image Processing, 2010, pp. 1545–1548.

[6] S. Fazli and H. Fathi, “Video Image Sequence Super Resolution using Optical Flow Motion Estimation,” Int. J. Adv. Stud. Comput. Sci. Eng., vol. 4, no. 8, pp. 22–26, 2015.

[7] Y. Benabbas, S. Amir, A. Lablack, and C. Djeraba, “Human Action Recognition Using Direction and Magnitude Models of Motion,” in International Conference on Computer Vision Theory and Applications (VISAPP) 2011, 2011, pp. 277–285.

[8] B. D. Lucas and T. Kanade, “An Iterative Image Registration Technique with an Application to Stereo Vision,” in Proc. 7th International Conference on Artificial Intelligence (IJCAI) 1981, 1981, pp. 121–130.

[9] C. Tomasi, “Detection and Tracking of Point Features,” 1991.

[10] N. Ihaddadene and C. Djeraba, “Real-time crowd motion analysis,” in 2008 19th International Conference on Pattern Recognition, 2008, pp. 1–4.

[11] Y. Benabbas, “Human behavior analysis from videos using optical flow,” no. 33. UNIVERSITE DES SCIENCES ET TECHNOLGIES DE LILLE, pp. 1–49, 2011.

[12] R. V. H. M. Colque, C. A. C. Junior, and W. R. Schwartz, “Histograms of Optical Flow Orientation and Magnitude to Detect Anomalous Events in Videos,” in Conference on Graphics, Patterns and Images (SIBGRAPI 2015), 2015, pp. 1–9.

[13] T. Li, H. Chang, M. Wang, B. Ni, and R. Hong, “Crowded Scene Analysis : A Survey,” IEEE Trans. CIRCUITS Syst. VIDEO Technol., vol. 25, no. 3, pp. 367–386, 2015.

[14] M. Singh, a. Basu, and M. K. Mandal, “Human Activity Recognition Based on Silhouette Directionality,” IEEE Trans. Circuits Syst. Video Technol., vol. 18, no. 9, pp. 1280–1292, Sep. 2008.

[15] Y. Benabbas, N. Ihaddadene, and C. Djeraba, “Motion pattern extraction and event detection for automatic visual surveillance,” EURASIP J. Image Video Process., vol. 2011, pp. 1–15, 2011.

[16] K. Lertniphonphan, S. Aramvith, and T. H. Chalidabhongse, “Human action recognition using direction histograms of optical flow,” 2011 11th International Symposium on Communications & Information Technologies (ISCIT). Ieee, pp. 574–579, Oct-2011.

[17] S. Cui, N. Li, and Z. Liu, “Multi-directional crowded objects segmentation based on optical flow histogram,” in Proceedings - 4th International Congress on Image and Signal Processing, CISP 2011, 2011, vol. 1, pp. 552–555.

[18] O. Smirg, Z. Smekal, M. K. Dutta, and B. Kakani, “Automatic Detection of the Direction and Speed of Moving Objects in the Video,” in Sixth International Conference on Contemporary Computing (IC3), 2013, pp. 86–90.

[19] S. D. Khan, “Estimating speeds and directions of pedestrians in real-time videos: A solution to road-safety problem,” in CEUR Workshop Proceedings, 2013, vol. 1122, pp. 1–14.

[20] A. S. Rao, J. Gubbi, S. Marusic, A. Maher, and M. Palaniswami, “Determination of object directions using optical flow for crowd monitoring,” Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 8034 LNCS, no. PART 2, pp. 613–622, 2013.

[21] R. V. H. M. Colque, C. Caetano, M. T. L. de Andrade, and W. R. Schwartz, “Histograms of Optical Flow Orientation and Magnitude and Entropy to Detect Anomalous Events in Videos,” IEEE Trans. Circuits Syst. Video Technol., vol. 27, no. 3, pp. 673–682, Mar. 2017.

[22] C. A. Caetano, V. H. C. D. Melo, J. A. dos Santos, and W. R. Schwartz, “Activity Recognition Based on a Magnitude-Orientation Stream Network,” in 2017 30th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), 2017, pp. 47–54.

[23] A. Solichin, A. Harjoko, and A. E. Putra, “Grid-based Histogram of Oriented Optical Flow for Analyzing Movements on Video Data,” in 2015 International Conference on Data and Software Engineering, 2015, pp. 114–119.

[24] P. Wang, C. Chen, C. Dong, H. Xu, and F. Tian, “The analysis method of video camera’s motion based on optical flow and slam,” in 2016 International Conference on Audio, Language and Image Processing (ICALIP), 2016, pp. 62–66.

[25] B. K. P. Horn and B. G. Schunck, “Determining optical flow,” Artif. Intell., vol. 17, no. 1–3, pp. 185–203, Aug. 1981.

[26] J. L. Barron, D. J. Fleet, S. S. Beauchemin, and T. A. Burkitt, “Performance of Optical Flow Techniques,” in 1992 IEEE Computer Society Conference on Computer Vision and Pattern Recognition., 1992, pp. 236–242.

[27] S. S. S. Beauchemin and J. L. L. Barron, “The Computation of Optical Flow,” ACM Comput. Surv., vol. 27, no. 3, pp. 433–467, 1995.

[28] UMN, “Unusual Crowd Activity Dataset of University of Minnesota, Department of Computer Science and Engineering,” UMN, 2006. [Online]. Available: http://mha.cs.umn.edu/movies/crowd- activity-all.avi. [Accessed: 10-Aug-2017].

[29] X. Zhang, D. Ma, X. Ouyang, S. Jiang, L. Gan, and G. Agam, “Layered Optical Flow Estimation Using a Deep Neural Network with a Soft Mask,” IJCAI, vol. May, no. 2018, 2018.

Secure user Authentication and File Transfer in Wireless Sensor Network using Improved AES Algorithm

Ishu Gupta
Research Scholar
Punjab Technical University

Dr Harsh SadaWarti
Professor
C.T. University Jalandhar

Dr S.N Panda
Professor
Chitkara University,Rajpura

Jatin Gupta
Assisant Professor
Chitkara University,Rajpura

***Abstract*—The WSN technology is a highly efficient and effective way of gathering highly sensitive information and is often deployed in mission-critical applications, which makes the security of its data transmission of vital significance. However, the previous research paper failed to distinguish the role of centralized server for it being the main controller of the entire network. The decision of nodes communicating with each other in the previous research paper was based on the information received from the adjacent node. However, the proposed research paper will take into account the centralized server to develop a new technique to prevent the black node from joining the wireless sensor network. Key distribution technique along with the implementation of improved AES algorithm double key encryption will play an important role in transferring the data between authorized nodes securely and preventing unauthorized user from accessing it.**

Keywords—Wireless sensor networks (WSN); centralized server; black node; encryption; security; key distribution technique

I. Introduction

The wireless sensor networks (WSN) are specialized transducers with spatially dispersed and dedicated autonomous sensor nodes for identifying, monitoring and recording the physical and environmental conditions at different locations. Of the most commonly monitored physical parameters include temperature, pressure, light, direction of wind and its speed, intensity of illumination, vibration and sound, chemical concentration in water, air, pollutant level, humidity, vital body functions and so on.

WSN is a revolutionary technology that comprises of several sensor nodes that are small in size, light in weigh and easily portable. These sensor nodes are laced with a radio transceiver, a microcontroller and a battery, which can either be embedded in it or located externally as an energy resource. The function of the radio transceiver is to connect the sensor nodes or neighbor nodes with an external link while the microcontroller is an electronic circuit that plays a significant role to interface the sensor nodes thereby forming a complete circuit to effectively process, store, receive and send data to the base station.

This hi-end information gathering technology was originated as an initiative of keeping surveillance checks in the wars or battlefields. With its great potential applications, today, it is widely and effectively deployed in wide-mission vital military operations, various industries monitoring the health of machineries, agriculture and commercial domains for monitoring and controlling various other applications. As the WSN deals in highly sensitive information, its processing, gathering and transmittance, therefore, security in this spatially dispersed nodal network is of crucial concern. This kind of highly sensitive data, which can be related to a patient for its medical background data, military operations related strategies or highly confidential mission data, data related to earthquake, or other such environment calamity and much more must be dispersed or transmitted in an encrypted format. As any leakage or misuse of this critical information or data can create serious issues and impact an individual or the entire nation badly, thus, it becomes of paramount importance to secure the data of the sensory nodal network by deploying effective and efficient network security techniques.

Wireless sensor networks are one of the most intriguing yet most challenging technologies of the current times due to its built-in complexity. The sensor nodes of WSN work under extreme resource constraints as the energy resource usually comprises of an embedded device with limited supply to transmit data in a highly unspecified environment. Being a wireless mode of network, the chances of data packets getting damaged due to an unwanted error, or conflict amongst the nodes or over congestion is very high. As the entire security mechanism of the network depends upon the cryptographic key distribution and reporting of critical events, the unreliable mode of communication amongst the sensor nodes poses grave threat to the security of the network. Additionally, sending and receiving secure data in highly callous conditions is not an easy task as the sensor nodes have the tendency to closely interact with their physical environment to process and blend data and create novel information that must be transmitted to the end-station. However, these uncontrolled operations in unattended environment may create accidental node failure.

The security of Wireless Sensor Network (WSN) is under grave threat due to the attacks on the sensor nodes, which are often categorized as goal-oriented attacks, performer-oriented attacks and layer-oriented attacks. Goal-oriented attacks are mainly against the data confidentiality wherein an attacker passively monitors the traffic, analyzes it for imperceptibly encrypted sensitive information and then gains authentication information to pass through the network. This type of attack is called passive attack which results in the revelation of sensitive information to the attacker without any knowledge at the user's part. However, in the active attack, the attacker actively assesses the entire network to gain control over it. The best and most common ways of active attack includes data modification, spoofing, sinkhole, flooding, jamming the network, warm hole, black hole, fabrication, lack of co-ordination, node subversion, false nodes, selective forwarding and so on. While in performer-oriented attacks, the attacks are either internal or external. Internal attackers are the trickiest ones as they are not only the legitimate node of the original network but also have direct access to all the sensitive network information. The internal attacks include modification, misrouting, eavesdropping and packet dropping attacks that leads to suppression of critical information reaching the base station, thereby degrading the network performance. On the other hand, external attackers are known for eavesdropping on transmittance of data along with injecting fake data in the network to exhaust energy resource, which will lead to denial of services. Another attack on WSN are related to its layered architecture, which makes it susceptible to node capturing, jamming of radio signals, violating redefined protocols, inducing collisions by disrupting a packet, depletion of energy due to recurring retransmission and new connection request to avert the sensor nodes from communicating effectively.

As of now, trust management system is considered to be the most effective and efficient way of dealing with the attacks on the sensor nodes of a wireless network. Trust factor is a very important and useful concept for WSN for detecting the attacks on the sensor nodes and accordingly support in the process of decision making. The concept has grown its relevance with the rising use of internet transaction and e-commerce. Considering trust as an important parameter in the relationship between two nodes, it becomes quite easy to identify the innate qualms in their co-operation process.

This concept of trust has been originated from the sociological and psychological environment, which makes it an essential element in any kind of network, be it social or computer related. Generally, a trust management system is broadly classified into two types: credential based trust management system and behavior based trust management system.

II. Literature Survey

A lot of researchers have worked in this field to provide fool proof algorithms to avoid the security breaches of the wireless sensor network. The unwanted and unnecessary challenges associated with the sensory nodes of wireless network makes it intricate to employ the proposed security approaches of the researchers in the past. However, careful and diligent understanding of these challenges along with the susceptible attacks on the sensor nodes can definitely aid the researchers in proposing or presenting an algorithm that would work efficiently and effectively in handling the security breaches and enhancing the security in the wireless sensor network.

Geetha D. Devanagavi et al. [1] in their research proposed an agent based Secured Routing using Trusted neighbors (ASERT) in WSN. The proposed technique ensures high security by selecting the trustworthy neighbors and formulating the secured routes in the network using probability based trust model and MAC model. In this task, software agents play a pivotal role. The entire process of identifying trusted neighbors is divided into two phases: the first phase involves agents visiting all the neighbors one by one and determining their probability using trust model and in the second phase, MAC model is used to ensure the trusted neighbors.

Monia, Sukhchandan Randhawa and Sushma Jain [2] in their research study proposed an improved algorithm in which the cluster heads are chosen based on the received signal. Calculation of trust values and malicious node detection is done by considering the packet forwarding factor. The proposed technique also takes into account the consistency of clusters and lifetime of the network.

Mukesh Kumar and Kamlesh Dutta [3] put forth a literature survey to elaborately discuss concerns that can cause the security breach in data aggregation. It vividly describes the basics of aggregating the data in WSN in a secured manner, important factors to be taken into account for classifying the secured data aggregation techniques for WSN, key aspects of the existing data aggregation techniques and a crisp comparison of these techniques based on parameters like aggregation function, cryptographic techniques used in WSNs etc.

Vinod Kumar Verma, Surinder Singh, N. P. Pathak [4] intended to investigate the repercussions of static, dynamic and oscillating modes by performing prevalent analysis of wireless networks. The parameters like accuracy, path length and energy consumption are taken into account to examine the impact of different WSN modes on the deployed trust and reputation models: Bio-inspired trust and reputation, Eigen trust, peer trust, power trust and linguistic fuzzy trust and reputation.

Weidong Fang, Chuanlei Zhang, Zhidong Shi, Qing Zhao, and Lianhai Shan [5] proposed a beta based trust and reputation evaluation method which employs the beta distribution to determine the credibility of nodes distribution. The calculated trust values are utilized to choose the relay nodes and counteract the internal attacks. Intensive experimental results exhibit higher information security and maximize shield against various types of internal attacks from malicious nodes.

X. Anita, M. A. Bhagyaveni, J. Martin Leo Manickam [6] suggested a minimal overhead trust management scheme in terms of memory and energy consumption. Instead of deriving the trust values of the neighboring nodes haphazardly, it employs a novel trust detector that monitors and alarms the nodes whose trust falls below a minimum threshold. This warning motivates the sensor nodes to improve its trust relationship with other nodes by analyzing and rectifying its packet forwarding behavior.

Yun Liu, Chen-xu Liu, Qing-An Zeng [7] formulated an improved trust management system derived from the trust model in the iRTEDA protocol which is utilized to attain the secured data aggregation pertaining to the nature of relationship between the nodes in the network. The proposed trust model intends to efficiently utilize the second –hand information from the neighboring nodes and to attain the maximum level of security for aggregating the data and evaluating the trust and reputation of the nodes.

Yannis Stelios, Nikos Papayanoulas, Panagiotis Trakadas, Sotiris Maniatis, Helen C. Leligou, and Theodore Zahariadis [8] proposed a novel trust and reputation system which detects a wide range of security threats. The proposed model exhibits effective estimation of malicious nodes and sustains the network connectivity even when the malicious nodes comprise majority of the network. It also incorporates the energy awareness in the network.

III. Problem Statement

The existing research study has a big drawback in terms of not explaining the role of centralized server in checking the trust value of the nodes. Despite centralized server being the main controller of the entire network, the decision of nodes communicating with each other is based on the information received from the adjacent node. In such cases, presence of a black node in the network can cause severe damage to the network. Therefore, a black node can pose a big threat if the centralized network fails to detect its entry in the network, joining and connecting with other nodes as black node is known for sending wrong information to other nodes, thereby affecting the trust value of the target node as a whole.

With this drawback in mind, the main emphasis of the proposed research study is to check the trust factor of node from the centralized server i.e. the cluster head (CH) that controls the whole network to prevent the problem of black node. Additionally, we have set a number of rules to prevent the black node enter and disrupt the wireless Sensor Network and its functioning, which are discussed in the proposed work.

IV. Proposed Work

1) Registration Phase: In registration phase, server will give Big Integer unique key, an ID to user.

2) Login Phase: For fresh node, there will be two phases to join the network:

a) Authentication phase, wherein the authentication phase server will check the following parameters:

- mac Address,
- username,
- password, and
- unique key.

If all these parameters are matched, only then authentication will be completed for the server to send an encrypted message to the Client.

b) Authorization Phase: In this, a user will decrypt the message received by server by using his / her designated key and send it to the server for matching. If the decrypted message gets matched with the sending message, only then the authorization phase will be completed for the user to log in.

3) File Sharing: When user wants to share the file, firstly key agreement phase will be placed which will be done by using IBE algorithm.

Algorithm steps:

Firstly, user requests to the server for receiver's ID, public key and a number.

a) User will encrypt the message using receiver's ID, public key, and a number that he gets from the server.

b) Receiver will decrypt this challenge using its private key, number, user ID that he gets from the server.

If receiver decrypted message matches with the sender message, only then file sharing is possible between them.

A. Encryption

On the server, file will be encrypted, using RSA with homomorphic + AES algorithm;

Steps:

1) Apply Homomorphic with RSA algorithm: The homomorphic property is meant to preserve the multiplication.

C(x1) â‹…C(x2) = (xe1modm)â‹…(xe2modm)

2) Now using encrypted key, the AES algorithm firstly enables the server to select a master key for the particular user and apply RSA Homomorphic on it as and it returns digital signature which is used as key (1244812334565456)

Key (1244812334565456) in hexadecimal is→ 31 32 34 34 38 31 32 33 33 34 35 36 35 34 35 36

Now suppose, we have plain text to be encrypted as: "Wireless sensor"

Change it to Hexadecimal code: 57 69 72 65 6c 65 73 73 20 20 73 65 6e 73 6f 72

Perform X-OR operation on it with key as:

$$\begin{pmatrix} 31\ 38\ 33\ 35 \\ 32\ 31\ 34\ 34 \\ 34\ 32\ 35\ 35 \\ 34\ 33\ 36\ 36 \end{pmatrix} \text{ X-OR with } \begin{pmatrix} 57\ 6c\ 20\ 6e \\ 69\ 65\ 20\ 73 \\ 72\ 73\ 73\ 6f \\ 65\ 73\ 65\ 72 \end{pmatrix}$$

Result:

Suppose it returns as -

54 77 6F 20 4F 6E 65 20 4E 69 6E 65 20 54 77 6F

Pass this cipher text as plain text into AES as - Improved AES:

Key will be generated by main server.

1) Add Round key

In this process, X-OR operation is performed between round key and state. In Aes algorithm the total rounds are 10 but in our Improved Aes algorithm we have reduced 2 rounds to generate cipher text. Round key is generated from the cipher key by using the key expansion process.

AES example - The first round key:

- Round 0: 54 68 61 74 73 20 GD 79 20 4B 75 6E 67 20 46 75
- Round 1: E2 32 FC Fl 91 12 91 88 B1 59 E4 E6 D6 79 A2 93

2) Add Round key

- State Matrix

$$\begin{pmatrix} 54\ 4F\ 4E\ 20 \\ 77\ 6E\ 69\ 54 \\ 6F\ 65\ 6E\ 77 \\ 20\ 20\ 65\ GF \end{pmatrix}$$

Round 0 Matrix:

$$\begin{pmatrix} 54\ 73\ 20\ 67 \\ 68\ 20\ 4B\ 20 \\ 61\ 6D\ 75\ 46 \\ 74\ 79\ GE\ 75 \end{pmatrix}$$

Perform X-OR operation on state matrix and round 0 matrix.

New Matrix:

$$\begin{pmatrix} 00\ 3C\ GE\ 47 \\ 1F\ 4E\ 22\ 74 \\ OE\ 08\ 1B\ 31 \\ 54\ 59\ OB\ 1A \end{pmatrix}$$

Here, we will apply ADD ROUND KEY to this resulted matrix, using digital signature provided by RSA Homomorphic as:

Digital Signature key is→ “1234123412341234”;

This key is in hexadecimal is:

57 A1 50 A1 70 80 A0 60 F0 30 AB D1 FF F1 5E F2

$$\begin{pmatrix} 00\ 3C\ GE\ 47 \\ 1F\ 4E\ 22\ 74 \\ OE\ 08\ 1B\ 31 \\ 54\ 59\ OB\ 1A \end{pmatrix} \text{ X-OR with}$$

$$\begin{pmatrix} 57\ 70\ F0\ FF \\ A1\ 80\ 30\ F1 \\ 50\ A0\ AB\ 5E \\ A1\ 60\ D1\ F2 \end{pmatrix}$$

Suppose Results as:

$$\begin{pmatrix} 00\ 3C\ GE\ 47 \\ 1F\ 4E\ 22\ 74 \\ OE\ 08\ 1B\ 31 \\ 54\ 59\ OB\ 1A \end{pmatrix}$$

3) Substitution Bytes

- Current State Matrix:

$$\begin{pmatrix} 00\ 3C\ 6E\ 47 \\ 1F\ 4E\ 22\ 74 \\ OE\ 08\ 1B\ 31 \\ 54\ 59\ OB\ lA \end{pmatrix}$$

Firstly, we will change this matrix into 1*4 orders so that we can reduce substitution time as following:

COL 1 COL 2 COL 3 COL4

ROW1 →

	COL 2	COL 3	COL 4
00 1F 0E 54	3C 4E 08 59	6E 22 1B 0B	47 74 31 1A

In order to increase the throughput, 4 rows are merged into 1 row.

Substitute each entry of current state matrix column wise and find the entries into S-box as following:

a) As 00 1F 0E 54, firstly it will be ordered in the ascending manner 00 0E 1F 54

b) For Example, take 1F to find the entries into S-Box. In 1st row and Fth column, the next entry loop will start from 1 onward, suppose now next value is 54, so loop will start from 1 instead of searching from 0th location.

- Substitute each entry (byte) of current state matrix by corresponding entry in AES S-Box. This leads to new State Matrix:

$$\begin{pmatrix} 63\ EB\ 9F\ AO \\ CO\ 2F\ 93\ 92 \\ AB\ 30\ AF\ C7 \\ 20\ CB\ 2B\ A2 \end{pmatrix}$$

4) Shift Rows

It is an operation that is applied to each row of the matrix state where the first row remains unchanged and the second, third and fourth rows are cyclically shifted on (K-1) basis.

Here k represents the order of the row.

5) Mix Columns

During the mix columns process, each column of the state array is considered as a polynomial. Multiplication using a predefined Matrix is carried out and output is obtained.

The binary calculation based method is used in conventional mix column transformation. The mix columns transformation operates on the state column-by-column. The multiplication method used in mix column transformation.

6) Download File

For downloading a file, user sends request to the server and if user is authenticated, then the server send all the file data to the user.

B. Decryption

At the time of decrypting a file, the request goes to the server and if user is authenticated then server send the encrypted key to the user to decrypt the message by using the following algorithm steps:

1) First select the file to be decrypted.

2) Use the key for decrypting the file.

3) Select the cipher mode for decryption.

4) Now decrypt the final result with the help of AES algorithm.

5) Get the file content into Bytes.

6) Decode this bytes using Base 64 Decoder.

V. Results and Discussion

The energy consumption is represented in Table I and Fig. 1, where with higher nodes consumption is increasing.

TABLE I. Energy Consumption

No. of nodes	Existing	Proposed
10	.008	.006
50	3	2
100	5	4
150	6	7
200	7	8
500	12	15
1000	32	38

Formula: Energy consumption= Load * time

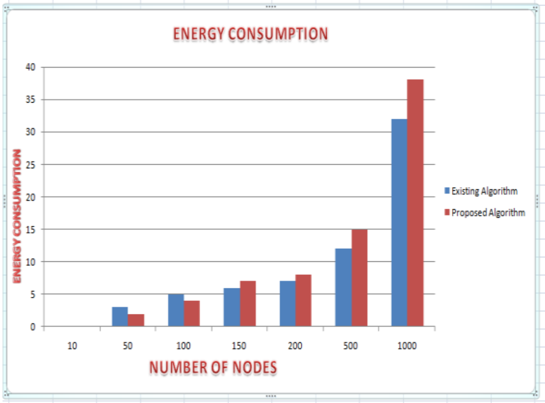


Fig. 1. Energy consumption.

TABLE II. Aggregating Energy

No. of nodes	Existing	Proposed
50	45	50
100	45.80	52
300	46	54.5
500	53	58.5
1000	60	65

Formula: Aggregation accuracy = (Average value of successful transaction/ Total no. of nodes)*100

The aggregating accuracy is represented in Table II and Fig. 2, where the aggregating accuracy for our proposed algorithm is quite better than existing.

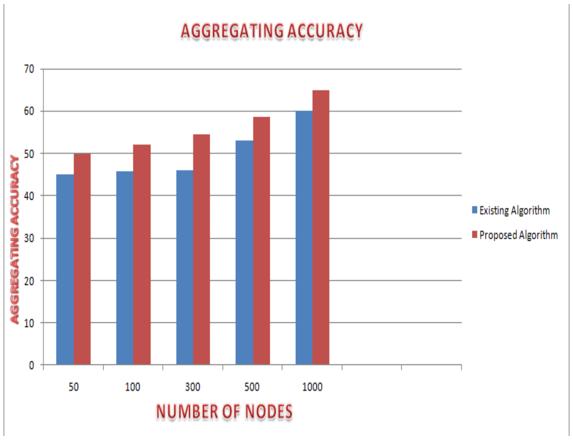


Fig. 2. Aggregating accuracy.

TABLE III. TRUST VALUE

No. of nodes	Existing	Proposed
50	0.5	0.5
100	0.68	0.70
300	0.70	0.72
500	0.75	0.78
1000	0.82	0.87

Formula: Trust value = No of successful transaction of active nodes / Total number of active nodes

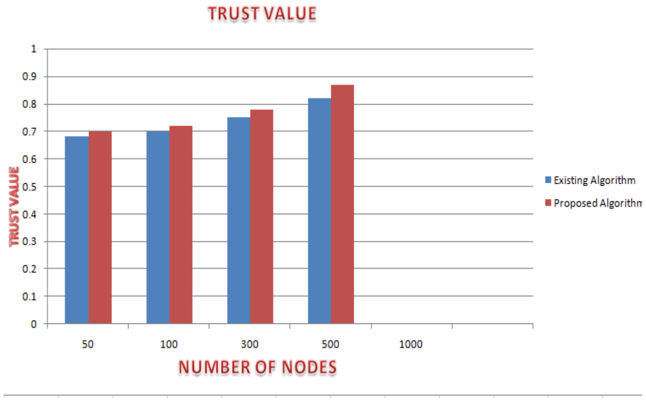


Fig. 3. Trust value of nodes.

The trust value of nodes is given in Table III and in Fig. 3, where the trust value is slightly better for higher number of nodes. The trust is one of the most important parameter of evaluating the routing performance.

TABLE IV. COMPROMISED VALUE

No. of nodes	Existing	Proposed
50	0.5	0.5
100	0.47	0.46
300	0.30	0.28
500	0.29	0.25
1000	0.18	0.15

Formula: Compromised value = 1- No. of successful transaction of every node

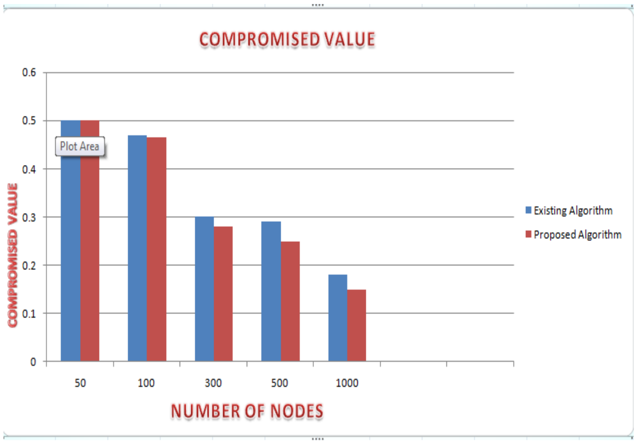


Fig. 4. Compromised value of nodes.

The value of compromised nodes is shown in Table IV and Fig. 4, the compromised nodes are lesser for our proposed algorithm.

VI. CONCLUSION AND FUTURE SCOPE

In our research work, we have developed a new technique to prevent the black node from joining the wireless sensor network. For this, we used key distribution technique along with the implementation of improved AES algorithm double key encryption to not only transfer the data between authorized nodes securely but also prevent unauthorized user from accessing it. Our results regarding trust value, energy consumption, and compromised nodes are enhanced as compared to the previous approaches. However, the only drawback with our research is that if by any chance a black node gets to enter the WSN, then it can easily receive the digital key from the main server and decrypt the data. Therefore, more research needs to be done in this particular area in the future.

REFERENCES

[1] Geetha D. Devanagavi, N. Nalini, Rajashekhar C. Biradar "Trusted Neighbors Based Secured Routing Scheme in Wireless Sensor Networks Using Agents". DOI 10.1007/s11277-014-1704-4. © Springer Science+Business Media New York 2014.

[2] Monia, Sukhchandan Randhawa and Sushma Jain "An Efficient Trust Management Algorithm in Wireless Sensor Network". © Springer Science+Business Media Singapore 2016. N.R. Shetty et al. (eds.), Emerging Research in Computing, Information, Communication and Applications, DOI 10.1007/978-981-10-0287-8_26.

[3] Mukesh Kumar and Kamlesh Dutta "A Survey of Security Concerns in Various Data Aggregation Techniques in Wireless Sensor Networks". Springer India 2015. L.C. Jain et al. (eds.), Intelligent Computing, Communication and Devices, Advances in Intelligent Systems and Computing 309, DOI 10.1007/978-81-322-2009-1_1.

[4] Vinod Kumar Verma, Surinder Singh, N. P. Pathak "Towards comparative evaluation of trust and reputation models over static, dynamic and oscillating wireless sensor networks". DOI 10.1007/s11276-015-1144-4. Springer Science+Business Media New York 2015.

[5] Weidong Fang, Chuanlei Zhang, Zhidong Shi, Qing Zhao, and Lianhai Shan "BTRES: Beta-based Trust and Reputation Evaluation System for wireless sensor networks". http://dx.doi.org/10.1016/j.jnca.2015.06.013. 1084-8045/& 2015 Published by Elsevier Ltd.

[6] X. Anita, M. A. Bhagyaveni, J. Martin Leo Manickam "Collaborative Lightweight Trust Management Scheme for Wireless Sensor Networks". DOI 10.1007/s11277-014-1998-2. © Springer Science+Business Media New York 2014.

[7] Yun Liu, Chen-xu Liu, Qing-An Zeng "Improved trust management based on the strength of ties for secure data aggregation in wireless sensor networks". DOI 10.1007/s11235-015-0078-6. © Springer Science+Business Media New York 2015.

[8] Yannis Stelios, Nikos Papayanoulas, Panagiotis Trakadas, Sotiris Maniatis, Helen C. Leligou, and Theodore Zahariadis "A Distributed Energy-Aware Trust Management System for Secure Routing in Wireless Sensor Networks". F. Granelli et al. (Eds.): MOBILIGHT 2009, LNICST 13, pp. 85–92, 2009. © ICST Institute for Computer Sciences, Social-Informatics and Telecommunication Engineering 2009.

Design of Android-Based Remote Patient Monitoring System

Salman ul Mouzam, Muhammad Daud, Salman Ali, Abdul Qadir Ansari
Department of Biomedical Engineering
Mehran University of Engineering and Technology, Jamshoro, Pakistan

***Abstract*—Efficient real-time monitoring systems for Patients with critical health condition have been always helpful for making timely decisions to save the lives. In such systems, the useful monitored factors include SPO2 (Oxygen Saturation in Blood), heart rate as well as temperature. Further, there are hundreds of patients in ICUs under monitoring systems in different hospitals and in different regions under fewer doctors/consultants on the move. Under above facts, a prototype for continuous monitoring of patient's health statistics such as SPO2 and temperature along with a bed-side desk using a PC/Laptop (bio instrumentation) working as Server Database with an application layer top transfer data on Android Application Server is successfully developed. This Android application accessing real-time monitored factors using Server Database allows the consultant to monitor patient's vitals data using his smart phone on move being at any hospital or city that creates easiness to handle any emergency and reduces Patient risks.**

Keywords—Monitoring system; SPO2; temperature; android application; bio instrumentation

I. Introduction

The Present research work on healthcare system aims to deliver better healthcare to patients anytime and anywhere in low cost and friendly manners. Therefore, for increasing the patient care efficiency, there is a need for continuous monitoring and consultation to avoid any life loss. Health care industry today faces some basic problems when it comes to patient monitoring. Like firstly, less number of Surgeons/Doctors, which is a reason that patient is unable to get timely attention and good treatment. Secondly, increased population, which also increases the number of patients and aged persons require health care. In order to achieve better quality patient care, there is a need for an effective monitoring system which can help the patient to seek utmost attention from doctors and similarly can help doctors to treat as many patients being at distance or without any distance.

Thanks to recent advances in Telemedicine and Bioinstrumentation, it is possible to receive, process, record and transmit patient's physiological vitals (Signals presenting actual vitals level) to computer servers and to any location via internet servers. This advancement has not only a blessing for patients at outreach to get a good healthcare efficiently but also advantageous for doctors who can treat many patients at the same time. Further, with the prevalent android OS based smartphones, it is motivating to make an android application for doctors and consultants to access real-time patient monitoring data on the mobile phone application.

Today, into eager every day life, such is challenging in conformity with preserve a wholesome lifestyle, that's why improper health cover/checkup leads to increase the number of persons getting sick for for long time. The regular health check can be achieved with monitoring systems [1]. Generally, a monitoring system for a patient is a procedure where a Doctor can constantly screen patient's vitals in a from distance location. Since Traditional healthcare technologies have been confined to hospitals providing no mobile healthcare (monitoring or consultancy) resulted with wastage of time, money and ease, several research teams have been working on this bioinstrumentation via remote monitoring using different methods [2].

In literature, it can be found that different possibilities for realization of such bio-instrumentation and remote monitoring like in [3], with ATMEGA8L microcontroller with sensor network – a healthcare monitoring is achieved. The system generates a buzzer if patient vitals exceed the nominal value with no remote transmission [4]. Monitors patient's Sp02 with MCU, ZigBee chip, and Sp02 sensor where sensor transmits data to the router, which is further linked to personal computer lacking every time monitoring. After Microcontroller and WSN Based system, Step further towards mobile healthcare system promotes the use most common android plate form for patient's mobile monitoring and alerts. These phones can easily monitor record and receive data collected by bioinstrumentation section. This method provides more flexibility, accuracy, ease, and analysis and reduces extra expenditures. Like in [5] a system with a combination of GSM and GPS first traces outpatient in pain or needs assistance and then send an alarm message to consultants for health care. Further, use of small single chip using Tran's impedance amplification, photodiode current source and photodetector in [6] implements a pulse oximeter that helps in bio-instrumentation with ultra-low power usage. This research work includes developing patient monitoring system using android application plate form with essentials vitals such as include SPO_2 (Oxygen Saturation in Blood), Heart Rate as well as Temperature. This module will assist doctors to monitor a patient using a simple android application on a mobile phone being at distance.

The rest of this paper is organized as follows: Section II presents an overview about system design, Section III defines results, outcomes and analysis about the project and Section IV presents conclusion and Section V presents future work.

II. System Design

This system aims to develop 1) Pulse Oximeter and with Temperature sensor: This part consists of a temperature sensor, red LED, infrared LED, a phototransistor. The Arduino UNO microcontroller interfaces the circuit with a computer and 2) Design for patient bedside monitoring Desk for Consultants (a Connection between sensor parts, web server, and Android application): This part consists of GUI program to connect to connect sensor circuit and computer. Arduino programming calculates the value of temperature and % oxygen in the blood (SpO2) with easy to access interface for the end user (Consultants) to access the measured patient health patient data around the globe. The complete system is depicted in block diagram in Fig. 1.

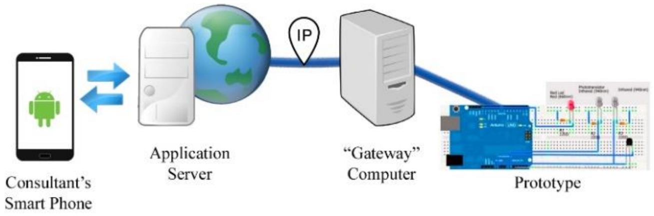


Fig. 1. Generalized block diagram.

The proposed architecture has following modules:

A. Temperature Sensor for Patient

LM35 sensor Fig. 2 whose output voltage is linearly proportional to Celsius is used with patient body contact; this is a precision integrated-circuit for temperature measurement with high preciseness, low self-heating and trimming less.

B. Pulse Oximetry Meter (POM) for Patient

The quantity of Oxygen saturation (Sp02) defines how much oxygen is present in the blood. Most common, reliable and non-invasive method based on Hemoglobin and Deoxyhemoglobin is Pulse Oximetry. Two different Light Wavelengths 660nm (red light spectrum) and 940 nm (infrared light spectrum) are normally used to determine the actual dissimilarity in the absorption spectrum of HbO2 and Hb [7]. At receiver, photodetector collects non-absorbed light from the two LEDs used and presents final output signal after OpAmp. The final output signal has both frequency and non-frequency parts which represent Intensity of Red LED (pulsatile arterial blood) and Intensity of IR LED (venous blood, tissue, and non-pulsatile arterial blood), respectively. The POM (Fig. 3) compares the two wavelengths and calculates Sp02 as follows: % Spo_2 = Intensity of Red LED/ Intensity of IR LED Equation (1).

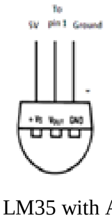


Fig. 2. LM35 with Arduino.

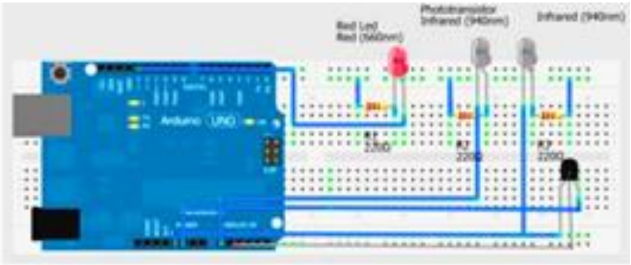
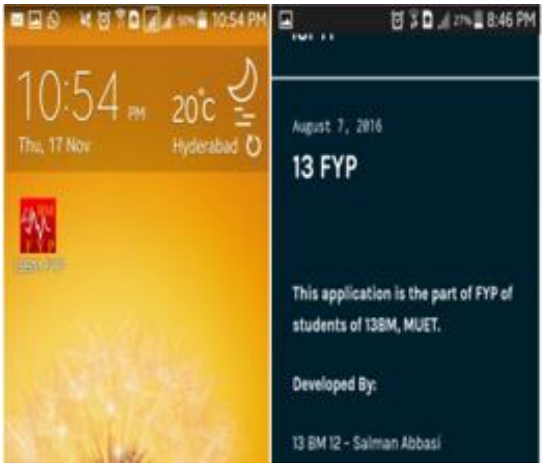


Fig. 3. POM circuit.

C. Arduino Interface with GUI and Android Application

The Arduino IDE is a cross-platform application written in Java and derived from the IDE with efficient compiling and uploading programs to the board (Android programming). Most common Arduino functions to make a program are Setup () and Loop (). The end part of this project is GUI for the end users. GUI connects the equipment to computer and simply click measures percentage of oxygen in blood and body temperature. Serial monitor panel will show the results of these parameters. Built-in GUI processing screen for real-time monitoring is present in the android application (Retomeier, Android Application), however, monitoring reading can also be performed LabVIEW or any other tool, and results can be exported to other software using the port.

Further, Android application (Fig. 4) has been developed for transmission of a vital signal from one device to another in real time. GUI is connected to the web using a port and same port address is given to the Android application. The android application lands on the given port address where the Acquired signals can be visualized.


10:54 PM

10:54 PM 20°C

Thu, 17 Nov Hyderabad

8:46 PM

August 7, 2016

13 FYP

This application is the part of FYP of students of 13BM, MUET.

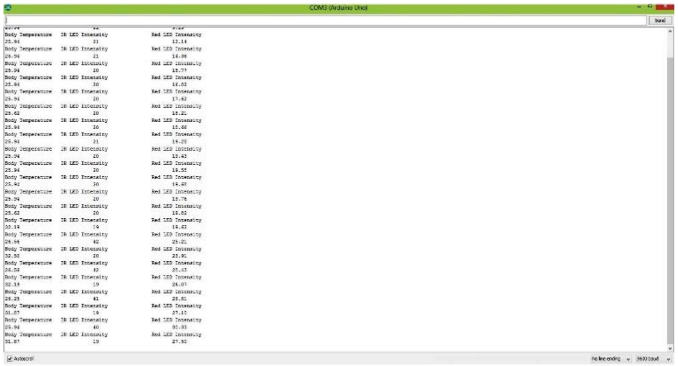
Developed By:

13 BM 12 - Salman Abbasi

Fig. 4. Android applications.

III. Result and Discussion

Results for temperature and readings are of Red and IR LED can be visualized in numerical format on Arduino's serial monitor (Fig. 5) and Graphical representation on android application for temperature and %SpO2 based on Equation (1).



```
Body Temperature   IR LED Intensity        Red LED Intensity
25.94                     21                      13.14
Body Temperature   IR LED Intensity        Red LED Intensity
25.94                     21                      14.86
Body Temperature   IR LED Intensity        Red LED Intensity
25.94                     20                      15.77
Body Temperature   IR LED Intensity        Red LED Intensity
25.94                     20                      16.02
Body Temperature   IR LED Intensity        Red LED Intensity
25.94                     20                      17.62
Body Temperature   IR LED Intensity        Red LED Intensity
25.62                     20                      18.21
Body Temperature   IR LED Intensity        Red LED Intensity
25.94                     20                      18.66
Body Temperature   IR LED Intensity        Red LED Intensity
25.94                     21                      19.25
Body Temperature   IR LED Intensity        Red LED Intensity
25.94                     20                      19.43
Body Temperature   IR LED Intensity        Red LED Intensity
25.94                     20                      18.55
Body Temperature   IR LED Intensity        Red LED Intensity
25.94                     20                      19.68
Body Temperature   IR LED Intensity        Red LED Intensity
25.94                     20                      19.76
Body Temperature   IR LED Intensity        Red LED Intensity
25.62                     20                      18.82
Body Temperature   IR LED Intensity        Red LED Intensity
33.19                     19                      16.42
Body Temperature   IR LED Intensity        Red LED Intensity
28.56                     42                      25.21
Body Temperature   IR LED Intensity        Red LED Intensity
32.50                     20                      23.91
Body Temperature   IR LED Intensity        Red LED Intensity
26.56                     42                      25.63
Body Temperature   IR LED Intensity        Red LED Intensity
32.19                     19                      26.07
Body Temperature   IR LED Intensity        Red LED Intensity
28.25                     41                      28.81
Body Temperature   IR LED Intensity        Red LED Intensity
31.87                     19                      27.10
Body Temperature   IR LED Intensity        Red LED Intensity
25.94                     40                      31.33
Body Temperature   IR LED Intensity        Red LED Intensity
31.87                     19                      27.90
```

Fig. 5. Numerical results for LEDs and temperature.

TABLE I. READINGS ON ARDUINO SOFTWARE

Body Temp	Red Intensity	IR Intensity	%SpO2
28.12ºC	33.84	34.44	98.25783972mg
27.8 ºC	33.12	34.89	94.92691316 mg
27.98 ºC	32.52	34.12	95.31066823 mg
26.3 ºC	32.78	34.76	94.30379747 mg
24.31 ºC	31.23	34.18	91.36922177 mg
25.84 ºC	31.22	33.9	92.09439528 mg
26.65 ºC	31.78	33.67	94.38669439 mg
29.47 ºC	32.19	34.41	93.5483871 mg
27.49 ºC	31.78	33.98	93.5256033 mg
27.4 ºC	32.42	33.87	95.7189253 mg
27.12 ºC	31.32	33.21	94.30894309 mg
28.35 ºC	31.88	34.11	93.46232776 mg
26.11 ºC	31.67	34.05	93.010279 mg
27.3 ºC	32.09	33.19	96.68574872 mg
27.9 ºC	31.83	34.39	92.55597557 mg
27.46 ºC	31.42	33.5	93.79104478 mg
26.52 ºC	31.97	34.12	93.69871043 mg
28.74 ºC	31.71	34.22	92.66510812 mg
27.34 ºC	31.55	33.86	93.1777909 mg
28.62 ºC	31.8	34.67	91.72194981 mg
28.9 ºC	31.39	34.45	91.11756168 mg
29.71 ºC	31.08	33.76	92.06161137 mg
29.11 ºC	31.42	33.87	92.76645999 mg
24.17 ºC	31.12	34.08	91.31455399 mg
24.49 ºC	15.13	45.11	33.54023498 mg
24.66 ºC	12.45	49.14	25.33577534 mg
24.76 ºC	10.33	55.22	18.70699022 mg
22.7 ºC	9.23	57.76	15.9799169 mg
22.34 ºC	12.88	58.9	21.86757216 mg

Table I shows the reading acquired using the prototype, were compiled within thirty seconds, one reading each second, the first column contains the body temperature, second and third columns show the Red and IR LEDs' intensity respectively and the fourth column is showing the SpO2 values calculated using Equation (1) and shown in Fig. 7 and results on android application is shown in Fig. 8.

Fig. 6 represents graphical representation of readings of intensity of Red and IR LED.

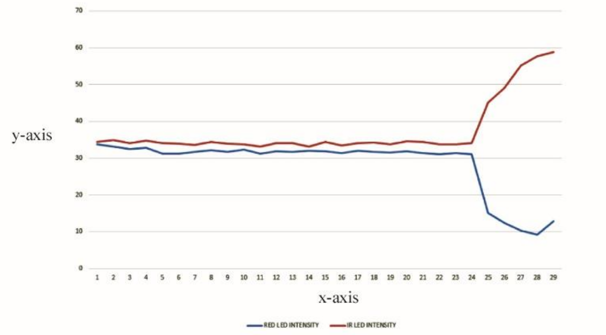


Fig. 6. Readings of intensity of red and IR LEDs.

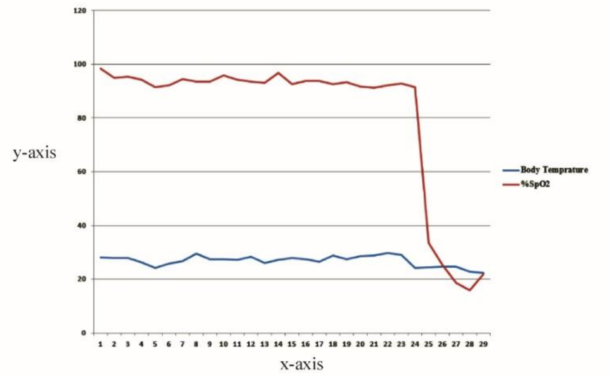


Fig. 7. Results for % SpO2.

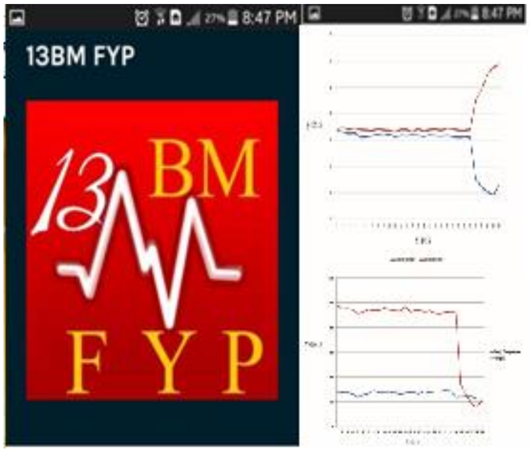


Fig. 8. Results on android application.

IV. CONCULSION

A Well Designed system for Android-based Bed-Side Monitoring Desk is presented and tested to access patient's data on smartphone android application with low complexity, low power consumptions, and high portability. The system has an android application (user- friendly GUI) for consultant smartphones, an Android application Access Server, Database Server and the indigenously designed and developed patient monitoring system having real- time temperature and SpO2 monitoring.

V. FUTURE WORK

In future, research work focuses to include more physiological vitals such as heart rate, Blood pressure, and ECG. Also, a local server can be established to store past and present history of patients so that Surgeons could have a quick analysis of all procedures and treatments patient has gone through.

REFERENCES

[1] S. Trivedi.2017.Android-based health parameter monitoring, 2017 International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, 2017, pp. 1145-1149

[2] Rubina A. Shaikh. 2012. Real-Time Health Monitoring System of Remote Patient Using ARM7 Md.Manirul Islam.2012.Microcontroller Based Health Care Monitoring System Using Sensor Network International Conference on Electrical & Computer Engineering (ICECE), PP. 272-275, Dec. 2012.

[3] Zhou Yan.2010.Design and Implementation of ZigBee Based Wireless Sensor Network for Remote Sp02 Monitor, International Conference on Future Computer and Communication(ICFCC), vol.2, pp. 278-281, May 2010J.

[4] Yuanyuan Du.2011. An Android-based emergency alarm and healthcare management system, IT in Medicine and Education (ITME), vo1.1, pp. 375-379, Dec.2011.

[5] Maziar Tavakoli.2010. An Ultra-low power pulse oximeter implemented with an energy efficient transmission impedance amplifier, IEEE transaction on biomedical circuits and systems, vol.4, No. I, pp. 27-38, Feb 2010.

[6] Santiago Lopez.2012.Pulse oximeter fundamentals and design, in Free scale semiconductor, 2012. Android programming and application development.

[7] Laufer, J., Delpy, D., Elwell, C., & Beard, P. (2007). Quantitative spatially resolved measurement of tissue chromophore concentrations using photoacoustic spectroscopy : application to the measurement of blood oxygenation and hemoglobin concentration. Physics in Medicine and Biology, 52(1), 141–168

Urdu Word Segmentation using Machine Learning Approaches

Sadiq Nawaz Khan[1], Khairullah Khan[2]
Department of Computer Science
University of Science & Technology Bannu, Bannu,
Pakistan

Asfandyar Khan[4]
Institute of Business and Management Sciences
University of Agriculture
Peshawar, Pakistan

Wahab Khan[3]
Department of Computer Science & Software Engineering
International Islamic University
Islamabad, Pakistan

Fazali Subhan[5]
Department of Computer Science
National University of Modern Languages
Islamabad, Pakistan

Aman Ullah Khan[6], Burhan Ullah[7]
Department of Computer Science
University of Science & Technology Bannu
Bannu, Pakistan

***Abstract*—Word Segmentation is considered a basic NLP task and in diverse NLP areas, it plays a significant role. The main areas which can be benefited from Word segmentation are IR, POS, NER, sentiment analysis, etc. Urdu Word Segmentation is a challenging task. There can be a number of reasons but Space Insertion Problem and Space Omission Problems are the major ones. Compared to Urdu, the tools and resources developed for word segmentation of English and English like other western languages have record-setting performance. Some languages provide a clear indication for words just like English which having space or capitalization of the first character in a word. But there are many languages which do not have proper delimitation in between words e.g. Thai, Lao, Urdu, etc. The objective of this research work is to present a machine learning based approach for Urdu word segmentation. We adopted the use of conditional random fields (CRF) to achieve the subject task. Some other challenges faced in Urdu text are compound words and reduplicated words. In this paper, we tried to overcome such challenges in Urdu text by machine learning methodology.**

Keywords—Part-of-speech (POS); NER; word segmentation; information retrieval; Natural Language Processing (NLP); conditional random fields (CRF)

I. Introduction

Natural Language Processing (NLP) is a key area for research in almost every language of the world. In NLP computers are trained in such a way that can easily understand and manipulate human language text or speech. NLP researchers are trying to produce such a knowledge that how human beings understand and use natural language. They use applicable tools and procedures that can be technologically advanced to make computer systems cognize and operate natural languages and achieve the desired tasks. NLP fundamentals lie in various disciplines such as information and computer sciences, electronic and electrical engineering, linguistics, artificial intelligence (AI), mathematics and psychology, etc. [1]. NLP applications consist of various fields of studies, such as text processing and summarization, user interfaces, CLIR (cross-language information retrieval), speech recognition, AI and word segmentation etc. Recognition of valuable and relevant documents from a large collection with respect to the desired query is information retrieval (IR). The technique which is used to process document or collection of documents for identification of events or entities which have been pre-specified is information extraction (IE). Information extraction (IE) is a technique which processes a document, or collection of documents, to identify pre-specified entities or events.

Word Segmentation has significant role in all NLP applications. It has the ability of dividing and separation of written text into meaningful units which are usually known as words. Words boundaries in a spoken language can be identified by word segmentation. Hindi like languages attracted researcher’s attention during recent years. Especially on web Urdu language is going to become a key part of Asian languages [2]. Informational retrieval (IR) and Data Mining (DM) need a detailed knowledge of NLP with responsibilities of the relationship exploration, topic categorization, event extraction and sentiment analysis, etc. NLP significance such as part-of-speech (POS) tagging, morphological analysis, named entity recognition, stop words removal, parsing and shallow parsing have significant importance in all NLP systems [3]. Urdu word segmentation problem is not unadorned as some of the other Asian languages, in which space is used for word demarcation, but it has not consistently been used. The use of space gives rise to both space omission and space insertion problems in Urdu text [4] and [5]. The Space omission problem e.g. the Urdu word "انکا" which is actually a combination of two words but the system treats it as a single word. Such Segmentation in Urdu text is handled with the application of Urdu-Devnagri transliteration system [6]. The Space Insertion problem e.g. the word "عقل مند"(Aqalmand, Intelligent) is actually one word but when segment it will be treated as two words i.e. عقلand مند which is handled by a two-stage system [7]. Hindi-Urdu transliteration issues are briefly discussed by [8] and [9]. Simple, compound and complex words are segmented for Sindi language using three layers [10]. A complete survey of techniques regarding Urdu-Arabic Word Segmentation and also their challenges have discussed by [11].

II. Literature Review

Nowadays different languages use different techniques for word segmentation problem so far. These techniques are used by NLP researchers and have deduced better results from each one. The existing techniques for word segmentation in NLP are Dictionary/rule-based, statistical/machine learning and hybrid approaches.

A. Existing Techniques

There are some techniques which are commonly used for word segmentation and some are not widely used yet. The detail of these techniques is given below:

1) Rule-Based Techniques: Rule-based techniques are set of rules or pattern which are used to perform various NLP tasks. Rule-based approaches were constructed manually by linguistics experts. This approach was used by [12] for chines word segmentation. They also show a transformation-based algorithm for improving the output of the system. As Urdu, Chines, Japanese and Thai etc have not delimited by spaces, therefore word segmentation is how much difficult as compared to other western languages like English etc. Word segmentation for Thai language using rule-based technique was presented by [13]. An Urdu stemmer namely "Assas Band" developed by [14] is based on rule-based. Assas-Band firstly removes the prefix from the stem and then postfix and finally stem is extracted with the accuracy of 91.2%. Urdu online handwriting recognition system provided by [15]. Author in [16] has used the rule-based technique for Name Entity Recognition in Urdu. Urdu word segmentation using this approach is done by [5].

2) Machine Learning/Statistical Techniques: Machine learning approach is much better than rule-based approaches although this technique is not commonly used for word segmentation. These techniques use learning algorithms which are capable of defining a function that takes input samples to a range of output values. A corpus is constructed for these approaches in which word boundaries are explicitly defined. Statistical models are formed containing features of the words which have been surrounded by boundaries. Supervised statistical learning is one of the most current dominant technique in NLP. This approach automatically induces rules from training data. Machine learning algorithms consist of intelligent modules. Different machine learning models have been discussed by [17]. In order to carry out major NLP task using statistical approaches, it incorporates stochastic and probabilistic methods. A two-stage word segmentation system for handling space insertion problem in Urdu by [7] is done using the statistical-based technique. The space omission problem in Urdu word segmentation using this approach has been used by [6].

3) Hybrid Approaches: Hybrid techniques are the combination of features of rule-based and statistical techniques. Authors in [18] presented a hybrid approach for Urdu sentence boundary disambiguation comprising of unigram statistical model and rule-based algorithm. These results better than rule-based and statistical based approaches. Hybrid technique for segmentation presented by [19] uses top-down mechanism for line segmentation and bottom-up design for segmenting the line into ligatures. The accuracy result was achieved 99.2% using this technique.

III. Urdu Language

Urdu is the National language of Pakistan. The hand-held devices such as mobiles phones, etc. have been successfully using everywhere but the software they provide for user input is mostly in English and in Pakistan, it is difficult for a common man to communicate in English easily. In order to facilitate Urdu speakers and writer and reduce the difference between the common man and the new technology, Urdu NLP systems are required. We have tried to bridge this gap by using machine learning approach for segmentation of Urdu text.

A. Urdu Writing Style

Urdu is not scripting language although it is written in cursive Arabic script. Arabic script has many traditional writing styles such as Naskh (mostly used for the Arabic language), Taleeq, Kufi, Divani, Sulus, Riqa, etc. As Nastalique is complex writing style but it is novel and robust and most commonly used for Urdu writing. Nastaleeq is character based, bidirectional (mainly right to left), diagonal, on-monotonic, cursive, context-sensitive writing system with a significant number of marks (dots and other diacritics).

1) Urdu Characters: Urdu has 50 consonants in which 35 are simple and 15 are aspirated. There are 15 diacritical marks and 1 character for nasal sound. Consonant letters, vowels, diacritic marks, numerals, punctuations and few superscripts signs support Urdu text. Urdu text can be written with simple characters or characters with diacritical marks. Both format conveying same meaning but the only difference is in writing and oral saying e.g. the Urdu word having simple character "***ڈولکفل***' is same in oral saying as "***ڈُولکِفل***' which have two diacritic marks i.e ◌ُ and ◌ِ. For segmentation, such diacritic marks will have to remove first. Table I shows the Urdu digits and characters, while Table II shows some other characters which are not counted as part of the alphabet, punctuation marks, signs, and symbols of Urdu text.

TABLE I. URDU DIGITS AND ALPHABETS

Urdu Writing Style	Digits & Alphabets		
	Numbers & Characters	*Numbers*	*Characters*
		٥ ٦ ٧ ٨ ٩ ١ ٢ ٣ ٤ ٠	ا ب بھ پ پھ ت تھ ٹ ٹھ ث ج جھ چ چھ ح خ د دھ ڈ ڈھ ر ڑ ڑھ ز ژ س ش ص ض ط ظ ع غ ف ق ک کھ گ گھ ل م ن و ہ ی ے

TABLE II. TABLE TYPE STYLES

Urdu Writing Style	Diacritics, punctuation mark, signs & symbols		
	Characters not counted as part of alphabet/diacritics	*Punct uation marks*	*Signs & Symbols*
	◌ٌ ◌ً ھ ں آ ۂ ۓ ئ ؤ ء ◌َ ◌ّ ◌ِ ◌ُ ◌ٍ ◌ٚ ◌ٔ ◌ٖ ◌ٰ ◌ٗ	٬ ۔ ؟ ؛ ، ‘	ع صـــہ ســـ ـــ ـہ ◌ؔ ◌ؓ ◌ؒ ◌ؑ ◌ؐ بِسۡمِ ٱللَّهِ ٱلرَّحۡمَٰنِ ٱلرَّحِيمِ

2) *Joiners*: Urdu script is cursive and characters are joined with neighbor within a word and acquire different shapes. Such characters are known as joiners. Joiners have four-way shaping i.e. initial, medial, final and isolated form. Table III shows some examples of four-way shaping form of joiners and Table IV shows joiner characters of Urdu text.

TABLE III. FOUR-WAY SHAPING OF JOINERS

Urdu Writing Style	Four-way Shaping of Urdu Joiners			
	Final	*Medial*	*Initial*	*Isolated*
	بب	ببب	با	ب
	بت	بتب	تا	ت
	بٹ	بٹب	ٹا	ٹ
	بی	بیب	یا	ی

TABLE IV. JOINERS IN URDU

Urdu Writing Style	Joiners in Urdu
	Joiners
	ب ت ة ث ٹ ج چ ح خ ہ س ش ص ض ط ظ ع غ ف ق ک گ ل م ن ء ی

3) *Non-Joiners*: Some Urdu characters are not joined with the neighbor ones, such characters are referred to as non-joiners. Non-joiners have only two forms i.e final and isolated. The following Table V shows some examples of the final and isolated forms of non-joiners whereas the Table VI shows non-joiner characters of Urdu text.

TABLE V. FORMS OF NON-JOINERS

Urdu Writing Style	Forms of Non-Joiners		
	Urdu non-joiners	*Final*	*Isolated*
		با	ا
		بد	د
		ظر	ر
		کو	و
		کے	ے

TABLE VI. NON-JOINERS IN URDU

Urdu Writing Style	Non-Joiners in Urdu
	Non-Joiners
	اآ د ڈ ذ ر ز ڑ ژ و ے

B. Urdu Linguistics Resources

Urdu lexical resources are a necessary part of every NLP system for the computational processing of Urdu language. In Pakistan the area of applied linguistic such as English language teaching (ELT) and sociolinguistic are the two highly focused fields by the researchers. Very trivial study has been reported in respect of descriptive and theory-based linguistics and there has an evenly finite capability in that area in Pakistan. For the purpose stated, one of the leading "Essential Urdu Linguistic Resources project" is concentrating on building up indispensable Urdu lingual resources and tools by ramping up research capability in grammatic and semantic studies. This coaction will assist research community to abloom the area of linguistics inside Pakistan.

The Urdu corpus and lexical resources developed for Urdu has been discussed by [20] are listed below:

1) *Urdu Encoding Scheme*: The computer keyboard is used as input device for entering data to the computer. It contains characters, numbers, functional keys and symbols etc. Special encoding technique is used when the computer gets input. Character encoding is the process of assigning a unique number to each character of the language. This code is generated internally in a computer system. For Urdu language, different encoding schemes have been developed but for standardization of encoding scheme, no effort was undertaken.

2) *UTF*: The Unicode organization is responsible to develop and assign a unique character encoding scheme for digital text of almost all languages of the globe. The most general Unicode character scheme that are commonly in use are UTF-8, UTF-16, and UTF-32. Majority operating systems are based on UTF-16 encoding scheme. This encoding scheme is adopted as worldwide encoding scheme and is capable to map all known characters.

3) *Urdu Zabta Tahti (UZT)*: As there are no industry standards for coding in Urdu, similar to ASCII standard for English, therefore, it needs much attention. For this purpose Urdu Zabta Tahti (UZT) version 1.01 by [21] is a standard code page for Urdu. The Government of Pakistan has accepted UZT version 1.01 as a standard code for Urdu.

4) *Urdu Text Corpus*: In 2002 Becker and Riaz released the first publicly freely available Urdu dataset to promote research activities in Urdu. In its development the contents of 7000 news articles were used which was extracted from BBC

Urdu URL. The Becker and Riaz dataset contains very reach contents and is considered feasible for majority of ULP tasks such as Part of speech tagging, named entity recognition and so on.

EMILLE project has made Urdu corpus for the first time by [22]. The corpus has 200,000 words of English text translated into Urdu etc. and 512000 words of spoken Urdu and 1640000 words of Urdu text.

5) CLT Conference: In Pakistan the Society for Natural Language Processing (SNLP) has taken initiative steps to arrange a series of international conference, namely, Conference on Language and Technology (CLT) with the objectives to abide students, researchers of various universities and research institutions to share research ideas and to promote research culture in Pakistani and South Asian languages.

6) SNLP: Recently researcher has shown growing interest in the computational processing of Urdu digital text in Pakistan. In Pakistan there are assorted number of organizations and individuals which perform research activities in isolated manners and there exists no coordination among various organizations and individuals.

An integrated exertion is necessary to bring in them in collaborative platform to present ideas and pass around information. SNLP renders a research platform for organizations and individual researchers for this aim.

These days more than 60 languages are mouthed in Pakistan; hence we can state that Pakistan symbolizes a diverse still adhesive lingual and cultural environment. Lot languages are interconnected and several are generally mouthed crosswise territorial bounds. Hence, there is a demand to build up a basic platform to draw together the research community processing these languages.

IV. Conditional Random Fields

CRF is a machine learning algorithm, which is widely used in Natural Language Processing (NLP) tasks e.g. word segmentation, sequential labeling, Name Entity Recognition and so on. Conditional Random Fields (CRFs) are undirected graphical models used to calculate the conditional probability of values on designated output nodes given values on designed input nodes. CRF has several advantages over Hidden Markov models and stochastic grammar models (Lafferty, McCallum, & Pereira, 2001) and defines a CRF on X and random variable Y as follows:

Let the graph $G = (V, E)$ such that $Y = \left(Y_y\right)$vεv) so that Y is indices by the vertices of G. Then (X, Y) is conditional random field when the random variable Y_v, conditioned on X, obey the Markov property with respect to the graph: $p\,(Y_y/X, Y_{W,}\, w{\sim}v)$ means that w and v are neighbors in G. For sequence tagging tasks, the LDCRF (Latent-dynamic random fields) or DPLVM (Discriminative Probabilistic Latent Variable Models) are a type of CRFs for sequence tagging tasks. These models are known as latent variable modela that are trained discriminatively. According to LDCRF let a given sequence of observations say, $X = x_1, x_2, x_3,\ x_n$ one of the tagging task but here the problem arises that how to assign sequence of labels and this problem should be solved by the model let $Y = y_1, y_2, y_3, \dots y_n$, be a labels sequence. In ordinary linear-chain CRF, latent variables 'h' is inserted between x and y rather than directly modeling $P(Y/X)$. It uses chain rule probability.

$$\mathrm{P(Y/X)} = \sum_{\mathrm{h}} \mathrm{p(Y/h\,, X)P(h/X)} \qquad (1)$$

Suppose $\mathrm{x_{1:n}}$ is a sequence of Urdu words in a sentence with name entities $\mathrm{z_{1:n}}$. According to linear chain CRF, the conditional probability is as:

$$\mathrm{P\,(z_{y:n}} \Big/ \mathrm{x_{1:n}}) = 1 \Big/ \mathrm{Z} \exp\left(\sum_{\mathrm{n=1}}^{\mathrm{N}} \sum_{\mathrm{i=1}}^{\mathrm{F}} \lambda_{\mathrm{i}}\ \mathrm{f_i}(\,\mathrm{z_{n-1}},\ \mathrm{z_n},\ \mathrm{z_{1:n}}, \mathrm{n})\right) \qquad (2)$$

Where the normalization factor Z is calculated as under

$$= \sum_{\mathrm{z_{1:n}}} \exp\left(\sum_{\mathrm{n=1}}^{\mathrm{N}} \sum_{\mathrm{i=1}}^{\mathrm{F}} \lambda_{\mathrm{i}}\ \mathrm{f_i}(\,\mathrm{z_{n-1}},\ \mathrm{z_n},\ \mathrm{z_{1:n}}, \mathrm{n})\right) \qquad (3)$$

V. Name Entity Recognition

NER was first introduced in 1995 as part of MUC-6 (Message Understanding Conference). Later on, in 1996, the MET-1 conference introduced the name entity recognition in the non-English text. Name entity is one of the prior tasks in NLP. Named entity recognition consists of identifying within sentence words or sequences of adjacent words belonging to a certain class of interest or it classifies proper nouns into its predefined categories such as a person, time, date, brand names, quantities, monetary values, percentages, abbreviations, location, organization, etc. For each class of interest, the labeling distinguishes between the first word in the named entity and the following words in the named entity. Words not belonging to any class of interest are labeled as O (other). Name entity recognizer is the software which labels sequence of words in a text. Word segmentation has been applied in several tasks e.g. NER, IR, automatic speech recognition, machine translation, etc. There are two types of approaches to utilize word segmentation in such tasks: pipelining and joint-learning. The pipeline approach creates word segmentation first and then feeds the segmented words into the subsequent task(s). The joint-learning approach trains a model to learn both word segmentation and the subsequent task(s) at the same time. Many NER types of research are based on word segmentation and even Part-Of-Speech (POS) tagging. The relationship between them is described in Fig. 1.

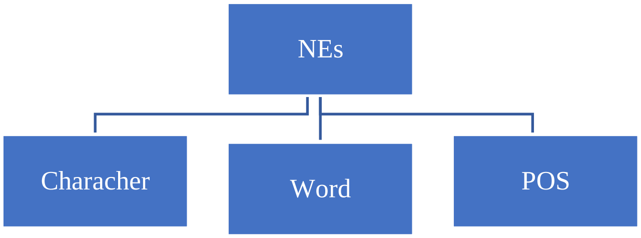


Fig. 1. NER model for segmentation.

The main goal of NER is to recognize the name entities and then resolve the ambiguities from them. Two types of ambiguities are common in names i.e. structural ambiguity and semantic ambiguity has been discussed briefly by [23]. They implemented a module for proper names recognition. Considerable work has been done for NER in western languages such as English, etc. but the interest for NER in South Asian languages has not been developed so far. The main reason is lack of technologies for South Asian languages. Urdu is one of the most important languages of South Asia and a lot of efforts are going on for the development of this language throughout the world especially in Pakistan because Urdu is a national language of Pakistan. The first effort in NER for South Asian languages was made by [24], who highlighted the main challenges facing NER for the Urdu language. They created Becker-Riaz Urdu corpus for the first time as there was no other resource available at that time. In IJCNLP conference 2008, a comprehensive attempt for NER was made for South Asian languages. Many experiments have done for NER in Urdu which uses CRF up to some extent, but need more attention and deep study while using CRF as a module for NER in Urdu.

CRF Classifier provides a general implementation of (arbitrary order) linear chain Conditional Random Field (CRF) sequence models for any task. In our work, the NER structured as to consider the following Urdu sentence.

پاناما کیس کے فیصلہ سے قبل گیلپ پاکستان نے عوامی رائےجانی۔

[پاناما کیس: NOR] [کے: NOR] [فیصلہ: NOR] [سے: NOR] [قبل: NOR] [گیلپ: PER] [پاکستان: LOC] [نے: NOR] [عوامی: PER] [رائے: NOR] [جانی: NOR]

VI. Challenges in Urdu Word Segmentation

Urdu word segmentation faces different challenges such as Space omission problem, space insertion problem, compound words, reduplicated words, affixations and English Abbreviations. All these challenges of word segmentation are briefly discussed below.

A. Space Insertion Problem

When space is inserted in between two words of Urdu then the space insertion problem arises. In handwritten Urdu text, there is no space inserted in between words and are briefly discussed by [8], [5] and [6]. When the ending character of the word is joiner then space must be inserted to separate the words otherwise they make a miss-understandable form which the system does not recognize it however the native speaker of Urdu can understand. E.g. consider the Urdu words داخلہ فارم (dahla form, Admission Form) ہم منصب (hum mansab, counterpart) is a combination of four words but semantically these are two words. Now if we remove the spaces in between the words then the above words look like داخلہفارم، ہممنصب which having visually incorrect shape, means in such like cases space must be inserted in between the words otherwise system will not recognize such words. But the problem hereby arises that if we put space in between such words then it is also difficult for a system to take it as a single word because such words are a combination of different words. Similarly, consider the whole sentence, "اقرارکرنایاانکارکرنا" (iqrar karna ya inkar karna, accept or refuse) having five separate words اقرار، کرنا، یا، انکار، کرنا can easily understand and segmented by the native speaker of Urdu. But the system will take this whole sentence as a single word. Space insertion problem causes due to multiple reasons which have been briefly discussed by [5].

B. Space Omission Problem

When space is omitted in such a place where it should be inserted for the appropriate form of the word, then space omission or space exclusion problem arises. Space omission in Urdu text is also a challenging task for word segmentation. If a word ends with a joiner character then it should be separated by a space otherwise it will append to next word which then gives visually incorrect shape. Consider the word شاہی قلعہ (Shahi Qilla), if space is omitted then it will look like شاہیقلعہ having a visually incorrect shape for reader and system as well. But there are some words in which if space is omitted then they do not lose their meaning and have correct shape also. Consider the words: آپ کا(yours) ،(will do) کرے گی کے لئے، (for), (narrate) بیان کریں , اس وقت(at that time), after omitting the space in between these words they make the forms: آپکا، کریگی، کیلئے، بیانکریں، اسوقت all these shapes are acceptable and understandable by the system and the native speakers (Durani & Hussain, 2010). Thus we can say that space is not always used as a word boundary in Urdu. One of the considerable approach for handling space omission problem in Urdu word segmentation is used by (Lehal, 2010), which is based on Urdu-Devnagri transliteration system, in which Urdu words are translated into Hindi Devnagri and then segmented.

C. Compound Words

Compound word is the combination of two or more lexemes to form another lexeme [25]. Compounding is the process in which new units of thought are formed. [8] have categorized the compound words into three categories.

- AB
- A o B
- A e B

The examples of Urdu words in the above formats are جیل خانہ(jail khana, Prison), مخنت و عظمت(mehnat o azmat, hardworking and greatness) and حالتِ زار(halat-e-zar, bad condition). In our system, these compound words are handled while doing word segmentation.

D. Reduplicated Words

Reduplicated are those words in which one word/morpheme occurs twice consecutively. Jawaid & Ahmed, 2009 has discussed the Urdu reduplicated words: دن بدن(din ba din, day by day), کبھی کبھی(Kabhi Kabhi, whenever). By observing the above two reduplicated words it is concluded that in reduplication one word is repeated twice or a morpheme is added to that word and make reduplicated word e.g. in دن بدن word the morpheme ب is added to the repeated word دن. The reduplicated words will treat by the system as separate orthographic words (Durrani & Hussain, 2010).

In Urdu word segmentation such words need proper attention and in our work, these words are handled up to some extent.

E. Affixations

In Urdu text affixation (prefixes and suffixes) are used e.g انتھک(anthak, tireless) is an example of prefixes which should be a single word [3]. Similarly, the examples of words with suffixes بد اخلاق(bad akhlaq, bad character), با وقار(ba Waqar, honorable) etc should also consider as single words [14].

F. English Words

Urdu is a language which borrows words from other languages such as Arabic, Farsi, Greek, Latin, and English etc. Abbreviations of English in Urdu writing needs a space/dash character in between the words [8], e.g. Ph.D. (پی ایچ ڈی) or (پی-ایچ-ڈی), M.Phl (ایم فِل) etc.

VII. Urdu Word Segmentation Model

The proposed CRF based Urdu word segmentation model makes use of named entities and POS information of words as a feature for the subject task.

For POS tag information we used CLE POS tagged corpus and for NE information we used the UNER dataset [26]. The UNER dataset contains only NE tags since POS information of particular words provides important information about the basis of the word. Therefore, to make the UNER dataset more informative for feature learning task we first assigned POS tags to each word of the UNER dataset. For this purpose, we make legal use of CLE POS tagged corpus. The assignment of POS task is achieved with help of longest maximum matching technique.

After POS tag assigned to the whole UNER dataset CRF model is trained on this UNER dataset containing both POS and NE tags. This new UNER dataset is used to generate a model file with help of feature set provided in below table. The resultant training model file of CRF is then used along with lexical dictionary file for testing test data. The following Table VII shows the feature template for our proposed model.

TABLE VII. Feature Template for Proposed Model

Features	
Feature Template	***Description***
U01:%x[-1,0]	N-1 token
U02:%x[0,0]	Current token
U03:%x[1,0]	N+1 token
U04:%x[-1,0]/%x[0,0]	N-1 word and N+1 token
U05:%x[0,0]/%x[1,0]	Current token and N+1 token
U06:%x[-1,0]/%x[1,0]	N-1 token and N+1 token
U07:%x[-1,1]	POS tag of N1 token
U08:%x[0,1]	POS tag of the current token
U09:%x[1,1]	POS tag of N+1 token

Fig. 2 below shows the graphical depiction of proposed CRF.

A brief summary of the steps is below:

- UNER dataset is pre-processed
- CLE corpus is pre-processed
- POS tags are assigned to UNER dataset using Longest maximum matching techniques
- The new UNER dataset is then modeled in CRFSharp package requirements
- CRF is trained using the feature template
- The model file is generated
- Test data is tested for word segmentation task against the model files and dictionary files
- Output is generated
- Result is calculated
- Results are averaged

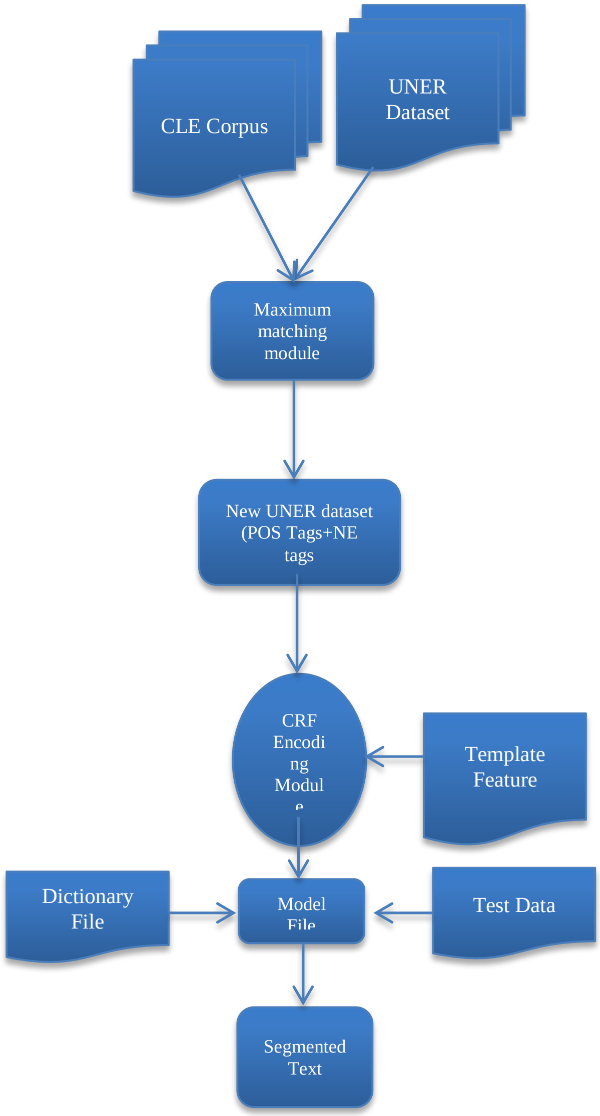


Fig. 2. Graphical depiction of proposed CRF model.

VIII. Experiments and Evaluations

To evaluate the performance of our proposed system we used WordSeg [1] libraries a C# implementation. Training corpus contains 320413 Urdu words, in which compound, reduplicated and foreign words are also included. The overall performance of the system is evaluated using Precision, Recall and F-measure (F-score). Precision and Recall are inversely related to each other as Precision increases, Recall decreases and vice versa. F-measure is the value gained from calculating the harmonic mean of Precision and Recall. For testing Urdu text was taken from [2] BBC site. The text was in the form of sentences in four cases. Table VII shows the Precision, Recall and F-score values for the test data. The tested Urdu text is in the form of sentences and the number of sentences and words for the four cases are given in Table VIII.

TABLE VIII. Tested Text Results

Results	**Precision, Recall & F-score values of Urdu tested text**			
Tested Text				
Sentence	***Words***	***Precision***	***Recall***	***F-Score***
1	23	100%	50%	67.5%
3	50	94%	51%	66%
6	99	94%	51%	66%
31	497	96%	50%	65.7%

The results show the average values of Precision, Recall and F-score for all the tested four cases are 96%, 50.7%, and 66.3%, respectively. It was observed that increasing the training data for Urdu word segmentation improves the results as well. The main challenges in Urdu word segmentation i.e. space insertion and omission problems, reduplication, compound words and foreign words are covered up to some extent depending on the training corpus.

In this study, we considered the research work of [27] as baseline work. The comparison of the proposed system with baseline work is shown in Table IX:

TABLE IX. Comparison of proposed CRF Model with Baseline Approach

Results Comparison	**Comparison of Proposed CRF Model with Baseline Approach**				
Approaches	***Problem Addressed***	***Tested Text***	***Correctly Segmented Words***	***Uncorrected Segmented Words***	***Accuracy***
Baseline Approach	Space Omission	11,995	11,723	272	97.2%
Proposed CRF Approach	Space Omission, Deletion, compound, Reduplicated, Abbreviation, English Words	3,161	3,118	43	98.6%

IX. Conclusions

In this paper we have presented a system for solving Urdu word segmentation using machine learning approaches i.e. CRF algorithms. The task of Urdu word segmentation is more challenging as compared to other Asian languages because of space problems in between the words. The objective of this study was to present ML based new system for Urdu word segmentation in which both the main issues of segmentation i.e. space insertion and space deletion as well as compound words and reduplicated words, are handled up to some extent. We believe that the proposed word segmentation system is more advanced system when compared to previous systems as it addresses simultaneously space insertion, space deletion, compound words and reduplicated words challenges.

References

[1] G. G. Chowdhury, "Natural language processing," Annual review of information science and technology, vol. 37, pp. 51-89, 2003.

[2] S. Mukund, R. Srihari, and E. Peterson, "An Information-Extraction System for Urdu---A Resource-Poor Language," ACM Transactions on Asian Language Information Processing (TALIP), vol. 9, p. 15, 2010.

[3] A. Daud, W. Khan, and D. Che, "Urdu language processing: a survey," Artificial Intelligence Review, pp. 1-33, 2016.

[4] N. Durrani, "Typology of word and automatic word Segmentation in Urdu text corpus," 2007.

[5] N. Durrani and S. Hussain, "Urdu word segmentation," in Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, 2010, pp. 528-536.

[6] G. S. Lehal, "A word segmentation system for handling space omission problem in urdu script," in 23rd International Conference on Computational Linguistics, 2010, p. 43.

[7] G. S. Lehal, "A two stage word segmentation system for handling space insertion problem in Urdu script," analysis, vol. 6, p. 7, 2009.

[8] B. Jawaid and T. Ahmed, "Hindi to Urdu conversion: beyond simple transliteration," in Conference on Language and Technology, 2009.

[9] A. Malik, L. Besacier, C. Boitet, and P. Bhattacharyya, "A hybrid model for Urdu Hindi transliteration," in Proceedings of 2009 Named Entities Workshop: Shared Task on Transliteration, 2009, pp. 177-185.

[10] J. Mahar, H. Shaikh, and G. Memon, "A Model for Sindhi Text Segmentation into Word Tokens," Sindh University Research Journal-SURJ (Science Series), vol. 44, 2012.

[11] A. Mahmood, "Arabic & Urdu Text Segmentation Challenges & Techniques," vol. IV, pp. 32-34, 2013.

[12] D. D. Palmer, "A trainable rule-based algorithm for word segmentation," in Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics, 1997, pp. 321-328.

[13] Y. El Hadj, I. Al-Sughayeir, and A. Al-Ansari, "Arabic part-of-speech tagging using the sentence structure," in Proceedings of the Second

[1] https://github.com/zhongkaifu/CRFSharp
[2] http://www.bbc.com/urdu/sport

International Conference on Arabic Language Resources and Tools, Cairo, Egypt, 2009.

[14] Q.-u.-A. Akram, A. Naseer, and S. Hussain, "Assas-Band, an affix-exception-list based Urdu stemmer," in Proceedings of the 7th workshop on Asian language resources, 2009, pp. 40-46.

[15] S. Malik and S. A. Khan, "Urdu online handwriting recognition," in Emerging Technologies, 2005. Proceedings of the IEEE Symposium on, 2005, pp. 27-31.

[16] K. Riaz, "Rule-based named entity recognition in Urdu," in Proceedings of the 2010 named entities workshop, 2010, pp. 126-135.

[17] W. Khan, A. Daud, J. A. Nasir, and T. Amjad, "A survey on the state-of-the-art machine learning models in the context of NLP," Kuwait Journal of Science, vol. 43, 2016.

[18] Z. Rehman and W. Anwar, "A hybrid approach for urdu sentence boundary disambiguation," Int. Arab J. Inf. Technol., vol. 9, pp. 250-255, 2012.

[19] G. S. Lehal, "Ligature segmentation for Urdu OCR," in 2013 12th International Conference on Document Analysis and Recognition, 2013, pp. 1130-1134.

[20] S. Hussain, "Resources for Urdu Language Processing," in IJCNLP, 2008, pp. 99-100.

[21] S. Hussain and M. Afzal, "Urdu computing standards: Urdu zabta takhti (uzt) 1.01," in Multi Topic Conference, 2001. IEEE INMIC 2001. Technology for the 21st Century. Proceedings. IEEE International, 2001, pp. 223-228.

[22] R. Carter and J. McRae, Language, literature and the learner: Creative classroom practice: Routledge, 2014.

[23] N. Wacholder, Y. Ravin, and M. Choi, "Disambiguation of proper names in text," in Proceedings of the fifth conference on Applied natural language processing, 1997, pp. 202-208.

[24] D. Becker and K. Riaz, "A study in urdu corpus construction," in Proceedings of the 3rd workshop on Asian language resources and international standardization-Volume 12, 2002, pp. 1-5.

[25] R. W. Sproat, Morphology and computation: MIT press, 1992.

[26] W. Khan, A. Daud, J. A. Nasir, and T. Amjad, "Named Entity Dataset for Urdu Named Entity Recognition Task," Organization, vol. 48, p. 282, 2016.

[27] R. Rashid and S. Latif, "A dictionary based urdu word segmentation using maximum matching algorithm for space omission problem," in Asian Language Processing (IALP), 2012 International Conference on, 2012, pp. 101-104.

A Systematic Review of Cyber Security and Classification of Attacks in Networks

Muhammad Kashif[1], Sheraz Arshad Malik[2], Muhammad Tahir Abdullah[3], Muhammad Umair[4]
Department of Information technology,
GC University,
Faisalabad, Pakistan

Prince Waqas Khan[5]
Department of Computer Science,
University of Agriculture,
Faisalabad, Pakistan

***Abstract*—Cyber security plays an important role to secure the people who use internet via different electronic devices in their daily life. Some causes occurred all over world that people face problems when they connect their devices and system via internet. There are some highly sensitive data like biotechnology and military assets which are highly threatened by the hackers; cyber security plays a vital role in securing such data. Misusing the internet becomes a current issue in different sectors of life especially in social media, universities and government organizations. Internet is very useful for students in study institutes and employees who work in different organizations. Internet source gives the facility to people to fetch some information via internet. However, they must be protected when use the internet and secure for any unauthorized access. In this paper we have covered the different aspect of cyber security and Network security in the modern era. We have also tried to cover the threats in Intranet of organizations.**

Keywords*—Cyber security; internet; intranet; network security; cybercrime and security alludes

I. Introduction

During the last decade, access to information technology (ICT) has surged across the world. Broadband technology is now available to billions more it was just years in the past. The access of computer internet technology has large benefits but on the other hand there are several issues to reaping these benefits. Among all those challenges one of the biggest challenge is cyber security. In any computer network environment successful security infrastructure must meet the three basics objectives that are CIA (confidential integrity, availability) without obtaining the goals of these requirements it's not feasible to achieve the goal of secure internet. Against a setting of the Internet of Things-based DDoS assaults researchers come to know that in the recent years, it is termed as "death of the internet". Now it is very easy for someone who wants to turn down the internet. They will start attacking the service provides [1]. In this paper we have discuss different type of Security attacks which can affect internet. Here we discuss about the major attacks of daily routine which are the reason of security breach. We can also mention the counter measure how we can secure our Internet from those attacks Internet security is one of the biggest challenge of computer era as we know that internet is also provided Internet As A Services (IAAS). Security is becoming challenge for both internet consumers and internet provider computer networks are part of our daily life their security as important as on the off chance that we have a learning of different web assaults as we can shield yourself from these assaults. Purpose of implementing CIA triad is to make internet secure and the whole network. The purpose of this model is to save from damage of internet security and to stop misuse of data. There are many ability pitfalls which could stand up if network protection isn't implemented well.

Internet is a telecommunications network that use guided and unguided media (phone lines, satellites, and links) to associate PCs and different gadgets to the internet. All PCs, mobiles, TV, and diverse electronic gadgets can associate with the Internet. Internet is becoming an important source for people of world who use in different kinds of work in their daily life. The graph of using internet in all over world including developed and developing countries has increasing day by day. The latest technologies are also introducing with passage of time the people's requirement related to internet and fast access of any thing is also increase. Web was utilized as a part of military, guard working activities, and for one of a kind college's examination purposes. Presently in this century it has developed in all sectors of life including information, business, social life, education, entertainment, data sharing service, and shopping. There are many advantages and disadvantages. In disadvantages of internet is online fraud, cyber terrorism, identity theft, and other cybercrimes are mostly commit by any unauthorized third party. This reason cyber security is very important for users of internet and they feel protected when they will use internet [2]. Cyber security gives the assurance and guarantees to internet users their communication way/internet is secure and protected against any attack of unauthorized third person.

II. Cyber Security

Cyber security is the set of rules, body of technologies, processes and procedures to protect the electronic data, networks, computers, and programs from any attack and unauthorized access. Cyber security must satisfy three points:

1) Measure amount of data for the protection of information technology.

2) The Level of protection as an outcome from application of those taken measures.

3) The field associated with the professional endeavor.

These three aspects of cyber security play a vital role to prevent and secure a personal data of every user of internet, business, and government [3]. Those data are essential because they can be hacked by other person for illegal activities. Various powers oversee the lofty ascent in hostile cyber intrusions and unapproved network breaks. The blast of new advancements and development of societal reliance on all-inclusive interconnected innovation, joined with the robotization and commoditization of cyberattack tools, digital aggressor modernity, and low passage hindrances into the cybercrime market 10 are no uncertainty among the key ones [4].

III. Digital Security Alludes

The significance of cyber security tends to be chosen in various settings. At times, it alludes to monetary terms or in social and social terms or even in politics and military terms. As it is regularly utilized, “digital security” alludes to 3 (three) things:

1) A course of action of activities and diverse measures wanted to guarantee from strike, intrusion, or distinctive risks PCs and unique segments of the web.

2) The state or nature of being protected from such risks. The activities can join security surveys, settle organization, approval systems, get to organization, and so on.

3) The extensive field of take a stab at, including investigation and examination, went for realizing and improving those activities and quality [3].

Web client is developing significantly in assortment age and the motivation behind utilizing web at that point is done in different courses as clarified previously. They can incorporate, such as, investigate and estimate the feature and susceptibility of the gear and programming used as a piece of the country's political and money related electronic structure. They furthermore incorporate revelation and response to safety occasion, lightening of impacts, and repossession of pretentious portions. Diverse compute can join such things as gear and programming hardware & software firewalls, physically protection, for instance, cemented workplaces, and personnel getting ready and commitments. The quantity of web clients in Indonesia for instance is expanding each year. As indicated by Internet [5] business web administrations started in Bahasa Indonesia at 1995 and imminent in 2008, Malay had a normally more than 25 million Internet users. It is foreseen that in starting of 2013, the amount of web users in Indonesian is getting the chance to be observably more noteworthy than in 2008. [6] also communicates that web customers in Indonesia is scramble definitely finished the latest two calendar years, “20 ratio of Indonesians 14 at another time and more settled now get to the Internet reliably. That is more than million people and growing reliably consistently. Regardless, we need to review two fundamental facts that depict the utilization. In the first place, around ten of the thirty million customers get to the Internet by methods for their cell phones. Secondly, around 70 ratios of those 30 million customers visit web Facebook, YouTube and twitter every month, creating in the most predominant location in the country” [7]. This reality truly isn't amazing considering the way that PC and its ability containing web as preface has been displayed since the adolescents in review school. It infers that the Indonesian youngsters particularly college understudies have expertise to get to web yet they are additionally possibly to abuse or to be abused by the web. Nearness of web for understudies in college really encourages them to get a considerable measure of data identified with their assignments. The data is given in types of book on the web and diaries. The two books and diaries give a simple assignment for the understudies to complete their undertakings specific when they lead their last paper to be graduated. In any case, a great deal of instances of abuse the web work additionally led by college understudies. Counterfeiting is a standout amongst the most web abuse led by the understudies. They tend to duplicate a few materials to their undertakings however they don't say the creator's name. Other web abuse can be discovered, for example, unlawful substance, online extortion, and wholesale fraud.

A. Classes of Strike

As indicated by National Research Council in the year 2003, there are 3 (three) classes of strike that directed to web, as following:

1) Organization unsettling influence: It causes lost organization and can come to fruition because of debilitating of frameworks through a grouping of attacks, for instance, repudiation of organization (DoS) and pummeling of information [8].

2) Theft of favorable circumstances: It mainly handles essential information on an adequately immense scale to have genuine impact.

3) Catch and control: It’s incorporates catching all controls of the web and apply this as an armament. These levels of strike or attacks are label as a cyber-crime and moreover have been balanced in various manners. That modus in certainty undermine every single individual exercise including framework.

B. Dealing with the Violations

To deal with and to keep those violations, digital security assumes vital part to ensure individuals to utilize web securely. As we known, the internet aggregates a gigantic scope of related components of the internet exercises and it is in this way the internet exercises are conceivably in danger. To take out or expel the hazard, insurance of the internet framework is required keeping in mind the end goal to stop programmers to perpetrate their violations. The assurance of the foundation must cover web equipment, media communications framework, processing gadgets as control framework and figuring gadgets as personal computer [9] moreover stipulates that to take out the hazard isn't just security to the foundation (equipment) yet in addition must ensure the product. Insurance of programming is proposed to help everyone to utilize PC/web securely. It is because such many PCs are utilized as a part of homes and organizations. The PC working frameworks and email programs are two parts of PC/web that is powerless against be assaulted and misused. Instance of PC worms that smacked MS Windows working structure in 2003 was a evidence to see that the prevention of write the computer programs is relied upon to guarantee the web customer; or other case of a worm happened in 2010, when a worm called Stuxnet was impelled to strike the Iran nuclear power plant agenda. [10].

Both protection of gear and writing computer programs are the essential reason for advanced security. They can surety people to use web soundly and safely. People will use web to help their activities with no agonize too pessimistic for impact of web. Regardless, the two preservations must be executed and introduced in domestic and widespread master plan or policy (control) to accomplish its targets. In United States, e.g., it can be found National Strategic plan for motherland protection. The inspirations driving this system are to check computerized ambushes against fundamental structure; to reduce domestic vulnerableness to advanced attack; and, to restrain the mischief and regaining duration from computerized attacks that do happen or another case in Canada, it’s domestic procedure is determined to 3 segments: protecting governing body of a nation structures; working together with the privately owned fragment; and serving its citizens to protect connected through care boost and raising [3].

Internet misuse is mostly committed by the younger generation their ages between 16 to 22 years. They are all belongs to different study’s institutes. They are use internet for financial fraud. These types of people hacked the other people accounts, ATM card PIN numbers, and important data. Social media is also a major and critical platform whose misuse by many illegal persons. The cyber terrorism activities are also occurred via this source in all over country. The misusers post the unethical posts on Facebook, WhatsApp, and Twitter, etc. For harmed the people. In forecast countries e.g., Indonesia, etc. has a law against cybercrimes but in Pakistan there is no any law against cybercrime introduced or implemented in country. That’s why the misuse of internet is increase day by day all over country. If any Law implemented in all over country against these kinds of people and Pakistan telecommunication authority, other agencies monitor the all activities and communication with help of cyber security techniques. Then we will overcome on this problem in very few time of period.

IV. Intranet and Multilevel Security Management

Many organizations are moving toward a private internet or personalized source of internet within the organization mostly called intranet. Intranet is an inside a vital piece of information system based on Internet knowledge, TCP/IP, HTTP transmission rules and website services. The intranet is an innovation that enables your association to characterize itself all in all element, a gathering and a family, where everybody knows their parts and everybody is chipping away at the change of the association.

As on the Internet, there will be specific goals and information and data on the intranet that ought to be kept puzzle from work drive, e.g. fund and helpful information. Measures are hence vital to ensure that these advantages are used and gotten to in a secured and orderly form. In a circumstance where there are just two sorts of assembled information, e.g. 'portrayed' and 'unclassified', access to these different sorts of information can be controlled by strategies for an ordinary access control procedure, e.g. a mystery word. If you know the mystery word, you will get to the more unstable information. Else, you may be allowed to see the non-sensitive information. Associations differentiate nevertheless, and military and government circumstances make the protected securing of information living on these intranets fundamentally more troublesome. The clarification behind this is the request of information as showed by its substance [11]. Out of the blue there aren't just two specific requests of data any more, however in the military condition for example, there are more game plans, i.e. Constrained, Classified, Secret and Top Secret. The number of groupings of data is settled only by the earth and can be practically than in the already said representation.

An intranet could comprise of a WAN (Wide Area Network) with many LAN (Local Area Networks) associated with it. Each of these LAN's alone can bolster an expansive number of workstations, servers, work area PC's, fringe gadgets and so on. Every one of these contraptions related with the intranet could have a security gathering. Heartbreakingly it's not just the devices that have a security gathering, yet the information they store or process as well. Every one of the information on a LAN can have the same or unmistakable arrangements. These elements befuddle security matters truly. Work power will in like manner have plans which will allow them access to information with orders equal or lower than their own specific gathering. This oversee is gotten from the Bell-Lapadula Model which is a famous security model. Clearly indicates for a relationship to empower most of its work power to use such an intranet direct is a critical troublesome task. If an affiliation is made to the web as well, the issue ends up being altogether more troublesome. Clearly the extended number of requests - i.e. Multi-Level Security - of data makes a charming issue in respects security in an intranet space. The change of modern and advanced Internet of Things has pulled in an impressive measure of enthusiasm from different research schools. The recognizing verification of people and things has permitted their portrayal in a propelled world through radio frequency distinguishing proof developments. This has enabled numerous applications to be delivered for key traceability and aerating and access control in various zones, for instance, transportation, mechanical or building [12].

V. Advances of Overseeing Multilevel Security

There are a few advances are utilized for overseeing multilevel security in military condition.

A. Onion Routing

Onion routing is a technique for unusual correspondence on a computer network. In an onion network, messages are summarized in coatings of encoding, closely resembling layers of an onion [13]. This scrambled information is communicated through a progression of system hubs also called onion switches, every one of which "peels" left an unsociable layer, revealing the information's next target. At the point when the past layer is decoded, the message lands at its target. The source stays unknown because every delegate identifies just the area of the promptly going before and following hubs.

B. Intelligent Agents

A keen specialist is an item that enables people and take after to up for their purpose. Savvy administrator's effort by empowering persons to name work that they could have finished, to the pro programming. Pros can accomplish dull errands, remember things you disregarded, keenly diagram complicated information, pick up from you and even influence proposal to you. To see how clever operators function, it is best to inspect a portion of the down to earth issues that shrewd specialist can help comprehend. A canny specialist can enable you to discover and channel info when you are viewing data or browsing the Internet and don't recognize where the correct information is. It could likewise re-try information to your slants, consequently save you time of dealing with it as additional latest information arrived every day on the Internet.

VI. Requirements for Network Security

Network security is the procedure through which we can ensure the computerized data. It is so urgent for all systems must be shielded from dangers and the dangers with the goal that a business can accomplish its fullest potential. With the advances in microelectronics, embedded processing, and wireless communications, the enthusiasm for Body Sensor Networks has risen pointedly and has empowered the improvement and usage of such systems There are no models or rules on estimating a scheme`s productivity. Security investigation is infrequently performed with formal strategies; rather, spellbinding examination is typical [14]. The target of system security is

1) To secure the secrecy the data must be gotten to and examined just by the endorsed individuals or social occasions. It is the protection of the individual information. We can differentiate order and security. Data encryption, User Ids and passwords, bio metric checks are a bit of the methodologies through which characterization can be secured.

2) To keep up Integrity it is the affirmation of not only the information can be gotten to or changes by the endorsed individuals presently moreover the data must be exact, unfaltering completed on the off chance that it can recall cycle. Measures taken to ensure respectability consolidate controlling the physical state of sorted out servers, limited access to data, and develop intensive check practices. Cryptography expect a to a great degree genuine part in ensuring the data uprightness. Hashing the data, you get and differentiating it and the hash of extraordinary message is another system to ensure data uprightness.

3) To ensure that the Availability of Data must be open to the endorsed individuals at the perfect time. It can be ensured by completely keeping up all hardware, preparing gear repairs rapidly and keeping up an adequately working system condition. Standard fortification must be taken, for information benefits that are exceedingly essential, abundance is fitting procedure to ensure availability.

VII. Complications in Network Security

Major Cyber Attacks and their counter measure as internet is facing number of security problems. That is the job of the network security to keep the system ensure against malevolent programming, worms, and dangers and different assaults. An assault is a data security risk through which the social criminal endeavor to get past, change, evacuate, embed or screen essential data without approved get right of section or consents.

A. Malware

Malware is a truncated term implying "malevolent programming". This is programming that is especially expected to get passageway or mischief a PC. Unmistakable sorts of malware are spyware, key loggers, honest to goodness contaminations, worms, or any sort of malignant code that mischief a PC. Overall, writing computer programs is recognizing malware in perspective of the desire of the creator as contradicted to its real highlights. Prior to the term malware was begat by [15], malignant programming was alluded to as PC infections. Malware is habitually familiar with a machine through email associations, programming program downloads or working structure vulnerabilities.

1) Counter measure for malware: The most ideal path against malware is to stop downloading contents from the untrusted internet sources and users just try to improve security. That is now and again accomplished through deploying strong and up to date firewalls, which prevent the unwanted traffic and it allows only the rusted traffic. While applying the different access control lists. It allows you to segregate the desired IP addresses which you want to allow in your Network. It must be checked out after rapid interval that operating system (for example Ubuntu, Windows, Mac OS X, centos and Linux) you use has the most updated security policy. Software developers update their software frequently to overcome the weak points for this purpose one should use up to date version of software. For example, the OS corporations like Microsoft, launch the updates for their OS often.

B. Phishing

Much of the time acting like an interest in data from a put stock in pariah, phishing attacks are sent by methods for email and demand that customers tap on an association and enter their own data. Phishing messages have turned out to be fundamentally more refined recently, making it troublesome for a couple of individuals to perceive a real interest for information from a false one. Phishing messages as often as possible fall into an undefined grouping from spam. yet are more destructive than only a straightforward promotion.

Research done by indicated that users effectively identified just 53% of phishing sites notwithstanding when prepared to recognize them and that they for the most part invest almost no energy looking at security markers contrasted with site content when making appraisals [16].

Phishing messages consolidate an association that aides the customer to a false site that will take a customer's information. From time to time, every one of a client desires to do is tap on the association. For example, in the past few years social media are the main and easy target of Pashing. Hackers create a fake page for social media site and the Victim think that it is original site, so he provides the credentials and get hacked easily Check any requirements from associations that touch base by means of email via telephone. If the email itself has a telephone number, don't call that number, however rather one you find unreservedly on the web or inside documentation you've gotten from that organization. User must check the URL of the accessing website before entering their credentials. The protocol must be checked that it is HTTP or HTTPS, HTTPS is more secure to give the personal information. Many companies warn users not to give any personal information over the Email or Telephone without verification. So, a user must double check before providing his sensitive information over the web.

C. Attacks on System via Password

The password is the most sensitive information a user could have over the internet. But most of the times user use the easy and simple words as their password which are easy to remember for him. For example, 1234, his mobile number, Birthdate, his own name or country name. Sometimes user set the same username and password for his convenience. The hacker tries to break your password using your public information.

This does not require any special algorithm or programming technique. It works just by entering Words which user could use as password.

To avert this, a user must use the complex password. His password must be at least 8 charters long. It should contain upper case, lower case letters, numbers and special characters. User should avoid using his personal pubic information like his own name as password. Do not use the words available in the dictionary. Password must be changes frequently.

D. Attacks on System via Denial-of-Service (DoS)

A DoS assault centers around entering the excessive information to break the security. To slow down the performance of a server or in some cases completely down the services, the hacker sends the fake requests in a large number of amount.

There are two or three different ways aggressors can achieve DoS assaults, however the most widely recognized is the circulated dissent of-advantage (DDoS) assault. This includes installing a third-party software to send the large number of fake requests to access the server. Due to this server stop to respond to the real or actual requests.

Using the advance level firewall could help to protect your system from such attacks. If a MAC address is broadcasting enormous number of requests in small time intelligent firewall will automatically block that traffic. Unless your association is massive, it's uncommon that you would be focused by an outside collecting or attacker for a DoS attack. Your site or system could even now surrender to one, be that as it may, if one more association on your organization is focused on. The most perfect method to keep an extra rupture is to keep your framework as secure as possible with normal program writing energizes, online safety detecting and examination your information stream to identify any bizarre or incapacitating points in rush hour jam before they turn into a problem. DoS attacks can similarly be performed by fundamentally cutting a link or removing a fitting that interfaces your site's server to the web, so due persistence in physically detecting your associations is suggested too. Simply cutting a cable or dislodging a plug that connects your website’s server to the internet, so due assiduousness in physically monitoring your networks is suggested as well.

E. “MAN in the Mid” (MITM)

By copying the endpoints in an online data exchange (For example the connection from your mobile phone to a web site) the MITM can gain your info from the end client and the module he or she is talking with. For occurrence, in case you are spending money on the different website, the man in the inside would conversation with you by impersonating your bank, and talk with the bank by rivalling you. The man in the inside would then get most of the info traded between the two gatherings, which could incorporate sensitive information, for example, financial balances and specific individual data.

Often, a MITM gains entry over a non-encoded remote entree point (For example one that doesn't utilize WAP, WPA, WPA2 or other protection efforts). They would then move toward most of the information being swaped among the two gatherings.

The greatest perfect method to counter them is to just apply encoded remote access attentions that Applying WPA & WPA2 safety or more prominent. On the off casual that you must interface with a site, ensure it uses a HTTPS association or, for excellent or effective security, consider inserting resources into a VPN (virtual private network). HTTPS uses validations that check the identity of the server machine you're interfacing with using an outsider association, for occurrence, VeriSign, while VPN (Virtual private network) allow you to link with areas through virtual private network.

F. Drive-By Downloads

Through malware on a true-blue site, a program is downloaded to a client's structure just by passing by the site. It doesn't need any type of motion by the customer to download.

Generally, a little part of code is downloaded to the client's outline and that code at that point links with one more PC to get the rest and download the package. It frequently manipulates vulnerabilities in the customer's working framework or in various plans, for example, Adobe and Java.

The most perfect path is to make certain the bigger portion of your working systems and programming agendas are flow of blood edge. This cuts down your danger of inadequacy. Furthermore, try to confine the amount of program additional things you use as these can be viably replaced off. For example, if your PCs needn't waste time with Flash or the Java component, consider uninstalling them.

G. Malvertising

A method to job off your PC with malicious code is downloaded to your outline when you tap on a partial announcement. Malvertising strategies are tormenting the web promoting business—offenders are procuring benefits by posting authentic commercials at content robbery destinations or utilizing a multitude of botnets to counterfeit publicizing movement [17].

Computerized aggressors exchange debased show advancements to different targets using an ad compose. These commercials are then passed on to goals that match definite amount watchwords and chase specification. Once a client taps on one of those advertisements, malware will have downloaded. Any webpage or web provider can be put through to malvertising, and several don't know they've been exchanged off.

The best way to deal with thwart surrendering to Malvern is to use sound to decide. Any improvement that confirms resources, free PCs or goes to the Bahamas is likely pipe dream, and thusly could be disguising malware. Of course, in the current style programmed and working structures are your best first line of protection.

H. Rogue Software

Malware that pretenses as genuine and essential safety programming will protect your framework.

Revolutionary safety programming fashioners make fly up windows and alerts that saw true blue. These alerts urge the customer to download safety programming, agree to terms or revive their existing system with an eventual objective to remain sure. By clicking "yes" to any of these conditions, the dissident writing computer programs is downloaded to the customer's PC.

The best protection is an average offense—for this circumstance, a revived firewall. Guarantee you have a working single in your workplace that protections you and your agents from these categories of attacks. It is also a savvy assumed to acquaint a put stock in antagonistic with contamination or against spyware programming program that can identify threat like free intellectual programming. Similarly, as with furthermost categories of bad actions, circumspection is one of the way to revolution. As computerized stranger ends up being more advanced and more trades relocate on the web, the quantity of risks to persons and federation will carry on developing. Prepare by own and its own trade put separately the chance to protect own systems and made computerized security a need. If you appreciate about some exceptional ways to agreement with keep on wary opposed computerized attacks, it's continuously best to start at home. At this time are eight methods to guarantee your organization's data is protected.

VIII. Conclusion

Internet has become a basic part of life in all over world. There are many advantages and disadvantages. The critical advantage is misuse of internet by criminal persons via unauthorized access and sources. People urges secure and protected platform who use internet. Cyber security gives the facility to internet users access/use the secure and protected source and way. Intranet has a private network used in government organizations and especially in military. It is more secure as compared to local internet but some drawbacks are occurred in this network. Then onion routing and intelligent agents control/protect the all over system by any threats and attacks. The security is the fundamental issue in the versatile specially appointed system [18]. In MANNET hub looks like self-centeredness. A hub can utilize the assets of other hub and safeguard the assets of possess. This sort of hub makes the issue in MANET there is various ways, which ensure for the wellbeing and security of your system [19]. Play out the accompanying to keep away from security provisos must have a refreshed antivirus program; try not to give progressively or undesirable access to any system client. Working framework ought to be consistently refreshed.

References

[1] Newman, S. (2017). Service providers: the gatekeepers of Internet security. Network Security, 2017, 5-7.

[2] Boshoff, W. H., & von Solms, S. H. (1989). A path context model for addressing security in potentially non-secure environments. Computers \& Security, 8, 417-425.

[3] Deibert, R. (2012). Distributed Security as Cyber Strategy: Outlining a Comprehensive Approach for Canada in Cyberspace. Journal of military and strategic studies, 14.

[4] Weber, R. H., & Studer, E. (2016). Cybersecurity in the Internet of Things: Legal aspects. Computer Law \& Security Review, 32, 715-728.

[5] de Argaez, E. (2011). Internet world stats: Usage and population statistics. Internet world stats.

[6] Suhariyanto, B. (2012). Information and Technology Crime (Cybercrime). Jakarta, PT. RajaGrafindo Persada.

[7] Tawar, M., & Keshari, V. (2013). The Impact of Information Technology on Work and Society. Pioneer Journal, 12, 00.

[8] Goldschlag, D. M., Reed, M. G., & Syverson, P. F. (1996). Hiding routing information. International Workshop on Information Hiding, (pp. 137-150).

[9] Andress, J., & Winterfeld, S. (2013). Cyber warfare: techniques, tactics and tools for security practitioners. Elsevier.

[10] Farwell, J. P., & Rohozinski, R. (2011). Stuxnet and the future of cyber war. Survival, 53, 23-40.

[11] Ghaffari, A. (2006). Vulnerability and security of mobile ad hoc networks. Proceedings of the 6th WSEAS international conference on simulation, modelling and optimization, (pp. 124-129).

[12] Zhu, N., & Zhao, H. (2017). IoT applications in the ecological industry chain from information security and smart city perspectives. Computers \& Electrical Engineering.

[13] Reed, M. G., Syverson, P. F., & Goldschlag, D. M. (1996). Proxies for anonymous routing. Computer Security Applications Conference, 1996., 12th Annual, (pp. 95-104).

[14] Kompara, M., & Hölbl, M. (2018). Survey on security in intra-body area network communication. Ad Hoc Networks, 70, 23-43.

[15] Boshoff, W. H., & Von Solms, S. H. (1990). Application of a path context approach to computer security fundamentals. Information Age, 12, 83-90.

[16] Alsharnouby, M., Alaca, F., & Chiasson, S. (2015). Why phishing still works: User strategies for combating phishing attacks. International Journal of Human-Computer Studies, 82, 69-82.

[17] Chaudhry, P. E. (2017). The looming shadow of illicit trade on the internet. Business Horizons, 60, 77-89.

[18] Li, W., & Joshi, A. (2008). Security issues in mobile ad hoc networks-a survey. Department of Computer Science and Electrical Engineering, University of Maryland, Baltimore County, 1-23.

[19] Sharma, K., Khandelwal, N., & Prabhakar, M. (2010). An overview of security problems in manet. Proceedings of the International Conference on Network Protocols (ICNP).

Data Mining: Web Data Mining Techniques, Tools and Algorithms: An Overview

Muhammd Jawad Hamid Mughal
Department of Computer Science
SZABIST Dubai Campus
Dubai, United Arab Emirates

***Abstract*—Web data mining became an easy and important platform for retrieval of useful information. Users prefer World Wide Web more to upload and download data. As increasing growth of data over the internet, it is getting difficult and time consuming for discovering informative knowledge and patterns. Digging knowledgeable and user queried information from unstructured and inconsistent data over the web is not an easy task to perform. Different mining techniques are used to fetch relevant information from web (hyperlinks, contents, web usage logs). Web data mining is a sub discipline of data mining which mainly deals with web. Web data mining is divided into three different types: web structure, web content and web usage mining. All these types use different techniques, tools, approaches, algorithms for discover information from huge bulks of data over the web.**

***Keywords*—Web data mining; hyperlinks; usage logs; contents; patterns**

I. Introduction

Now a day's data over the internet is enormous and increasing frequently day by day. It is must to manage that massive information and display most related queried information on user's screen. Analyzing and fetching relevant data from large data bases is not possible manually, for this automated extraction tools are required through which user queried data can be fetch from billions of pages over the internet and discovers relevant information. Usually users find data from world wild web WWW by using different search engines like Yahoo, Bing, MSN, Google etc. Data mining is a process of analyzing usable information and extract data from large data warehouses, involving different patterns, intelligent methods, algorithms and tools. This process can help business to analyze data, user behavior and predict future trends. Data mining includes four strategies steps for relevant data extraction. Data source is a set on data in large data base which can have problem definition in it. Data exploration is a step of investigation true information from bulks of unfamiliar data. Third step is modeling, in this different models are designed and then evaluate. At the end tested models are deployed, that occurs in final step of data mining strategies. Organizations can use data mining techniques to change raw data into convenient information. It can also help business to improve their marketing strategies and increase the profit by learning more about customer's behavior.

Web mining is one of the types of techniques use in data mining. The main purpose of web mining is to automatically extract information from the web. For discovering useful data (videos, tables, audio, images etc.) from the web different techniques and tools are used. Information over the internet is huge and increasing with passage to time due to which size of data bases are also growing. Digging knowledgeable information and analyzing the data sets for relevant data is much difficult because data over the internet in not in plain text. It could be unstructured data, multimedia, table, tag.

Purpose of this paper is to describe web mining, its three different types, tools and techniques. All three types are explained in detailed and main focus is on web usage mining, its techniques. Summarization table is detailed for all three types.

II. Literature Review

Data mining is a process of discovering knowledge from data warehouse. This knowledge can be classified in different rules and patterns that can help user/organization to analyze collective data and predicted decision processes [9]. Centralized database of any organization is known as Data warehouse, where all data is stored in a single huge database. Data mining is a method that is used by organization to get useful information from raw data. Software's are implemented to look for needed patterns in huge amount of data (data warehouse) that can help business to learn about their customers, predict behavior and improve marketing strategies.

Web mining is actually an area of data mining related to the information available on internet. It is a concept of extracting informative data available on web pages over the internet [1]. Users use different search engines to fetch their required data from the internet, that informative and user needed data is discovered through mining technique called Web Mining. Different tools and algorithms are used for extraction of data from web pages that includes web documents, images etc. Web mining is rapidly becoming very important due to size of text documents increasing over the internet and finding relevant patterns, knowledge and informative data is very hard and time consuming if it is done manually. Structure (Hyperlinks), Usage (visited pages, data use), content (text document, pages) are included in information gathered through Web mining [2], [5]. Term World Wide Web is related to the combination of web documents, videos, audios etc. Some processes included in web mining are:

Information Retrieval is a process of retrieving relevant and useful information over the web. Information retrieval has more focuses on selection of relevant data from large collection of database and discovering new knowledge from large quantity of data to response user query.IR steps includes searching, filtering and matching [5], [6].

Information extraction is an automatic process of extracting analyzed data (structured). IE is a task that work same like information retrieval but more focuses on extracting relevant facts [5].

Machine Learning is support process that helps in mining data from web. Machine learning can improve the web search by knowing user behavior (interest). Different machine learning methods are used in search engine to provide intelligent web service. It is much more efficient than traditional approach i.e. information retrieval. It is a process that has ability to learn user behavior and enhance the performance on specific task.

III. Web Mining Categories

Web Mining is sub categorized in to three types as shown in Fig. 1:

A. Web Content Mining

B. Web Structure Mining

C. Web Usage Mining

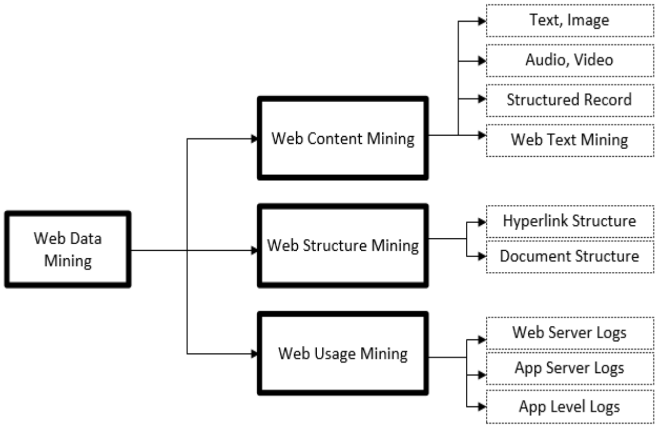


Fig. 1. Web mining taxonomy [8], [15].

Web Mining consists of massive, dynamic, diverse and mostly unstructured data that provides big amount of data. Explosive growth of web leads to some problems like finding relevant data over the internet, observing user behavior. To solve such kind of problem efforts were made to provide relevant data in structure form (table) that is easy to understand and useful for organizations to predict customer's needs [4].

A. Web Content Mining

Content Mining is a process of Web Mining in which needful informative data is extracted from web sites (WWW). Content includes audio, video, text documents, hyperlinks and structured record [1]. Web contents are designed to deliver data to users in the form of text, list, images, videos and tables. Over last few decades the amount of web pages (HTML) increases to billions and still continues to grow. Searching query into billions of web documents is very difficult and time consuming task, content mining extracts queried data by performing different mining techniques and narrow down the search data which become easy to find required user data [3].

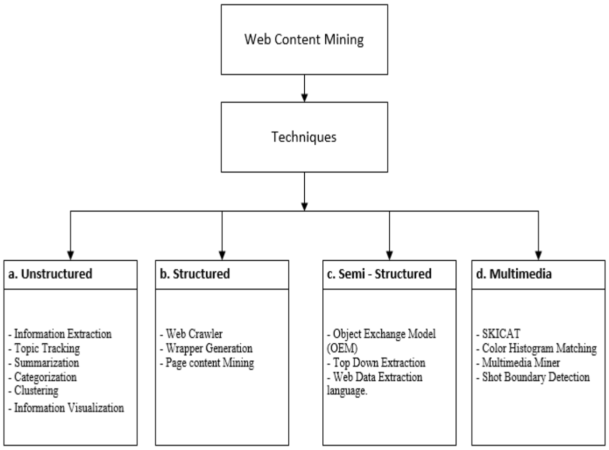


Fig. 2. Web content techniques [7].

1) Web Content Mining Techniques: Web content mining uses different techniques Fig. 2 to dig data. Following are four techniques described used by web content mining.

Mostly in web contents data is in unstructured text form. For extraction of unstructured data, web content mining requires text mining and data mining approaches [5]. Text documents are related to text mining, machine learning and natural language. Main purpose of text mining is to extract previous information from content source [7]. Text mining is a part of web content mining and hence different techniques are used for text data mining from web contents over the internet/website to provide unknown data, some of them are mentioned below:

- Information Extraction
- Summarization
- Information Visualization
- Topic Tracking
- Categorization
- Clustering

Structured is a technique that mines structured data on the web. Structure data mining is an important technique because it represents the host page on the web. Compare to unstructured, in structured data mining it is always easy to extract data [8]. Following are some techniques used for structured data mining:

- Web Crawler
- Page Content Mining
- Wrapper Generation

Semi-Structured is form of structured data but not full, text in semi-structured data is grammatical. Its structure is Hierarchical, not predefined. Representation of semi-structured data is in form of tags (such as HTML, XML). HTML is an intra-document structure case [4]. Techniques used to extract semi-structured data are:

- OEM - Object Exchange Model
- Web Data Extraction Language
- Top Down Extraction [5]

Traditionally computing data was consider as text and numbers but now a days there are different computing data types of multimedia data like videos, images, audios etc. This mining process is use for extracting interesting multimedia data sets and also converted data set types in to digital media [11]. Techniques uses for multimedia data mining are:

- Multimedia Miner
- Shot Boundary Detection
- SKICAT
- Color Histogram Matching [10]

2) Web Content Mining Algorithms: Multiple techniques are used by web mining to extract information from huge amount of data bases. There are different types of algorithms that are used to fetch knowledge information, below are some classification algorithms are described:

Decision tress is a classification and structured based approach which consist of root node, branches and leaf nodes. It is hierarchical process in which root node is split into sub branches and leaf node contains class label. Decision tress is a powerful technique [10].

Naïve Bayes is an easy, simple, powerful algorithm for classification and also known as Native Bayes classifier. It is based on Bayes' Theorem. From predefined dataset values, probabilities are calculated for each class by counting combinations on values. Most likely class is the one with highest probability [12].

Bayes' theorem: [13]

$$P(A \mid B) = \frac{P(B|A)P(A)}{P(B)}$$

Support Vector Machine is a well-known and simple machine learning and classification algorithm. SVM is a method that can be used for linear and non-linear data sets [10]. Optimal separating hyper plane (decision boundary) is just a line that is used to draw to separate the two classes depends on the different classification features.

Neural network is another web content mining approach which use back propagation algorithm. The algorithm consist of multiple layers i.e. input layer, some hidden layers and then output layer, each feeds the next layer till last layer (output). Neuron is the basic unit of neural network. Inputs are fed simultaneously to units. From input layer, inputs are simultaneously feeding to hidden layers. Usually there is one hidden layer but numbers of hidden layers are arbitrary [10]. Last hidden layer fed the input and make up the output layer.

B. Web Structure Mining

Now a day's massive amount of data is increasing on web. World Wide Web is one of the most loved resources for information retrieval. Web mining techniques are very useful to discover knowledgeable data from web. Structure mining is one of the core techniques of web mining which deals with hyperlinks structure [14]. Structure mining basically shows the structured summary of the website. It identifies relationship between linked web pages of websites. Continues growth of data over the internet become a challenging task to find informative and required data [15]. Web mining is just a data mining which digs data from the web. Different algorithmic techniques are used to discover data from web. Structure mining analyzes hyperlinks of the website to collect informative data and sort out in categories like similarities and relationship. Intra-page is a type of mining that is performed at document level and at hyperlink level mining is known as inter-page mining. Link analysis is an old but very useful method that is way its value increases in the research area of web mining – Structure analysis is also called as Link-mining [16]. Few of the tasks of link-mining Fig. 3 are summarized as:

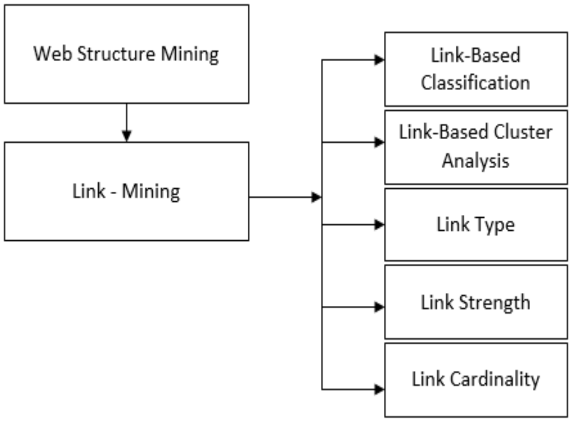


Fig. 3. Web structure mining.

- ***Link-based Classification:*** It is an upgrade classification version of classic data mining and its task is to link domains. Main focus is to predict webpage categories – based on text, HTML tags, link between web pages and other attributes [15].
- ***Link-based Cluster Analysis:*** Primary focus is on data segmentation. In cluster analysis data is categorized or grouped together [16]. Similar objects are grouped in a single group and dissimilar data objects are grouped separately. To dig hidden patterns from datasets link-based cluster analysis can be used [15].
- ***Link Type:*** It helps to guess link type between entities (two or more) [16].
- ***Link Strength:*** Link strength shows that links might be related to weights [16].

- ***Link Cardinality:*** Main focus of link cardinality is to find duplicated website, finding comparison between them, predicts links between objects, also page categorization [15].

3) Web Structure Mining Algorithms: There are various web structure mining algorithms as mentioned in Table I, the paper describes two of them i.e. Page rank algorithm and HITS algorithms. Both of them focuses on link structure of web and how it gives importance to web pages.

Page rank algorithm was developed in 1998 [16] by two famous authors L. Page and S. Brain. The idea was proposed in their PhD research. Both the authors suggested that well known search engine Google was formed by page rank algorithm. It is an algorithm that is frequently used to rank pages. Page rank approach leads to number of pages linking to a specific web page indicates, calculates or describes the importance of that page. Above calculated links are known as backlinks. If backlink is produced from key page or an important page then weightage of this link will be higher than those whose links are coming from non-important pages. Link from page A to page D is considered as a vote (Shown in Fig. 4: Back link Structure). More the vote receives by the page more the importance of that specific page will be. If vote produced from a high weightage page then the importance of linking page will become higher.

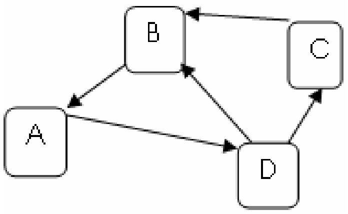


Fig. 4. Back link structure [16].

Following is the formula [14] to find page rank of page A:

$$PR(A) = (1 - d) + \frac{d(PR(T1))}{C(T1)} + \cdots + \frac{PR(Tn)}{C(Tn)(1)}$$

Where:

PR (Ti) = Rank of Pages

Ti = links to A

C (Ti) = No. of outbound links

d = damping factor (0 to 1)

HITS is an algorithm that stands for Hyperlink Induced Topic Search and is use for web structure (hyperlink analysis) mining. HITS concept was developed by Jon Kleinberg [16] to rank pages. Two terminologies are used in HITS algorithm i.e. authorities and hubs. Good authority is a page that is pointed by high hub weights and good hubs are pages that points to many authority pages with high weights Fig. 5. It is not easy to differentiate in between these two attributes as some sites can be hubs as well as authorities at the same time.

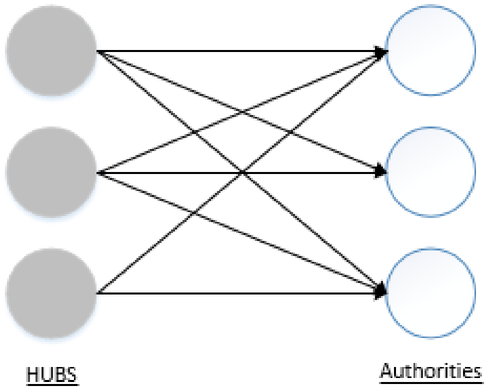


Fig. 5. HITS (Hubs and Authorities) [17].

HITS algorithm includes two steps. First is sampling in which related pages are collected for certain queries. In iterative step authorities and hubs are found with the help of sampling output. Because of the equal weights of pages HITS don't find the relevant pages requested by user queries [17].

4) Web Structure Mining Tools: Following mention tools are used for web structure mining. Google PR checker is a tool designed for page rank and is used for Page rank algorithm. It helps to rank of web pages in search engine result. It is simple to find page rank just by pasting website URL and click search – it will show rank of each page of website [22]. Link viewer is used for HITS to visualize analyzes links process [23].

- Google PR Checker (for PageRank)
- Lin Viewer (for HITS)
- Web Mining Categories Summarization

5) Web Usage Mining: Web usage mining also called log mining is a process of recording user access data on the web and collect data in form of logs. After visiting any website user leaves some information behind such as visiting time, IP address, visited pages etc. This information is collected, analyzed and store in logs. Which helps to understand user behavior and later can improves website structure [18]. Web usage mining is a technique that automatically archives access patterns of user and this information is mostly provided by web servers which are later collected in access logs. Logs stores much needed information like URL address, visiting time, Internet Protocol addresses etc. which can help an organization to understand their customer's behavior and insure good service quality. Web usage mining dig and analyze data present in log files which contains user access patterns. Main purpose of web usage mining is to observer user behavior at the time of his interacting with web. There are two types of pattern tracking i.e. general tracking and customized tracking. In general tracking information is collected from web page history. In customized tracking the information is gathered for specific user [19].

TABLE I. SUMMARIZATION TABLE FOR WEB DATA MINING CATEGORIES

Web Mining Categories	**Techniques**	**Tools**	**Algorithms**
Web Content Mining	- Unstructured Data Mining - Structured Data Mining - Semi – Structure Data Mining - Multimedia Data Mining	- Screen Scaper - Mozenda - Automation Anywhere7 - Web Content Extractor - Web Info Extractor - Rapid Miner	- Decision Tree - Naive Bayes - Support Vector Machine - Neural Network
Web Structure Mining	- Link-based Classification - Link-based Cluster Analysis - Link Type - Link Strength - Link Cardinality	- Google PR Checker - Link Viewer	- Page Rank Algorithm - HITS algorithms (Hyperlink Induced Topic Search) - Weighted Page Rank Algorithm - Distance Rank Algorithm - Weighted Page Content Rank Algorithm - Webpage Ranking Using Link Attributes - Eigen Rumor Algorithm - Time Rank Algorithm -Tag Rank Algorithm - Query Dependent Ranking Algorithm
Web Usage Mining	- **Data Preprocessing** • Data Cleaning • User & Session Identification - **Pattern Discovery** • Statistical Analysis • Association Rules • Clustering • Classification • Sequential Patterns - **Pattern Analysis** • Knowledge Query Mechanism • OLAP (Online Analytical processing) • Intelligent Agents	- **Data Preprocessing Tools** • Data Preparator • Sumatra TT • Lisp Miner • SpeedTracer - **Pattern Discovery Tools** • SEWEBAR-CMS • i-Miner • Argunaut • MiDas(Mining In-ternet Data for As-sociative Sequenc-es) - **Pattern Analysis Tools** • Webalizer • Naviz • WebViz • WebMiner • Stratdyn	- **Association Rules** • Apriori Algorithm • Maxi-mal Forward References • Markov Chains • FP Growth • Prefix Span - **Clustering** • Self-Organized Maps • Graph Partitioning • Ant Based Technique • K-means with Genetic algorithms • Fuzzy c-mean Algorithm - **Classification** • Decision Trees • Naïve Bayesian Classifiers • K-nearest Neighbor Classifiers • Support Vector Machine - **Sequential Patterns** • MIDAS (Mining Internet Data for Association Sequences) algorithm

6) Web Usage Mining Techniques: Following three techniques are described in detail with their sub approaches use in web usage mining. Each technique performs different tasks in a hierarchy.

- Data Preprocessing

Real world data and some data bases are incomplete, inconsistent and not understandable. Data preprocessing is a mining technique that integrate databases and make raw data understandable and consistent [18]. In data preprocessing information stored web logs are processed because of insufficient and noisy nature. Raw data cleaning is done is done in early step by removing redundant, useless, error, incomplete, inconsistent data [19]. Preprocessing task is to clean, correct the data and ready input data for mining. There are many e-sources in web usage mining from data can be collected and analyze such as data logs, website, users login information, web access logs, cache, cookies etc. The reliable source for usage mining is considered as web access logs because they use standard logs format (Common LF and Extended CLF) for recording [20]. Data preprocessing includes methods like Data cleaning, User and session identification are describe as follow.

Data cleaning is not only important for usage mining but important for other analysis techniques as well. Purpose is to remove irrelevant and no needed information from logs. Graphics and videos needs to be removed from web logs as they are unnecessary for usage mining [21]. When user requests for a web server for a particular web page, multiple entries are stored in log file. Those records that are not useful for usage mining must be removed.

User and Session identification technique is used to find user sessions from access log file. After data cleaning next step is to identify users. Different approaches are used for user identification like user login information, cookies to detect visitors with unique ID for specific webpage. Session identification is to know number of pages visited by a single user in a row on one visit to a website. Session is a set of webpage visited by users, new IP mean new user. Difficult step is when proxy server is used, same IP addresses for different users in log file. Referrer method is suggested as a solution to this problem. As different IP indicates new users, if IP's are same then different browsers / OS can identify new users. If OS, IP and browsers are same then Referrer approach consider URL account information. If account in URL was not accessed before it will consider it as a new user [18].

- Pattern Discovery

Consider as key component in web data mining. After data cleaning and user identification, some web usage pattern discovery techniques are used to discover interesting patterns. Main and tough task is to discover patterns produced by preprocessing section and extract useful knowledge [19]. Pattern discovery techniques are describe as follow.

Statistical analysis is a powerful technique used for extracting knowledge about webpage visitors. Analysts perform to describe statistical analysis on session log while analyzing using different variables. Knowledge obtained by statistical analyzing result may help to improve performance and enhance the system security as well as marketing strategies [24]. Frequency, median and mode are three statistical analyses are used mostly on sessions to show length of page, recently accessed pages and view time [19].

Association rule is one of the basic rules of data mining and is mostly use in web usage mining. Association rule helps to find correlations between webpages that appears in a user session repeatedly. The rule describes the relationship between pages visited one after another by user at the time of his visit session. The rule X => Y (where X and Y are pages) state that items (transaction) includes in page X also contain in page Y [26]. Rule format can be shown as:

X.html, Y.html => Z.html

It means that if user will observe page X and Y, most probable he will also observe page Z in the same session.

Clustering is a method of grouping items (users and pages) with similar features together. Usage mining consist of two types of clusters i.e. users and pages cluster. Users cluster provides information about set of users with a similar activities or browsing patterns [25]. Similar webpage content can be discovered from pages clusters. Different algorithms are used for clustering technique as shown in Table I.

Classification technique is use to classify data items and map them to different predefined classes. In usage mining, one with an interest of generating user profile will use this technique to establish user profile of user fitting to particular class [24]. Classification can be performed by use of different algorithms as mentioned in Table I.

Sequential sessions are discovered in sequential patterns. Many algorithms are used to find sequential patterns in usage mining, some of them are listed in Table I. MIDAS is commonly used algorithm for finding sequential sessions [19]. This technique catches patterns like one or multiple bulks of pages visited or accessed one after another in same time sequence. It is helpful for web admin/marketer to predict trends and prepare advertisements, place them to target group of users [25].

- Pattern Analysis

Pattern analysis is considered as a last and final step of usage mining. In this step all not interesting, irrelevant rules or patterns discovered in above phases are separated and interesting or relevant rules or patterns are extracted. This can help to improve system performance [24]. Following are the approaches uses for pattern analysis:

For query mechanism the most commonly language used is SQL. SQL stands for Structured Query Language and is use to extract useful information from patterns discovered [18].

After the pattern discovery data is receive into OLAP phase. In this phase data is store in data cube (multi-dimensional database) format and OLAP operation (roll up etc.) is performed. In OLAP measure term refers to dimensions (tables) [27].

An agent can be defining as an assistant that can help to perform some tasks on user's behalf. An intelligent agent can sense receiving element, recognizer them and determines which task should be performed. In usage mining agents analysis the pattern that are discovered at previously phases [18], [28].

7) Web Usage Mining Algorithms: For usage mining there are numbers of algorithms that can be used as few of them are listed in above Table I. This section will describe three important algorithms i.e. Apriori Algorithm, FP Growth Algorithm and Fuzzy c-means algorithm.

Apriori algorithm is an important and supervised algorithm mostly use for association rule (describe above) to find frequent sets of items during transaction. At first apriori algorithm observe initial database and captures those data sets which are large, then uses result of first captured data sets as a base or model to discover other data sets (large). In apriori algorithm there is a pre-defined support level, if the support level is greater than minimum then item sets are called large or frequent and if support level is below then item sets are known as small. Before AIS algorithm was used for mining regular item sets and association rules but after some time algorithm was modified and given a name Apriori Algorithm [31]. Example: Suppose we have two transactions $A1 = \{1, 2, 3\}$ and $A2 = \{2, 3, 4\}$ where 1,2,3,4 are item sets and 2, 3 are frequent items in both transactions because of repetition.

FP growth is another efficient algorithm use for association rule. FP-Growth discovers frequent sets of data from FP tree without candidate generation and use bottom-up approach. FP tree is complete data structure, contains one root node "null" and sub tree nodes (prefix) as children. FP growth search FP tree and fetch frequent sets of data [31].

Both Apriori and FP growth algorithms are suitable and scalable for association rule but FP growth is considered more efficient that Apriori algorithm but in Apriori full database needs to be scan for frequent sets of data where as in FP growth, FP tree is made and new sets are updated while transaction.

Fuzzy cMean is an algorithm use in usage mining using clustering approach. It was developed by Bezdek [29]. Fuzzy in an unsupervised algorithm that is applied to a wide range to connected data. FCM task is to group n number objects n number of clusters. In every cluster there is center point which describes features and importance of that cluster [30]. Objects close to the center of cluster become member of the cluster.

FCM Algorithm formula: [29], [30]

$$J(U, c_1, \dots c_c) = \sum_{i=1}^{C} Ji = \sum_{i=1}^{C} \sum_{j=1}^{n} u_{ij}^{m} d_{ij}^{2}$$

Where: Ci = Cluster Center

Uij = Numerical Value [0, 1]

i = Euclidian Distance = $d_{ij} = \| c_i - x_j \|$

ith , jth = Cluster Center , data points

8) Web Usage Mining Tools: Speed Tracer is an analysis tool and use for usage mining. This tool help to discover users surfing behavior and analyze with entries stored in server logs by using different data mining techniques. Cookies are not required for identifying user session, speed tracer used different kind for information like: IP address, URL of page, agent etc. Collection of browsing patterns helps to understand user behavior in a better way [18]. Three types of understanding are generated by speed tracer: User based which refers to user access time duration. Path based relates to process of frequently visited path in web. Group based generates information of repeatedly visited groups of web pages.

Suggest 3.0 is a system that provides familiar information to user about web pages they might have interest in. Customers/user needs are successfully achieved by set of constant changes in page links. Suggest 3.0 uses graph partitioning algorithm for historical information to be maintained. Main purpose is to keep record of incremental change [18].

WebViz is a tool that is used for statistical analysis of web access logs. The main idea of developing this tool was to provide WWW database designers graphical outlook of their local db and access patterns. Relation between access logs and databases (local) is displayed by use of Web-Path paradigm [32]. It presents local database documents and association of documents in graph structure. Information about accessed documents is collected from access log. Number of visited paths by users is also collected for display.

9) Web Usage Mining Techinques Comparison

IV. Conclusion

Data mining is a concept that helps to find information which is needed from large data warehouses by using different techniques. It is also used to analyze past data and improve future strategies. Web data mining is considered as sub approach of data mining that focuses on gathering information from web. Web is a large domain that contains data in various forms i.e.: images, tables, text, videos, etc. As size of web is continuously increasing; it is becoming very challenging task to extract information. In this paper we described three important types of web data mining that can help in finding informative data. Each type has different algorithms, tools and techniques that are used for data retrieval. Various algorithms, tools and techniques for each type are described. Table I summarizes all types and Table II shows comparison for web usage mining techniques. Web content mining is useful in terms of exploring data from text, table, images etc. Web structure mining classifies relationships between linked web pages. Web usage mining is also an important type that stores user access data and get information about specific user from logs. All techniques may have some advantages and disadvantages but drawbacks can be improved by further studies.

TABLE II. Usage Mining Techniques Comparison

Usage Mining Techniques	**Methods**	**Data Gathering**	**Data Store**	**Advantages**	**Important Algorithms**
Data Preprocessing	- Web status codes	- Data logs - Website - Users login information - Web access logs - Cache - Cookies etc.	- Web logs	- Convert raw data to understandable Common LF and Extended CLF for recording	- Apriori algorithm - FP Growth
Pattern Discovery	- Frequency, median, mode used to show length, recently accessed, view time of pages	- Filtered data from preprocessing section	- Session logs	- Extract useful information from discovered patterns correlations	- K-means with Genetic algorithms - Fuzzy c-mean Algorithm
Pattern Analysis	- Roll-up - Drill Down/Up	- Pattern discovery	- Data cube (multi-dimensional database)	- Irrelevant rules and patterns are separated	- SQL Language - OLAP

REFERENCES

[1] Anurag Kumar and Ravi Kumar Singh, "Web Mining Overview, Techniques, Tools and Applications: A Survey," International Research Journal of Engineering and Technology (IRJET), vol. 03, no. 12, pp. 1543-1547, December 2016.

[2] Simranjeet Kaur and Kiranbir Kaur, "Web Mining and Data Mining: A Comparative Approach," *International Journal of Novel Research in Computer Science and Software Engineering*, vol. 2, no. 1, pp. 36-42, January - April 2015.

[3] Ahmad Tasnim Siddiqui and Sultan Aljahdali, "Web Mining Techniques in E-Commerce Applications," International Journal of Computer Applications, vol. 69– No.8, pp. 39-43, May 2013.

[4] Kshitija Pol, Nita Patil, Shreya Patankar, and Chhaya Das, "A Survey on Web Content Mining and extraction of Structured and Semistructured data," Emerging Trends in Engineering and Technology, pp. 543-546, July 2008.

[5] R. Malarvizhi and K Saraswathi, "Web Content Mining Techniques Tools & Algorithms – A Comprehensive Study," International Journal of Computer Trends and Technology (IJCTT), vol. 4, no. 8, pp. 2940-2945, Augest 2013.

[6] Raymond Kosala and Hendrik Blockeel, "Web Mining Research: A Survey," SIGKDD Explorations, vol. 2, no. 1, pp. 1-15, July 2000.

[7] Faustina Johnson and Kumar Santosh Gupta, "Web Content Mining Techniques: A Survey," International Journal of Computer Applications (0975 – 888), vol. Volume 47– No.11, pp. 44-50, June 2012.

[8] Abdelhakim Herrouz, Chabane Khentout, and Mahieddine Djoudi, "Overview of Web Content Mining Tools," The International Journal of Engineering And Science (IJES), vol. 2, no. 6, June 2013.

[9] Claus Pahl and Dave Donnellan, "Data Mining Technology for the Evaluation of Web-based Teaching and Learning Systems," 7th Int. Conference on E-Learning in Business, Government and Higher Education, October 2002.

[10] Anurag kumar and Kumar Ravi Singh, "A Study on Web Content Mining," International Journal Of Engineering And Computer Science, vol. 6, no. 1, pp. 20003-20006, January 2017.

[11] Dr. S. Vijayarani and Ms. A. Sakila, "MULTIMEDIA MINING RESEARCH – AN OVERVIEW," International Journal of Computer Graphics & Animation (IJCGA), vol. 5, pp. 69-77, January 2015.

[12] Tina R. Patil and Mrs. S. S. Sherekar, "16. Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification," International Journal Of Computer Science And Applications, vol. 6, pp. 256-261, April 2013.

[13] M. Bilal, P. M. L. Chan, and W. Khan, "Cooperative Network for Vehicular Communications: Game Theoretic Distribution of Reward among Contributing Vehicles," Cyber Journals: Multidisciplinary Journals in Science and Technology, Journal of Selected Areas in Telecommunications (JSAT), vol. 3, no. 8, pp. 11-25, Augest 2013.

[14] Miguel Gomes da Costa Júnior and Zhiguo Gong, "Web Structure Mining: An Introduction," International Conference on Information Acquisition, pp. 590-595, June 27 - July 3 2005.

[15] Anurag Kumar and Kumar Ravi Singh, "A Study on Web Structure Mining," International Research Journal of Engineering and Technology (IRJET), vol. 04, no. 1, pp. 715-720, January 2017.

[16] B. L. Shivakumar and T. Mylsami, "SURVEY ON WEB STRUCTURE MINING," ARPN Journal of Engineering and Applied Sciences, vol. 9, pp. 1914-1923, October 2014.

[17] Monica Sehgal, "Analysis of Link Algorithms for Web Mining," International Journal of Scientific and Research Publications, vol. 4, no. 5, May 2014.

[18] Pranit Bari and P.M. Chawan, "Web Usage Mining," Journal of Engineering, Computers & Applied Sciences (JEC&AS), vol. 2, pp. 34-38, June 2013.

[19] Kamika Chaudhary and Santosh Kumar Gupta, "Web Usage Mining Tools & Techniques: A Survey," International Journal of Scientific & Engineering Research, vol. 4, no. 6, pp. 1762-1768, June 2013.

[20] Saša Bošnjak, Mirjana Marić, and Zita Bošnjak, "The Role of Web Usage Mining in Web Applications Evaluation," Management Information Systems, vol. 5, October 2009.

[21] Prabha.K and Suganya.T, "A Guesstimate on Web Usage Mining Algorithms and Techniques," International Journals of Advanced Research in Computer Science and Software Engineering, vol. 7, no. 6, pp. 518-521, June 2017.

[22] Liupu Wang et al., "Using Internet Search Engines to Obtain Medical Information: A Comparative Study," Journal of Medical Internet Research, May 2012.

[23] Saeko Nomura, Satoshi Oyama, Tetsuo Hayamizu, and Toru Ishida, "Analysis and Improvement of HITS Algorithm for DetectingWeb Communities," Applications and the Internet, February 2002.

[24] Yan Wang, Web Mining and Knowledge Discovery of Usage Patterns., February 2000.

[25] Parth Suthar and Prof. Bhavesh Oza, "A Survey of Web Usage Mining Techniques," (IJCSIT) International Journal of Computer Science and Information Technologies, vol. 6, pp. 5073-5076, 2015.

[26] Nasrin JOKAR, Reza Ali HONARVAR, Shima AgHAMIRZADEH, and Khadijeh ESFANDIARI, "Web mining and Web usage mining techniques," Bulletin de la Société des Sciences de Liège, vol. 85, pp. 321 - 328, 2016.

[27] Surajit Chaudhuri and Umeshwar Dayal, "An Overview of Data Warehousing and OLAP Technology," ACM SIGMOD, vol. 26, no. 1, pp. 65-74, March 1997.

[28] Ayse Yasemin SEYDIM, INTELLIGENT AGENTS: A DATA MINING PERSPECTIVE. Dallas, May 1999.

[29] Ajith Abraham, "BUSINESS INTELLIGENCE FROM WEB USAGE MINING," Journal of Information & Knowledge Management, vol. 2, no. 4, December 2003.

[30] M.SANTHANAKUMAR and C.CHRISTOPHER COLUMBUS, "Web Usage Based Analysis of Web Pages Using RapidMiner," WSEAS TRANSACTIONS on COMPUTERS, vol. 14, pp. 455-464, 2015.

[31] Aanum Shaikh, "Web Usage Mining Using Apriori and FP Growth Algorithm," (IJCSIT) International Journal of Computer Science and Information Technologies, vol. 6, pp. 354-357, 2015.

[32] James E. Pitkow and Krishna A. Bharat, "WEBVIZ: A TOOL FOR WORLD-WIDE WEB ACCESS LOG ANALYSIS," In Proceedings of the First International WWW Conference, January 1994.

Power Management of a Stand-Alone Hybrid (Wind/Solar/Battery) Energy System: An Experimental Investigation

Saindad Mahesar,
Mazhar H. Baloch
Mehran University of Engg & Tech,
Sindh Pakistan

Ghulam S. Kaloi
QUEST Larkana,
Campus Sindh Pakistan

Mahesh Kumar,
Aamir M. Soomro,
Asif A. Solangi, Yasir A. Memon
Mehran University of Engg & Tech,
Sindh Pakistan

Abstract—**In this manuscript, a hybrid wind/solar/battery energy system is proposed for a stand-alone applications. Wind-solar energy sources are used as power generation source in the proposed hybrid energy system (HES), whereas battery is used as energy storing system in order to manage the power flow among various power generation sources and energy storing system. Power management control strategy is also presented for a suggested hybrid system. Through the real load demand and practical weather data (proposed area is Jamshoro, Sindh Pakistan), the system performance is verified under different situations. It is observed that the hybrid system produces maximum power in summer season as compared to other seasons throughout the year. Moreover, the power generated from wind and solar energy contributes 77.88% and 22.12%, respectively. However, it is clearly observed that the HES is cost effective and can be used in remotely rural areas which are isolated from power grid. In future work, the HES can be integrated with the power grid in order to meet the load demand during shortage of power.**

Keywords—Hybrid; stand-alone; wind; solar; battery; power management; Pakistan

I. Introduction

The growing energy utilization, expanding environmental contamination, high rate and the prompt decline of fossil fuels have made increasing concern towards renewable energy generation sources, i.e., wind energy, solar energy, geothermal, biomass, and tidal energy [1]-[5, [18]. In Pakistan, the renewable energy sources have a huge potential to produce electric power and these resources can play a vital role in overcoming energy crisis [6], [7]. Among these renewable energy resources, solar and wind energy sources are more attractive power producing technologies [1], [2], [8]. The wind and solar energy are abundantly available, free, inexhaustible, and have no emission of greenhouse gases, therefore they are atmosphere friendly [3], [5], [6], [10]. Due to such features, these renewable energy resources can be utilized on large scale to produce power in order to meet load requirement [6]. A huge amount of population are living at rural areas around the globe, where still no power for their social life, and they have very minimum load demand, and still not connected with the power grid system. In order to supply the power to these remotely located areas, alternative energy sources such as wind and solar are economical and efficient substitutes [6].

Globally, wind and solar energy resources have their own deficiencies, such as wind is not accessible at all times and solar is accessible only day time, therefore in order to make sure the continuity of power supply to meet the load requirements, it is essential to make wind and solar energy resources along with storage device as a hybrid energy system [9]. The energy storing devices may be battery, flywheel, super capacitor or fuel cell [1], [4], [10]. Hybrid energy system is the integration of two or more than two power producing technologies together with some energy storage system to supply power to the load. Various unconventional energy sources i.e., wind, solar, diesel, fuel cell, gas turbine can be utilized to make a hybrid system [1]. A hybrid energy system is more efficient and provides continuous power to consumers with more reliability than a single source based system [1], [2]. The hybrid energy system is more appropriate for off-grid services to supply remotely located rural regions [10].

In this paper, a hybrid wind/solar/battery energy system is suggested for a stand-alone applications. Solar-wind system is primarily resources of generation in our proposed system, while battery is used as energy storage system and these sources are coupled with AC bus via appropriate converters. The DC load can be supplied through ac/dc rectification [15].

This paper is ordered as follows: in section 2, hybrid system configuration and description is given. In section 3, modeling and overall power management strategy is given. In section 4, result and discussions is given. While the last section 5 concludes the manuscript.

II. Hybrid System Configuration and Description

A. Hybrid System Configuration

The proposed configuration of the hybrid i.e., wind/solar/battery system can be shown in Fig. 1. In this system, renewable wind and solar resources are utilized as major energy sources, whereas battery is used as storage device. When the power generation from wind and solar system is excessive, the excessive power is delivered to the battery. When battery becomes fully charged, then extra power will be sent to dump load. While during the deficiencies of

Support of the Higher Education Commission Pakistan & Mehran University of Engineering and Technology, Pakistan is greatly acknowledged.

power generated from wind and solar energy system due to weather issues, the battery backup will supply the power to meet load demand. Through suitable interfacing circuits, various energy sources are coupled with the bus. The proposed hybrid system can be extended easily when other energy generation resources are accessible.

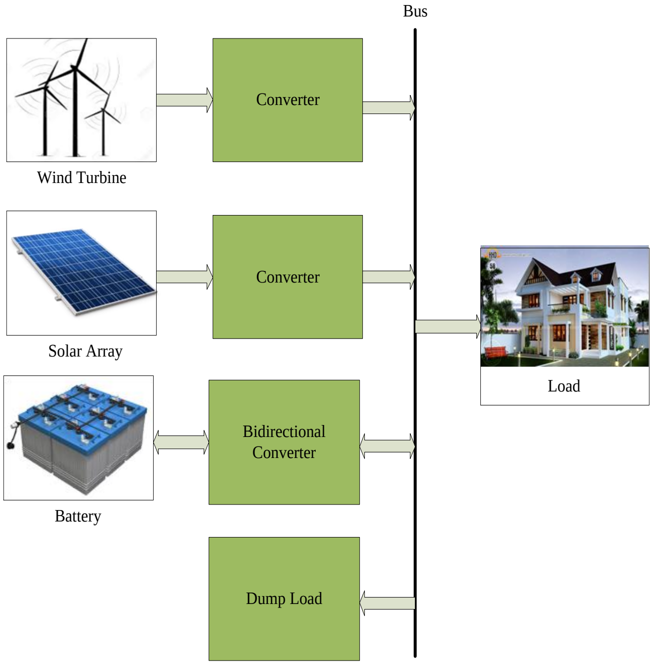


Fig. 1. Schematic diagram of proposed hybrid energy system.

B. System Unit-Sizing

The process of unit-sizing is considered for a proposed hybrid energy system with the recommended arrangements (as shown in Fig. 1) for domestic electricity supply in Jamshoro. It is clear from the Fig. 2 that minimum load demand is 5.5 kW while maximum demand is 12.8 kW. The core objective is to correctly size the elements of the hybrid system in order to confirm the reliable power supply to meet the load requirements.

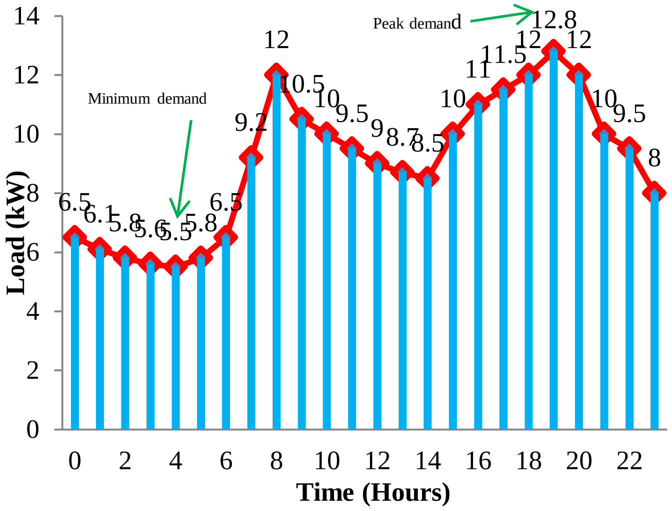


Fig. 2. Hourly average load of five homes in Jamshoro response.

The HES is planned in order to supply five homes in Jamshoro. In this analytical study an hourly average domestic load of five homes in Jamshoro is considered, as shown in Fig. 2. The technique for determining the sizing of solar array is given below:

The capacity factor (K_{cf}) can be characterized as follows [1, 12]:

$$K_{cf} = \frac{P}{P_{rated}} \qquad \text{“(1)”}$$

Where (P) and (P_{rated}) are the average power output and rated power of a renewable energy resource respectively. The theoretical capacity factor for wind is varies from 0-100 %, while in practice it is mostly up to 30 %. However, in this study authors have assumed 14 % and 10 % capacity factor for wind solar respectively[1].

The purpose of unit sizing is to reduce the difference between generated power (P_{gen}) from alternative energy sources and load demand (P_{dem}) for a time period (T), and time considered in this study is only one year data.

$$\Delta P = P_{gen} - P_{dem} = K_{cf\,wtg} * P_{wtg\,rated} + K_{cf\,solar} P_{solar\,rated} - P_{dem} \qquad (2)$$

Where $P_{wtg\,rated}$ and $P_{solar\,rated}$ are the power ratings of wind turbine and solar array respectively.

In order to balance the power generation and load demand, the power rating of solar array is given as:

$$P_{solar\,rated} = \frac{P_{dem} - K_{cf\,wtg} * P_{wtg\,rated}}{K_{cf\,solar}} \qquad (3)$$

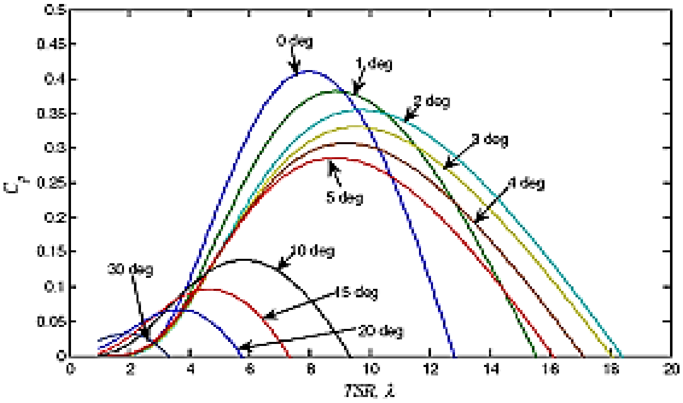


Fig. 3. Characteristics of the WT at different pitch angle response [11], [12], [15].

As shown in Fig. 2, the average load requirement is 9 kW approx., therefore, according to equation (3), the solar array will be of 20 kW in size.

The battery is used as storage device and offers backup for the system. The battery should be capable to store surplus power and to supply power to load during deficiency of power generation. The battery bank of 1800 AH rating is utilized in this stand-alone hybrid energy system as shown in Table I.

TABLE I. SYSTEM COMPONENT PARAMETERS

Wind Energy System Components	**Ratings**
Rated power of wind turbine	50 kW
Cut in speed	3 m/s
Rated speed	14 m/s
Cut out speed	25 m/s
Solar Array	
Power rating	20 kW
Efficiency	18 %
Battery	
Capacity	1800 AH
Battery rated voltage	48

C. Wind Energy System Model

The mechanical power taken out from wind turbine (P_{wind}) is expressed as follows [1], [2], [11], [13], [15] and [19]:

$$P_{wind} = 0.5\rho A v^3 C_p(\lambda,\theta) \quad (4)$$

Where ρ is the air density in *kg/m*^{3}, and its equal to 1.225 *kg/m*^{3}, A is the swept area of blades in *m*^{2}, its value is taken as 10 *m*^{2}, v is the wind speed in *m/s* and C_p is the power coefficient and it is the function of the tip speed ratio and pitch angle. The theoretical value of C_p is 0.59 (see [20], [21]), while its practical value is varies from 0.2 to 0.4 [11], [12], [15]. In this study C_p = 0.3 is considered.

In this paper, a variable-speed pitch controlled wind turbine is used, in which pitch angle controller have a significant role. The $C_p - \lambda$ characteristics of the wind turbine at different pitch angles as shown in Fig. 3. It can be noted from Fig. 3 that the value of C_p changes with the change in pitch angle (θ). Therefore, wind turbine output power can be controlled by means of pitch angle control.

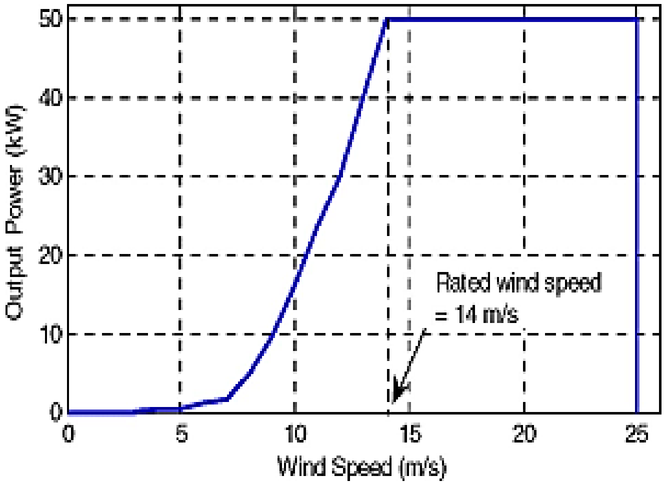


Fig. 4. WT output power vs. Wind speed response [11, 19].

The wind turbine output power versus wind speed characteristics are shown in Fig. 4. It can be observed from Fig. 4 that wind turbine output power retains constant when wind speed is higher than rated speed (i.e., 14 m/s), despite the wind turbine can generate furthermore power. This can be done with pitch angle control in order to avoid rotor over speeding and to safeguard electrical system. In this proposed study, the system will be taken out of operation when the wind speed is greater than cut-out speed (i.e., 25 m/s) in order to protect system.

D. Solar Energy System Model

In solar power generation system solar energy is directly transformed into electrical energy. A solar power generation system comprises of one or more than one photovoltaic panels in series or parallel in order to deliver required voltage and current. Solar power generation system is the most favorable source due to eco-friendly. The output power of the solar array depends upon the area of the solar array, solar irradiance and efficiency of the solar array.

The power extracted from solar array can be calculated as follows [15], [16]:

$$P_{solar} = A * H * \eta \quad (5)$$

Where A is the area of the solar array, H is solar irradiance and η is the efficiency of solar array. The Fig. 5 shows the I-V characteristic curves at different solar irradiance (at 25 ^{0}C). It can be observed from Fig. 5 that greater the solar irradiance, higher will be the short circuit current I_{sc} and open circuit V_{oc} voltage. Hence higher will be the solar output power.

Fig. 6 shows the impact of temperature on solar model performance, it can be shown that lesser the temperature, higher will be the solar output power and larger will be the V_{oc} [4], [14].

E. Battery System

The battery is the necessary component of the hybrid system in order to store surplus energy produced by the hybrid system and to deliver stored energy when the energy generated by the wind and solar energy sources is insufficient to meet the load requirements. Various types of batteries are available i.e., Nickel-Cadmium, Lithium-Iron, Lead-Acid, whereas Lead-Acid battery is commonly used to store and supply the energy. For hybrid energy system, the Lead-Acid batteries are less costly and more efficiently [4], [13]-[15].

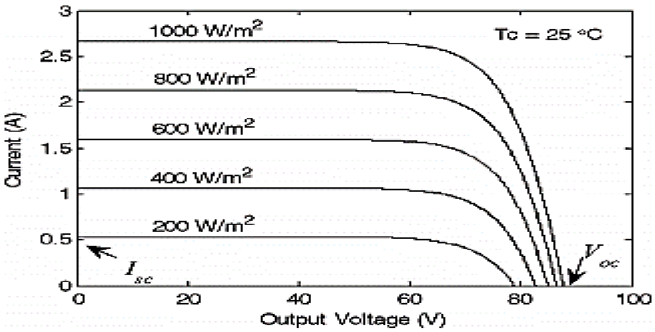


Fig. 5. V curves of solar model at different irradiances response [15], [16].

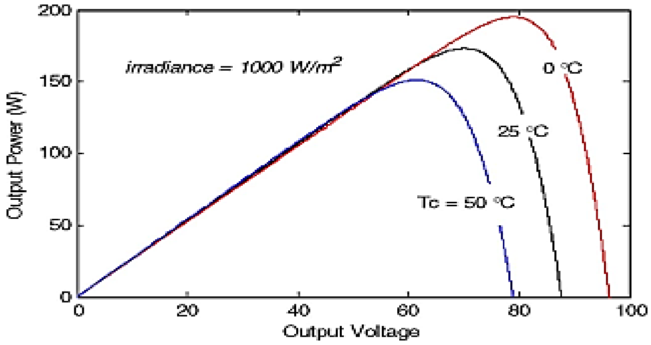


Fig. 6. P-V curves of solar model at different temperature response [4, 14].

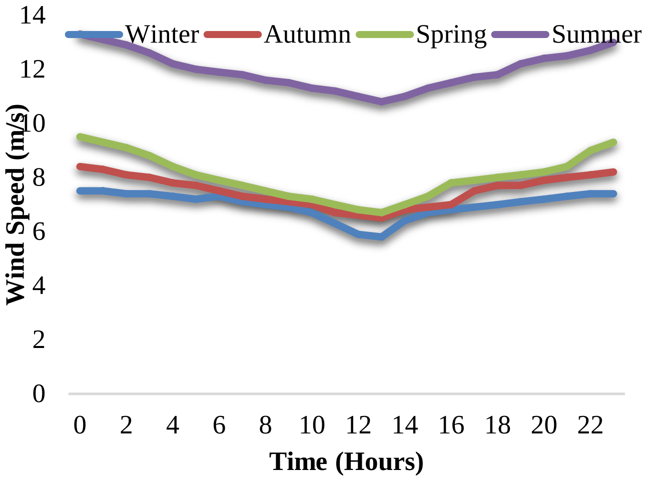


Fig. 7. Wind speed data response (at a height of 50m).

III. Modeling and Power Management Strategy

In this proposed hybrid system, wind and solar energy sources are utilized as chief power producing sources. The power produced by these energy sources are weather dependent, therefore battery is used to store extra power and to supply it when required by the load [2]. In multisource hybrid energy system, it is required to have an overall control and power management strategy among various energy sources. The difference between power produced by energy sources and requirement of the load is given as:

$$P_{net} = P_{wind} + P_{solar} - P_{load} \tag{6}$$

Where P_{wind} , P_{solar} and P_{load} are wind output power, solar power and load demand in *kW* respectively.

The main control strategy is that, the surplus power generated by the wind and solar energy sources $(P_{net} > 0)$ is delivered to the battery. Therefore, (6) will become:

$$P_{wind} + P_{solar} = P_{load} + P_{battery} \quad ; \quad P_{net} > 0 \tag{7}$$

while during shortage of power generation from wind and solar sources $(P_{net} < 0)$, the battery supplies power to the load. Therefore, the equation will become:

$$P_{wind} + P_{solar} + P_{battery} = P_{load} \quad ; \quad P_{net} < 0 \tag{8}$$

IV. Result and Discussions

Analytical studies has been done in order to verify the performance of the proposed system under diverse conditions by using experimental weather data collected at proposed site. In this analytical study an hourly average domestic load of the five homes in Jamshoro is used. The wind speed and weather data is obtained from Pakistan Metrological Department (PMD) [17]. This study carried out for the management of power during four seasons throughout the year. The load demand data is kept similar for the four seasons of the year. The results for the four seasons of the year are discussed in the following section.

A. Four Seasons (Winter/ Spring/ Summer/ Autumn) Descriptions

The wind speed and solar irradiance are lower in winter season (December-February), spring season higher than winter season (March-May), summer season (June-August) is higher than winter, spring and autumn seasons of the year, and autumn season (September-November) is higher than winter season but lesser than spring and summer seasons. From four seasons, the hourly wind speed (at a height of 50 m), solar irradiance and temperature data for a period of 24 hours are shown in Fig. 7, 8 and 9, respectively. The generated wind power and solar power over 24 hour's period of the day are shown in Figs 10 and 11 respectively. During excessive power generated from wind and solar sources ($P_{net} > 0$), the extra power available is supplied to the battery to store while during shortage of power generated ($P_{net} < 0$), than battery will supplies the power to the load, as shown in Fig. 12. In Fig. 12, positive value shows the excessive power available, which can be supplied to the battery while negative value shows the power supplied by the battery bank to meet load requirement.

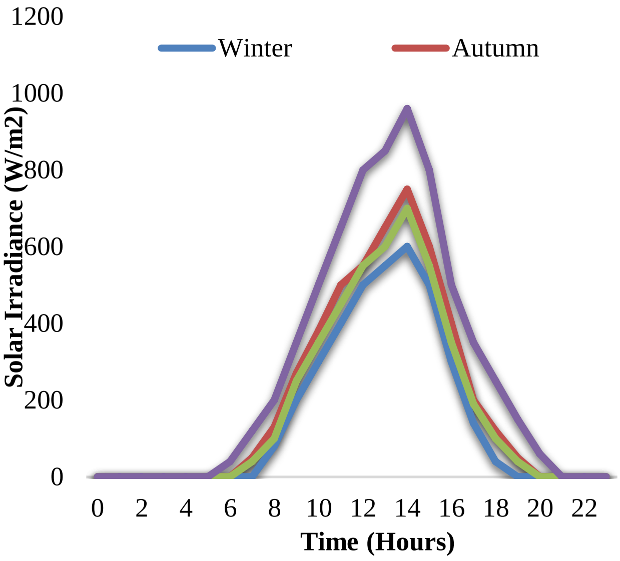


Fig. 8. Solar irradiance data response.

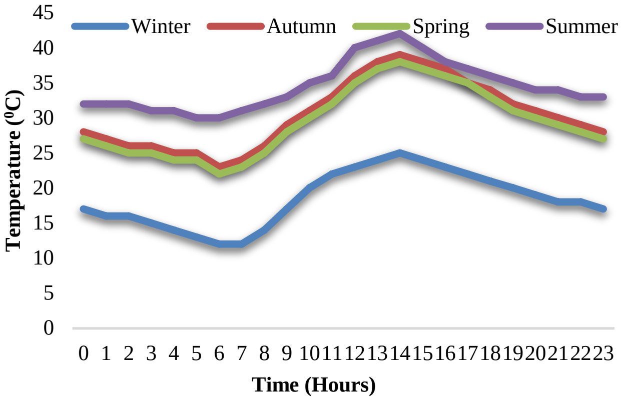


Fig. 9. Temperature data response.

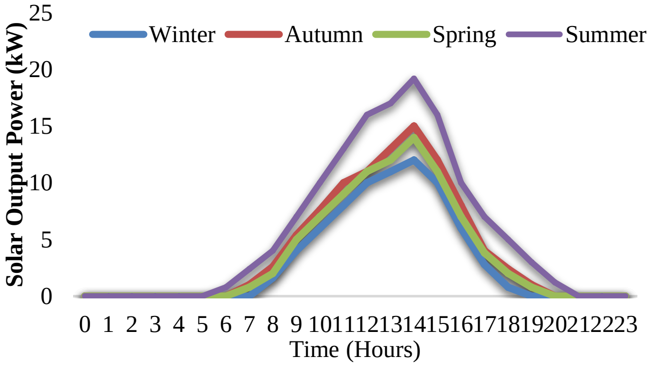


Fig. 10. Wind power response.

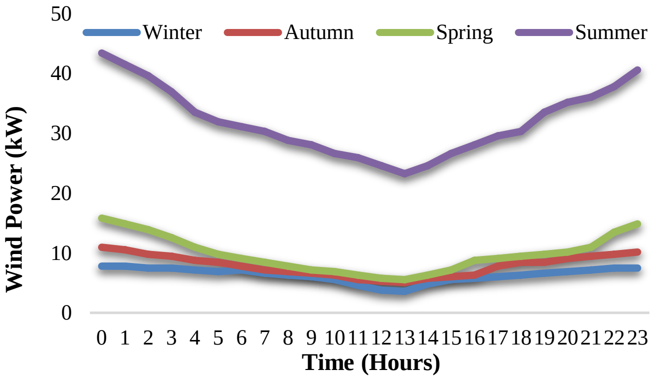


Fig. 11. Solar output power for four seasons response.

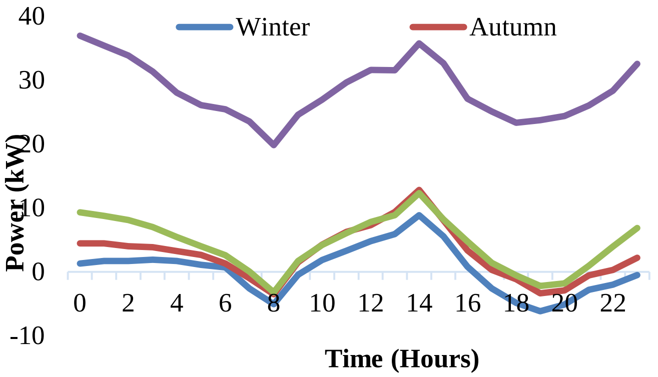


Fig. 12. Excessive power available for storage & power supplied by the battery during four seasons' response.

V. Conclusion

This paper proposes a stand-alone hybrid wind/solar/battery energy system. The HES configuration, system unit-sizing, characteristics of the key system components, modeling and overall power management strategy of the suggested stand-alone hybrid system is discussed. The wind and solar energy sources are utilized as main power producing systems while battery is used as energy storing system. The battery stores surplus power during excessive power generation from wind and solar sources while it supplies power when there is shortage of power generation to meet load requirement. Analytical studies are carried out to verify performance of a proposed system. The analytical studies of the four seasons i.e., winter, spring, summer and autumn, shows the usefulness and feasibility of a suggested stand-alone hybrid wind/solar/battery energy system. Through the experimental investigation, it is clearly observed that the Hybrid system can full the demand of the consumer where yet no power.

References

[1] Wang, Caisheng, and M. Hashem Nehrir. "Power management of a stand-alone wind/photovoltaic/fuel cell energy system." *IEEE transactions on energy conversion* 23.3 (2008): pp.957-967.

[2] Krishnan, M. Sivaram, M. Siva Ramkumar, and M. Sownthara. "Power management of hybrid renewable energy system by frequency deviation control." *Int. J. Innov. Res. Sci. Eng. Technol* 3.3 (2014): pp.763-769.

[3] Ramkumar, M. Sivaram Krishnan M. Siva, and A. Amudha. "Frequency Deviation Control in Hybrid Renewable Energy System Using FC-UC." *International Journal of Control Theory and Applications* 10.2 (2017): pp.333-344.

[4] Nehrir, M. H., et al. "A review of hybrid renewable/alternative energy systems for electric power generation: Configurations, control, and applications." *IEEE transactions on sustainable energy* 2.4 (2011): pp.392-403.

[5] Baloch, Mazhar Hussain, et al. "A Research on Electricity Generation from Wind Corridors of Pakistan (Two Provinces): A Technical Proposal for Remote Zones", *Sustainability*. 2017; 9(9):1611**.**

[6] Baloch, Mazhar Hussain, et al. "Feasible Wind Power Potential from Costal Line of Sindh Pakistan." *Research Journal of Applied Sciences, Engineering and Technology* 10.4 (2015): 393-400.

[7] Siddique, Muhammad Noman, et al. "Optimal integration of hybrid (wind--solar) system with diesel power plant\newline using HOMER." *Turkish Journal of Electrical Engineering & Computer Sciences* 23.6 (2015): pp.1547-1557.

[8] Rehman, Saif Ur, et al. "Feasibility study of hybrid energy system for off-grid rural electrification in southern Pakistan." *Energy Exploration & Exploitation* 34.3 (2016): pp.468-482.

[9] Kant, Jyoti, and Hari Kr Singh. "Scope and Potential of a Hybrid Solar & Wind Energy System for Jodhpur Region Case study." *International Journal of Science and Research (IJSR)* 3.6 (2014): pp.1603-1606.

[10] Negi, Swati, and Lini Mathew. "Hybrid renewable energy system: A review." *International Journal of Electronic and Electrical Engineering* 7.5 (2014): pp.535-42.

[11] Baloch, Mazhar Hussain, et al. "Active and Reactive Power Control of a Variable Speed Wind Energy Conversion System based on Cage Generator." *International Journal of Advanced Computer Science and Applications*8.9 (2017): pp.197-202.

[12] Baloch, Mazhar H., Ghulam S. Kaloi, and Zubair A. Memon. "Current scenario of the wind energy in Pakistan challenges and future perspectives: A case study." *Energy Reports* 2 (2016): pp.201-210.

[13] Baloch, Mazhar Hussain, et al. "A Review of the State of the Art Control Techniques for Wind Energy Conversion System." *International Journal of Renewable Energy Research (IJRER)* 6.4 (2016): 1276-1295.

[14] Chedid, R., H. Akiki, and Saifur Rahman. "A decision support technique for the design of hybrid solar-wind power systems." *IEEE transactions on Energy conversion* 13.1 (1998): pp.76-83.

[15] Marisarla, Chaitanya, and K. Ravi Kumar. "A hybrid wind and solar energy system with battery energy storage for an isolated system." *International Journal of Engineering and Innovative Technology (IJEIT) Volume* 3 (2013): pp.99-104.

[16] Ingole, Ashish S., and Bhushan S. Rakhonde. "Hybrid power generation system using wind energy and solar energy." *International Journal of Scientific and Research Publications* 5.3 (2015): pp.1-4.

[17] Report: Pakistan Metrological Department (PMD) www.pmd.gov.pk.

[18] Ghulam .S. Kaloi et.al., "Wind Energy Potential at Badin and Pasni Costal Line of Pakistan." *International Journal of Renewable Energy Development* 6.2 (2017): 103.

[19] Baloch, Mazhar Hussain, et al. "Stability and nonlinear controller analysis of wind energy conversion system with random wind speed." *International Journal of Electrical Power & Energy Systems* 79 (2016): 75-83.

[20] Ghulam .S. Kaloi et.al., "Dynamic Modeling and Control of DFIG for Wind Energy Conversion System Using Feedback Linearization, *Journal of Electrical Engg. and Technology (JEET), 11.5(2016): pp-1137-1146.*

[21] Ghulam .S. Kaloi et.al., "Active and reactive power control of the doubly fed induction generator based on wind energy conversion system." *Energy Reports* 2 (2016): 194-200

Text Separation from Graphics by Analyzing Stroke Width Variety in Persian City Maps

Ali Ghafari-Beranghar
Department of Computer Engineering, Science and Research Branch, Islamic Azad University,
Tehran, Iran

Ehsanollah Kabir
Department of Electrical and Computer Engineering, Tarbiat Modares University,
Tehran, Iran

Kaveh Kangarloo
Department of Electrical Engineering, Islamic Azad University,
Central Branch, Tehran, Iran

***Abstract*—Text segmentation is a live research field with vast new areas to be explored. Separating text layer from graphics is a fundamental step to exploit text and graphics information. The language used in the map is a challenging issue in text layer separation problem. All current methods are proposed for non-Persian language maps. In Persian, text strings are composed of one or more subwords. Each subword is also composed of one to several letters connected together. Therefore, the components of the text strings in Persian are more diverse in terms of size and geometric form than in English. Thus, the overlapping of the Persian text and the lines usually produces a complex structure that the existing methods cannot handle with the necessary efficiency. For this purpose, the stroke width variety of the input map is calculated, and then the average line width of graphics is estimated by analyzing the content of stroke width. After finding the average width of graphical lines, we classify the complex structure into text and graphics in pixel level. We evaluate our method on some variety of full crossing text and graphics in Persian maps and show that some promising results in terms of precision and recall (above 80% and 90%, respectively) are obtained.**

Keywords—Document image analysis; text/graphics separation; stroke width; raster map; Farsi; Persian; text segmentation; text label

I. Introduction

Text extraction is a fundamental task in graphical document image analysis. This problem frequently occurs in many applications like the map, form processing and engineering drawing interpretation where text and graphics are processed in mainly different ways [1]-[20]. Text and graphics are usually separated for later analysis and recognition; indeed, text recognition is completely different from graphics recognition in general. Current OCR systems cannot recognize text labels in complex mixed text and graphics. Moreover, both government and business organizations must frequently convert existing paper maps of raster maps into a machine-readable form that can be interfaced with the current geographical information systems (GIS) or optical character recognition (OCR).

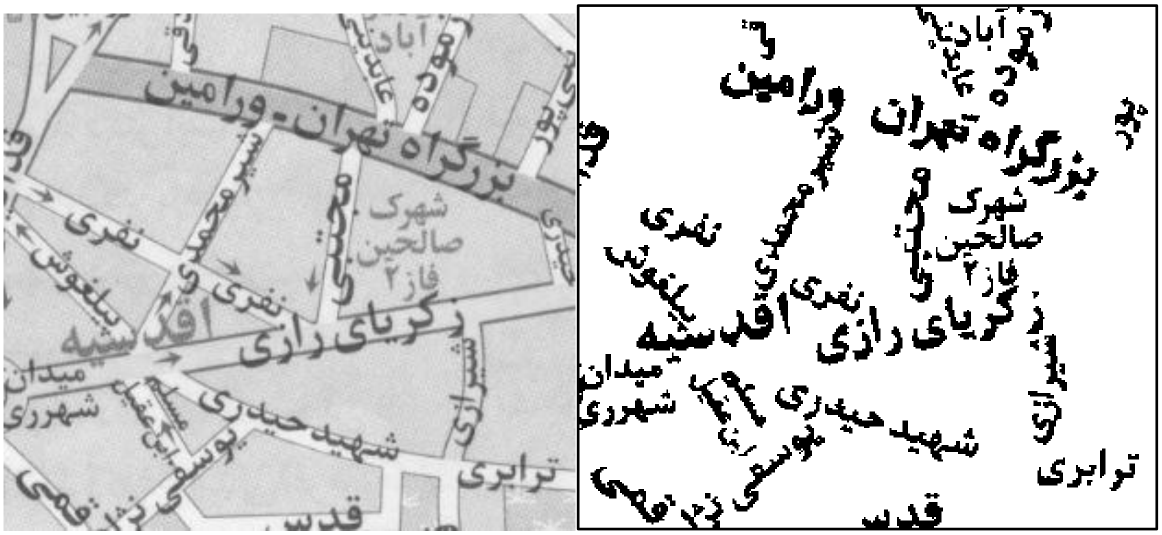
Despite the many studies that have been reported on the text layer extraction from the map, there does not exist any study on the effect of Persian language in the map processing research area [15]. Recent research suggests that language is one of the influential factors in the process of extracting text layers from a map [18]. In Persian, unlike English, one or more characters can be connected to each other and create a subword. In fact, the word (text string) consists of one to several subwords. Therefore, the components of text strings in Farsi have variable dimensions. On the other hand, the overlapping of subwords and lines creates a large and complex structure as shown in Fig. 1.

On the other hand, there are very rich and diverse sources of maps in which the information they contain is valuable. Information from past land and geographic areas that remain untouched without study. Therefore, in order to extract the Persian text layer from the map of the study, it is necessary to design a new method to solve these complex overlapping text and graphical lines. This work will open up a way for further study on the researchers in this area.

In this paper, we provide stroke width-based approach, which is a local descriptor of text, in urban maps that contain a wealth of text labels. Due to the complexity of the text and lines in these types of maps, macro-micro features are combined to separate the subwords and lines.

The rest of this paper is organized as follows: In Section 2, the proposed method is described in detail. And the experimental results and analysis of the performance are given in Section 3. Finally, the concluding remarks are given in Section 4.

A. Related Works



(a) (b)

Fig. 1. Part of street map of Tehran, b) Persian Text layer of map.

In this section, we review some related studies in the literature. Chiang has presented an extensive survey in map processing area [15]. Fletcher and Kasturi [1] proposed a method for text layer segmentation. The main assumption is based on not touching text and graphics. By analyzing the size and geometrical features of objects, text components are separated from graphical lines. The approach works well for simple maps in which text and graphics are not touched together. Such simple maps are not so common in varieties of applications like city maps in which text and graphical lines fully overlapped each other. In contrast, we propose the method not need to satisfy these limitations. In complex maps, these assumptions cannot be practical.

Cao and Tan [4] propose a method based on the assumption that simple touch occurs between a limited number of characters of a text string and associated lines. This method uses a thinning process of the input image to detect the region intersection. Then using some heuristic rules, the lines and the text are separated from each other. This method is able to separate a simple overlapping pattern between lines and text. Especially they assume that text characters are separate. In contrast with our method, we do not assume that the characters are necessarily separate from each other. In addition, our method can deal with complex patterns of overlapping text and lines as in Persian text words and lines in high-density city maps.

Tombre et al. [5] proposed a method to consolidation Fletcher and Kasturi's method. They assume that some characters of the text must be non-touched. Based on non-touching characters to lines, it looks for characters associated with graphic lines. Using skeletonization of big line structure, some heuristics rules are used to classify parts of the structure into text stroke and line segments. Tomber's method can be used in simple maps based on finding a major number of characters in a text string, while in Persian language maps, these conditions cannot be met. In the case of Persian text strings, the characters can be joined to each other, and it can not necessarily be assumed that a significant portion of the characters of the text can be separated from the associating lines, especially in complex maps with a high rate of text and lines overlapping as it is common in Persian city maps.

Cheng and Liu [6] proposed a method based on the assumption that a line as an interferential curve in the text image must be detected and then separated from it. The graph representation of input image is obtained using thinning process. Then, the shortest path algorithm is used to detect interferential curve and removes it from the input image. The text layer obtained from this process is an image thinned in which text quality has greatly reduced. The main limitation of this method is the separation of the lines depending on the thinning of the whole image, so the text image quality from the perspective of text recognition process is controversial. In addition, the method is assumed that the length of the lines should be greater than a predefined threshold and the curvature gradient should be smooth. Therefore, application of the method is not practical, especially in real maps in which these assumptions are not satisfactory. Compared to our proposed method, there are no limiting assumptions on the shape of lines or gradients, as well as line length.

Zhong's method [7] is based on the approximation of text and graphics intersection region. The intersection region is approximated by a polygonal shape. They detect the intersection of text stroke and graphical lines objects using heuristics rules. In complex text and graphics overlapping, the method has complex intersection regions and their heuristics rules cannot be applied efficiently to detect membership of intersection region. In contrast with our method, Zhong's method detects intersection region of text and lines, while our approach focuses on analyzing the complex big line structures that occur usually in complex Persian maps.

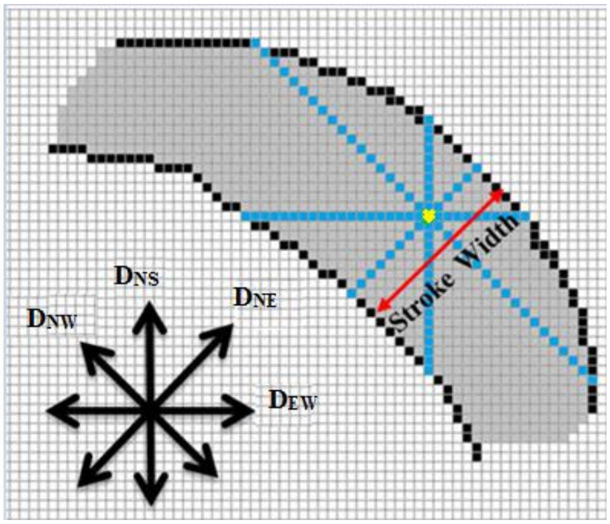


Fig. 2. Stroke width in the yellow pixel has been shown by the red double arrow as the smallest distance between the four directions passing through in the yellow pixel: DEW as east-west, DNS as north-east, DNE as north-east direction, and DNW as north-west direction.

Luo and Kasturi [8] designed some directional morphological operator to extract linear features from maps. Their method can separate lines touched to text. However, the process of morphological is based on manual iteration design. So the method is the difficult approach practically to apply and is dependent on expert knowledge of maps processed. In contrast, our method is not dependent on operator's effort to the analysis of maps.

Li et al. [9] proposed an OCR-based method by training prototypes of characters separated by an operator as training data using template matching approach. They assume some characters of text layer in the map are separable before completely text layer separation. Their approach is language and font dependent and requires user involved in text layer separation. In contrast, our method is not dependent on a specific font.

Tofani and Kasturi [10] proposed a method based on a priori knowledge about the text color of maps. So it is the map dependent method and it cannot apply to some variety of non-color maps. They assume text layer has two major colors, Black and Pink. So they experimentally by color thresholding separate text layer for color maps, then text separated from the graphical line. They find text region and then using line tracking, based on line thickness all pixels of line tracked are removed from the image. They assume that line width is constant and color of text layer is known. However, in contrast, we do not assume that the color of text layer is known.

Chiang and Knoblock [11] propose a user-centric approach to separate text layer of color map by some example as a sample text region or non-text sample to recognize text and non-text color. So based on these text colors found, text layer is extracted from map. In contrast, we propose a new method to extract text layer automatically, to generalize text layer separation from color map, complex text and graphics overlapping is consider using a micro level feature. In contrast, our proposed method is automatic.

Roy et al. proposed a method [12] and a system [13] based on color in raster maps. In [4], the color segmentation is applied to large graphics components to separate into smaller parts that differ in color. This method can be used for maps that have both high-level quality print and the color of text and graphics distinct from one another while color feature cannot be used in many maps, especially the historical maps. In Addition, the color of text and graphics cannot be distinguished even in color maps. In contrast, our proposed method is independent of the colors of the map. In [5], they perform the color segmentation to get different color layers. They assume that the dominant colors represent the layers of the map. This method works well for maps in which color of text layer is completely different from other layers of the map. However, this assumption is not common in all raster maps.

Levachkine et al. [14] proposed a global dynamic thresholding method to convert the RGB color map into the binary image to detect foreground layer of the map. The text layer is separated from the obtained binary image using connected component analysis. In fact, they use color as a global feature to segment foreground layer, they assume that binarization map is the suitable approach for foreground layer separation. This method can work on some maps, but in maps where the color contrast between the text layer and the rest of the layers is low, it causes loss of parts of the text.

Velázquez [15] proposed global thresholding method on the linear combination of RGB color space components (R, G, B) to obtain a binarized image map. Then V_line technique is used to detect a bounding box on the text string to feed to OCR that is trained for synthetic characters with help of Gazetteer (place names of geographical names list). They assume that some major parts of text strings are not touching to graphics. So the method can separate text string that only a limited number of characters are in touch with the associated line. In addition, the method is language dependent and it requires user's effort for the training of synthetic characters.

In contrast with above approaches [10], [12]-[15], our method is not dependent on the color of maps. It can handle the complex overlapping text and graphics situation. In addition, we do not assume about the shape or line with or line pattern to be separated from text objects as well as the isolation of characters, which is a hypothesized in the above approaches.

B. *Local Stroke Width Feature*

Stroke width of the text is defined as the distance between two parallel edges of a stroke. Fig. 2 shows part of a stroke in an image and the stroke width associated with it by a red double arrow in the yellow pixel of stroke [16].

II. Proposed Method

The block diagram of the proposed method has been shown in Fig. 3. Details of each step have been explained in the following sections.

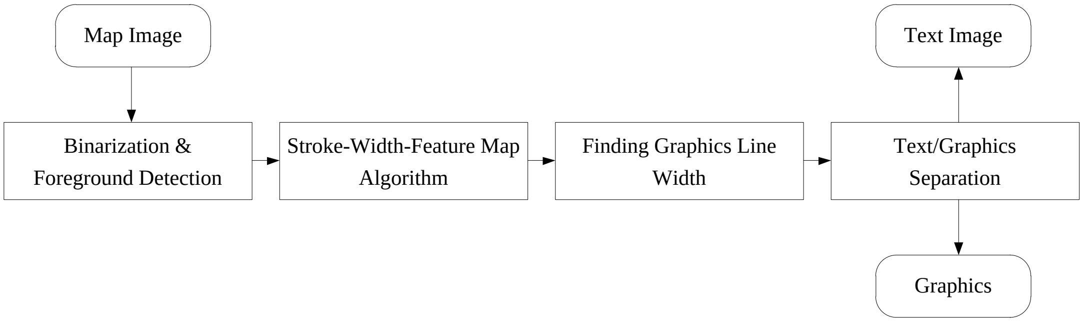

Map Image
Binarization & Foreground Detection
Stroke-Width-Feature Map Algorithm
Finding Graphics Line Width
Text/Graphics Separation
Text Image
Graphics

Fig. 3. The block diagram of the proposed method.

A. Binarization and Foreground Detection

Text and graphical lines are in the foreground layer of the map. So, it is used the best known local adaptive binarization method proposed by Sauvola [17] to detect foreground pixels. The Sauvola's method accurately keeps text labels and lines as foreground pixels while removing the background pixels. So, empirically, by setting binarization parameters the foreground of the input map image is detected properly to be used for further analysis. Without loss of generality, it is assumed that the foreground and background pixels have been assigned by '1' and '0' respectively.

B. Stroke-width-feature Map Algorithm

In this section, we present an algorithm for generating stroke width feature map. The Epshtein's method [18] was simplified to calculate local stroke width for each foreground pixel [16]. In Fig. 2 the local stroke width of the foreground yellow pixel has been shown by a red double arrow. So we will obtain a stroke-width feature map as an image of the equal size in which each pixel content has a stroke width. The stroke-width feature map is obtained using algorithm 1 as follows: for each foreground pixel, four distances are calculated. Each distance is the length of the line segment passing through the pixel in four directions of north-south, east-west, north-east and north-west. The smallest distance is obtained as the local stroke width of this foreground pixel.

```
ALGORITHM 1. Create SW-feature map
Input: BW, binary image of input map image
Output: SW image with the same size as of BW
SW = Zeroes (size of BW);
For each foreground pixel of BW, $p_i$ in position of (x, y)
    Calculate Four Distances as Follows:
    $D^i_{EW}$= number of adjacent connected pixels $p_i$ on the east-west direction until reaching to a background pixel
    $D^i_{NS}$ = number of adjacent neighbor pixels $p_i$ in north-south direction until reaching to a background pixel
    $D^i_{NE}$ = number of adjacent neighbor pixels $p_i$ in north-east direction until reaching to a background pixel
    $D^i_{NW}$= number of adjacent neighbor pixels $p_i$ in north-west direction until reaching to a background pixel
    $SW(x, y) = \min(D^i_{EW}, D^i_{NS}, D^i_{NE}, D^i_{NW})$;
End
```

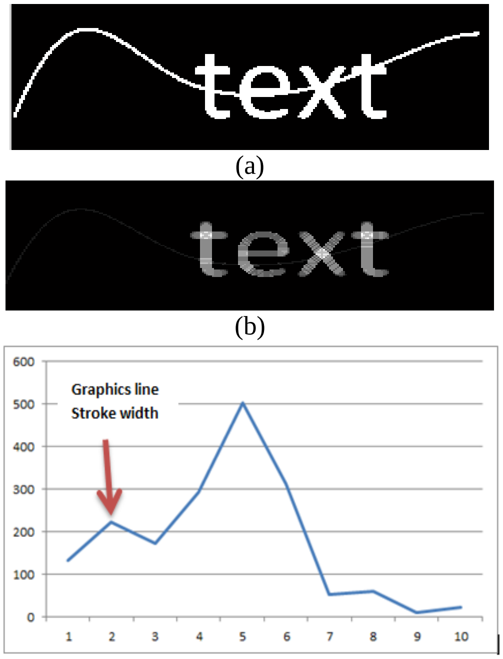
C. Finding Graphic Line Width

After obtaining a stroke-width feature map, the histogram of this feature map is calculated as follows: for each stroke width value in the feature map, we count the number of pixels which have the same stroke width. So, stroke width distribution is found as a histogram. Fig. 4(c) shows the stroke width histogram of the overlapped text and graphics shown in Fig. 4(a). The histogram shows the stroke width variety which is available in the mixed text/graphics input image. Obviously, there is two major stroke widths in this text image: the first stroke width belongs to graphical line overlapped with the text, and the second one belongs to the text. Therefore, the histogram shows the stroke width content of the input image.

By analyzing the stroke-width histogram we can find dominate stroke width of graphic lines and text labels on the map. In city maps, it is observed that the graphic lines on the map are finer than the text associated with it. So, it can be found the most frequent small stroke width as an estimation of the average width of graphic lines. For example, in Fig. 4(c) the smallest dominant stroke width appears at 2. So, it can be concluded that this stroke width is the average graphic line width of the input image.

D. Text/Graphics Separation

By analyzing the histogram of stroke width, the average width of graphic lines can be found. Since in city maps the widths of graphic lines are nearly fixed, so this estimation is used as the threshold of the text/graphics. In the stroke-width feature map, each pixel has the local stroke width associated with that point, so this threshold can be used to classify any point into two classes: graphics and texts. In addition, we observe that in maps usually, the widths of graphic lines are finer than the text stroke width. Therefore, we can classify each pixel of the stroke-width feature map using this threshold. For each point in stroke-width feature map, if pixel value, i.e., local stroke width, is less than the threshold, then it is classified as graphics; otherwise, it is recognized as text. So, two images will be obtained as follow: one for text and the other for graphics. For each pixel that is classified as text, it is assigned a '1' in the corresponding position in a new image (whose size equals the feature map image), and for each pixel classified as graphics, it is assigned '0' (in the same new image).



(a)

(b)

(c)

Fig. 4. a) text and overlapping line image, b) corresponding SW image, c) Stroke width variety histogram.

After classification of pixels into text and graphics, in the text layer, some points appear as noise. Size filter is applied to connected components of text layer extracted to remove them.

III. Experimental Results

A. Data Set

To the best of our knowledge, there exists no standard published data set on maps to evaluate our method. So, for our experiments, we gathered 5 real Persian city map images scanned at 300 dpi from sources like major map publishers, Sahab Geographic and Drafting Institute [19] and National Cartographic Center (NCC) [20]. The most important and dominant characteristics of these collected maps are full crossing where text and graphics are overlapping each other. In some collected city map images, multiple graphic lines have overlapped the text labels, as shown in Fig. 8(a). Graphic line patterns differ due to map publisher styles in map production processes. In some maps, lines are continuous (see Fig. 7(a)) and in some others are dotted or line-segment-dotted, as shown in see Fig. 6(a).

B. Evaluation Methodology

In this section, we evaluate our method on a collected data set. The results of the proposed method are shown on some varieties of collected city maps. To show quantitative performance evaluation, the common standard metrics like precision, recall, and f-measure have been measured using the corresponding map ground truth illustrated in the following section:

$$Precision = \frac{T_p}{T_p+F_p} \tag{1}$$

$$Recall = \frac{T_p}{T_p+F_n} \tag{2}$$

$$F_{measure} = 2.\frac{Precision.Recall}{Precision+Recall} \tag{3}$$

Here T_p is the true positive result, F_p is the false positive and F_n is the false negative result. To measure these factors, text ground truth of map was used as a true reference for the result and the parameters are defined as follow:

T_p : The set of text image pixels that is confirmed by the corresponding text ground truth

F_p: The set of text image that is not confirmed by the text ground truth

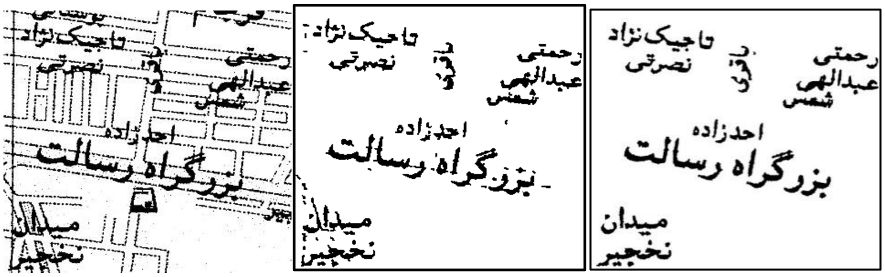
F_n: The set of text ground truth that is lost

$F_{measure}$: The harmonic mean of precision and recall

C. Ground Truth of Text Layer

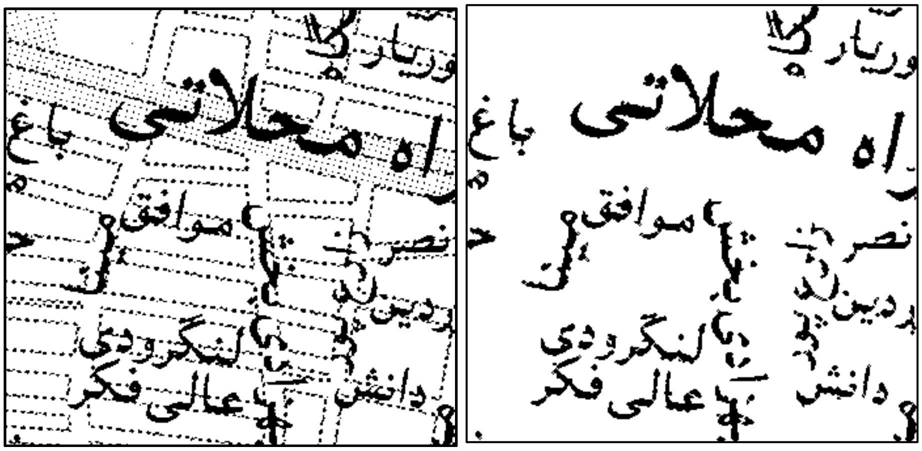
We manually provided text ground truth for some of the maps of our data set to evaluate quantitatively efficiency of our proposed method. At first, the map images were converted to binary images using Sauvola's. Then, we manually removed all the non-text pixels from the map like symbols, graphic lines and background texture that occur in binarization process. In addition, the fragmentary text labels in the map borders are neither true text nor graphic so these objects were also cleared. Fragmentary texts have been removed in the corresponding original map, as shown in Fig. 5(c).

Based on the results, the proposed method can extract the Persian text from a big line structure without dependence on the type of lines or the pattern shape of the lines.



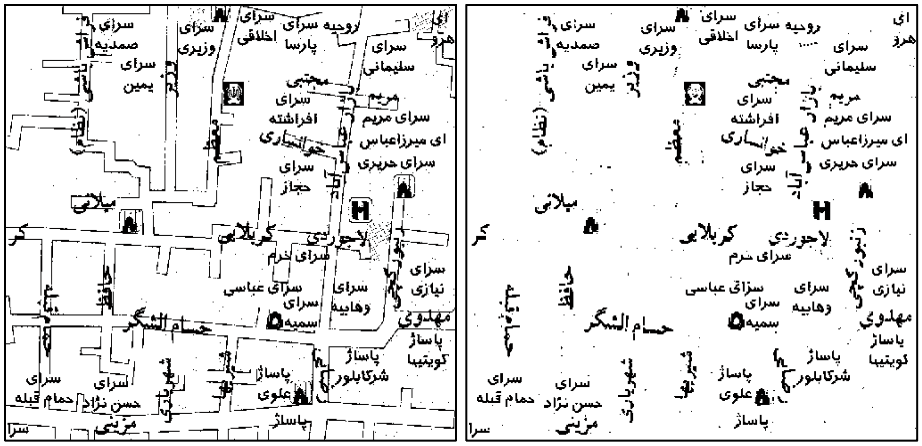
(a) (b) (c)

Fig. 5. a) Part of map from data set, b) Text layer extracted by proposed method, c) Corresponding grand truth; the fragmentary text labels in borders have been removed.



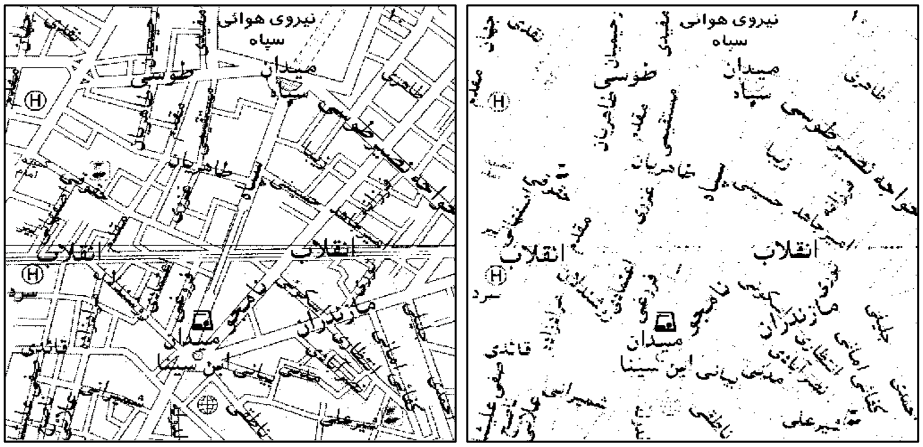
(a) (b)

Fig. 6. a) Part of Tehran map, b) Text layer extracted by proposed method.



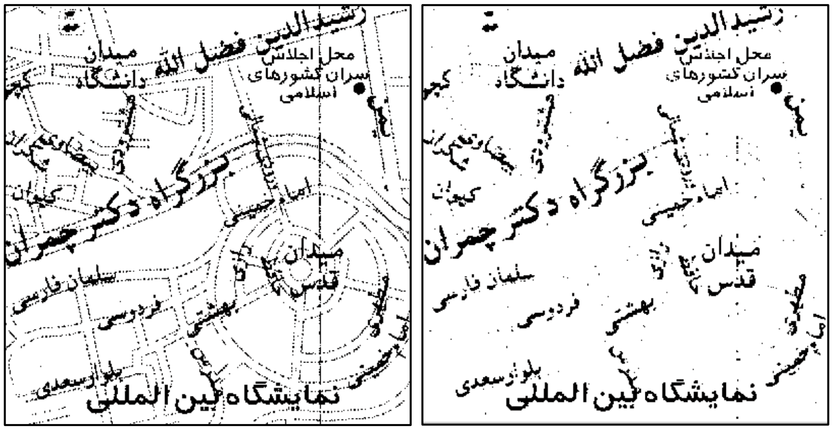
(a) (b)

Fig. 7. a) Original map, b) Text layer extracted by proposed method.



(a) (b)

Fig. 8. Original map, b) Separated text layer by proposed method.



(a) (b)

Fig. 9. a) Original street map, b) Text layer extracted by proposed method.

Corresponding text layer grand truth each input image as the evaluation reference was obtained. The quantitative evaluation results are obtained in Table I. The overlapping lines and text have created a whole unit, as shown in Fig. 5(a)-9(a). The proposed method categorizes the foreground pixels into text and graphics according to the average estimated thickness of line structure in the stroke width feature space.

TABLE I. THE PERFORMANCE EVALUATION OF PROPOSED METHOD

Map Image	Precision (%)	Recall (%)	Fmeasure (%)
Fig. 5(a)	79	96	84
Fig. 6(a)	99.6	98.9	99.3
Fig. 7(a)	86.4	91.4	88.8
Fig. 8(a)	86.5	89.6	88
Fig. 9(a)	92.6	92.1	92.3

Based on the results, the proposed method is able to separate the Persian text layer in different situations such as the continuous lines, as shown in Fig. 7(a), the high density of the lines, and even the overlap of several lines from the entire text, as shown in Fig. 8(a), or the curved lines as shown in Fig. 9(a).

The main limitation of the works is the cases where the quality of the map is low or the text with fine fonts, the quality of the extracted text is low. Also, in cases where lines have a high thickness than the average lines width estimated, the proposed method requires redesign

IV. CONCLUSION

All previous methods focus on extraction of English text string from maps. This paper proposes a new method for text layer extraction in Persian language maps. Our method uses local shape feature, stroke width, to separate text from graphics. Therefore, our method does not depend on the geometric shape of graphical lines, i.e. line length, smoothness, and line pattern as these are the important factor in most previous methods. In addition, it suggests the possibility of adding some features like intensity or color features of map objects to improve the performance of the method. This method can handle complex overlapping text and graphics in high-density maps. However, there is still much work to do. The proposed method cannot deal with poor quality of text in which text stroke width is near the average line with of graphical lines, so new methods should be designed further.

REFERENCES

[1] Fletcher, L.A. and R. Kasturi, A robust algorithm for text string separation from mixed text/graphics images. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 1988. 10(6): p. 910-918.

[2] Sauvola, J., et al. Adaptive document binarization. in Document Analysis and Recognition, 1997., Proceedings of the Fourth International Conference on. 1997.

[3] Luo, H. and R. Kasturi, Improved directional morphological operations for separation of characters from maps/graphics. Lecture Notes in Computer Science, 1998. 1389: p. 35-47.

[4] Tofani, P. and R. Kasturi. Segmentation of text from color map images. in Pattern Recognition, 1998. Proceedings. Fourteenth International Conference on. 1998.

[5] Li, L., et al., Integrated text and line-art extraction from a topographic map. International Journal on Document Analysis and Recognition, 2000. 2(4): p. 177-185.

[6] Cao, R. and C.L. Tan, Text/graphics separation in maps, in Graphics Recognition Algorithms and Applications. 2002. p. 167-177.

[7] Levachkine, S., et al. Semantic analysis and recognition of raster-scanned color cartographic images. in Graphics Recognition Algorithms and Applications. 2002. Springer.

[8] Tombre, K., et al., Text/graphics separation revisited, in Document Analysis System V, E.P. Lopresti, J. Hu, and R. Kashi, Editors. 2002. p. 200-211.

[9] Zhong, D.X., Extraction of embedded and/or line-touching character-like objects. Pattern Recognition, 2002. 35(11): p. 2453-2466.

[10] Velázquez, A. and S. Levachkine, Text/Graphics Separation and Recognition in Raster-Scanned Color Cartographic Maps, in Lecture Notes in Computer Science, Graphics Recognition. 2004. p. 63-74.

[11] Cheng, Z. and Y. Liu, A Graph-based Method to Remove Interferential Curve From Text Image. Machine Vision and Applications, 2006. 17(4): p. 219-228.

[12] Roy, P.P., J. Llados, and P. Umapada. Text/Graphics Separation in Color Maps. in Computing: Theory and Applications, 2007. ICCTA '07. International Conference on. 2007.

[13] Roy, P.P., et al. A System to Segment Text and Symbols from Color Maps. in Graphics Recognition. Recent Advances and New Opportunities, Lecture Notes in Computer Science. 2008.

[14] Epshtein, B., E. Ofek, and Y. Wexler. Detecting text in natural scenes with stroke width transform. in IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2010. San Francisco, CA: IEEE.

[15] Chiang, Y.-Y., S. Leyk, and C.A. Knoblock, A survey of digital map processing techniques. ACM Computing Surveys (CSUR), 2014. 47(1): p. 1.

[16] Ghafari-Beranghar, A., E. Kabir, and K. Kangarloo, Directional Stroke Width Transform to Separate Text and Graphics in City Maps. Journal of Computer & Robotics, 2014. 7(2): p. 1-7.

[17] Chiang, Y.-Y. and C.A. Knoblock, Recognizing text in raster maps. Geoinformatica, 2015. 19(1): p. 1-27.

[18] Chiang, Y.-Y., et al., Assessing the impact of graphical quality on automatic text recognition in digital maps. Computers & Geosciences, 2016. 93: p. 21-35.

[19] National Cartographic Center. NCC. 2018 [cited 2018 1 Jan]; Available from: www.NCC.org.ir.

[20] Sahab Geographic and Drafting Institute. 2018 [cited 2018 1 Jan]; Available from: http://www.sahabmap.com.

Impact of Anaphora Resolution on Opinion Target Identification

BiBi Saqia[1], Khairullah Khan[2], Aurangzeb Khan[3],
Department of Computer Science
University of Science & Technology Bannu
Bannu, Pakistan

Fazali Subhan[5]
Department of Computer Science
National University of Modern Languages
Islamabad, Pakistan

Wahab Khan[4]
Department of Computer Science and Software Engineering
International Islamic University,
Islamabad, Pakistan

Muhammad Abid[6]
Institute of Information Technology
Kohat University of Science & Technology Kohat,
Kohat, Pakistan

***Abstract*—Opinion mining is an interesting area of research because of its wide applications in the decision-making process. Opinion mining aims to extract user's perception from the text and to create a fast and accurate summary of people's opinion about anything. In this study, we have worked on opinion target identification and the impact of anaphora resolution on opinion target extraction. Anaphora resolution can be utilized to detect opinion target in sentences having prepositions instead of nouns. We empirically evaluated the impact of anaphora resolution using benchmark datasets. We have achieved accuracy such as precision: 88.14 recall: 71.45 and f-score: 72.12, respectively.**

Keywords—Opinion mining; machine learning; evaluative expression; anaphora resolution; opinion targets

I. Introduction

Opinion is a personal view, statement, or judgment of an individual about something [1]. People's view, knowledge, experience play an important role in human guidance and decision making [2]. For example in the sentence "Mahnoor Baloch is a good actress" is positive opinion regarding Mah,Noor Baloch, "Kamran is not a good player" is negative opinion, while "Milk is good for health but tea is not" is a neutral sentence. Opinion has different components. Identification of each component from free text is a challenging task [3], [4]. This study is about opinion target identification. OM attempts to find the evaluative perspective of natural language context [9]. The evaluative expression represents a source, an attitude, and a target or destination. For instance, in the sentence 'I disliked the rooms of the hotel, they were not well decorated', the speaker (the source) communicates a negative behavior regarding 'the rooms of a hotel' (the Target) [3].

Our research problem has different subproblem and has been approached in a different manner: In some papers, it is regarded as subjectivity analysis at the document or sentence level, and some has worked on opinion target and opinion words while some has tried correlation between the two. In this work, our goal is to investigate whether anaphora resolution (AR) can be potentially exploited to get improvement in domain-independent opinion-target pair extraction.

TABLE I. Anaphora Resolution

Sentences	Antecedent	Anaphor
Imran Khan is a good politician but his workers may be corrupted.	Imran Khan	His
Ali loves Maryam. He invites her to lunch.	Ali, Maryam	He, her

The word anaphora came from two ancient Greek words "Ana" and "phora". "Ana" means back, upstream, back in an upward direction whereas "phora" means the act of carrying. Anaphora is employed quite regularly in, both written and verbal discussion to ignore over - a reiteration of words for the purpose to enhance the continuity of terms [5], [6]. Table I shows an example of anaphora resolution:

The natural language processing consists of the various complicated demanding domain of learning in which anaphora resolution considered an important and interesting field of research [7]. The AR is necessary for the utilization of maximum certifiable NLP approaches and it is unavoidable incident in the estimation of sentence structure. In the discussion, an AR is an issue to recognized anaphors about prior else subsequent elements. The described elements (predecessors) may be specified or unspecified noun terms, verb terms, pronouns, the entire expressions, and phrases. There are three basic kinds of anaphor: (a) "pronominal anaphors", this is general sort of anaphor used to recognize antecedents of pronoun anaphors within sentence;(b) "definite noun-phrase anaphors", this sort of anaphor identify an antecedent via a noun terms; (c) "ordinal/quantifier anaphors", This kind of anaphor refers to an ordinal such as second and it may be represented to some unspecified quantity like few, some, etc. [8].

This paper is organized as follows: Section 2 presents the related work on opinion target identification and our problem. Section 3 discusses the proposed framework we employed for domain-independent opinion target extraction, while in Section 4 we explain experimental work. Section 5 concludes the paper.

II. LITERATURE REVIEW

The OM problem has been addressed in many research papers and diverse approaches have been employed for its solution. The OM has been categorized in subproblems as explained by [4]. Opinion words and opinion targets have been identified in different ways. Some work has purely ,used grammatical structure [30] of the language and some employed semantic features [2] and some has used both syntactic and semantic features [11], [18]. The combined approa,ch has shown proven results. In this work, we have adopted the combined approach. However, our goal is to the test the impact of anaphora resolution on opinion targets extraction. As explained in the introduction we regarded the problem of anaphora in context of inopinion target since the object and features in the free text are mostly referred by anaphora. There has been sound work on anaphora resolution. An anaphora resolution has been developed as a source of semantic evaluation with help of word features and Backus-Naur Form (BNF). This technique depends on coordinating restraints for the syntactical features of various wordthe s, opinion, and text. They get approximately 96% accuracy whereas the algorithm was also checked for complicated and composite sentences [12], [13]. The heuristic rules and WordNet ontology was used to enhance the accuracy of anaphora resolution. The intra-sentential and inter-sentential anaphora and pleonastic-it operation in English communication were utilized to improve the resolution accuracy [14]. The relevance scoring between context matrixes and WordNet glosses are used for calculating and extraction of the right sense of target word [12], [15]. The anaphora choice of pronoun has been essential for extracting general needs from the text of necessities report spontaneously [16]. The dependency and dialogue pattern was utilized to provide assistance in the resolution of particular kind of references. To resolved entity pronoun references in Hindi discourse a Paninian grammar dependent heuristic model were applied [17].

To improve unsupervised opinion targets extraction technique patterns and semantic analysis has been employed [18]. For the identification opinion targets, two steps are employed: candidate selection and opinion targets selection. The combined lexical based syntactic pattern was used for candidate selection while a hybrid likelihood ratio test approach with semantic base relatedness was employed for candidate selection [18]. For the annotation of opinions in unstructured text documents, a method was developed. Appositive instances were resolved by using Normalized Google Distance (NGD). Latterly the issue of anaphora resolved documents has been performed by employing the Vector Space Model [19]. The machine learning method has been employed to categorize subjective and objective sentences. They worked on rule-based domain independent opinion evaluation technique. They performed experiments on data collected from different websites [20]. Assigning pre-described categories to textual documents is referred to text classification. They build up a common method to evaluate the semantic relatedness of documents. To increase the semantic importance assign to every document anaphora resolution were used. The hidden meaning of the text was expressed more efficiently by word semantic and WordNet scientific categorization which provided an authentic description as compared to conventional Information method [21]. The rule-based technique was used by the proposed algorithm for Pashto dialect in their oblique, immediate as the well possessive state being the resolution of strong personal pronouns [22]. The pronominal anaphora resolution (PAR) was used with other conventional attributes along with global discourse knowledge. The referent of an anaphoric pronoun was evaluated locally by the attributes involved in searching. Usually, the sentence which includes the anaphor as well as several sentences quickly before structure the neighborhood setting of content. With the processing of discourse, the knowledge base gets were also improved [23]. The superlative entropy model and Random Forest classifier for the pronominal anaphora resolution using benchmarking technique provides precise features of Malay discourse like gender-neutral pronouns. They persist in a particular two steps procedure: First, Managing implantation to investigate the components of Malay anaphors. Second, In light of the investigated output, the pronominal resolution framework was outlining, actualizing, and assessing [8]. To determine the reciprocal pronouns in the Pashto language an algorithm has been developed depends on some specific principle. Since in the Pashto language, the NLP mechanism along with a collection of written explained texts were inaccessible, a little physically labeled and divided corpus was made for Pashto dialect [24]. Several issues were found in resolving pronouns in the Malayalam language compared to English discourse as its free phrase order language. The physical experiment was accomplished by settling anaphora on various stories about the data set. The execution of numerous NLP application like passage abstraction, Passage classification, and text retrieval has been enhanced through anaphora resolution system [25]. The individual pronoun anaphora resolution were assisted to accomplish website page data handling by a large number of paroxysmal text in the web [26]. An algorithm has been developed to resolve the distributive anaphoric connection by utilizing the global learning includes maximum characteristics of the noun in Urdu conversation [27].

The most relevant work to our problem is [10], [11]. They have worked on the improvement of opinion target extraction with anaphora resolution however, their approach is slightly different. Furthermore, their work is specifically for movie domain.

III. PROPOSED ARCHITECTURE

The whole procedure of the proposed structure of opinion target identification from unstructured reviews is discussed in this section. There are two main objectives of the proposed work; to identify opinion targets from evaluative expression and to improve opinion target identification by anaphora resolution. The procedure clarifies how opinion targets can be extracted from an input unstructured review. The following three phases used in this procedure as elaborated in the block diagram (Fig. 1). Every step describes a summary of the sub-steps included in the procedure.

A. Pre-Processing

The pre-processing phase applied for noise removal, sentence division and parts of speech tagging (POS). The POS tagging involves allocating exact grammatical category to every word of the text.

B. Candidate Selection

The identification of candidate features is a vital phase of opinion target extraction [28]. To find out evaluative expressions including opinion and targets the proposed algorithm is employed. This procedure utilized the following three basic steps.

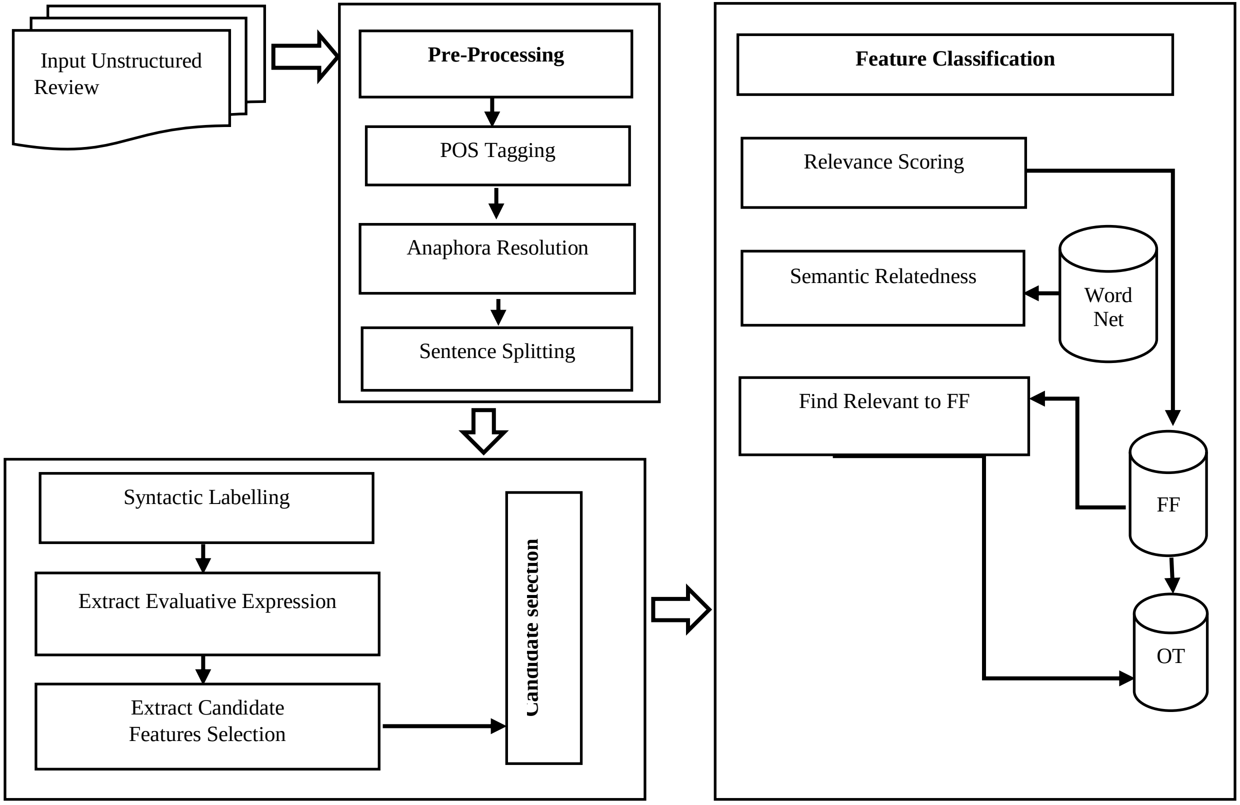


Fig. 1. Proposed architecture.

C. Regular Expressions

We adopted the Regular Expression (RE) pattern from [18]. These patterns are used for extraction of strings containing opinion and targets through base noun phrases along with various boundary conditions. The opinion lexicon dictionary is utilized by the proposed patterns for identification of opinionated expressions that consist of opinion and targets.

D. Candidate Selection

The candidate target features are selected in the extracted evaluative expressions by the pronoun phrase and also to obtain the relevance scoring arranged it according to their no. of occurrence. This algorithm consists of the following two steps.

- In this step, we look for constituents of the lexical patterns in the input sentence. If a sentence consists of any patterns of the proposed pBNP, at that point the sentence is named as opinionated, or then non-opinionated. The algorithm examines the pBNP constituent pattern on priority bases as vBNP, dBNP, iBNP, and sBNP, individually.
- At this stage, a set of a candidate features is produced from the extricated patterns. All pronoun phrases in the evaluative expression take out in step 1 are chosen as candidate features and the recurrence of each particular noun is determined.

E. Opinion Target Extraction

In this step semantic based likelihood ratio technique is derived from [18]. The relevance scoring technique is utilized to categorize candidate features into relevant and irrelevant. The LRT is used to extract opinion targets that happen maximum no of times while semantic based relation is applied to finds targets occurred infrequently. Table II describes sample product features.

F. Enhancement of Semantic-Based LRT through anaphora

In this step, we propose an enhancement of the semantic-based likelihood ratio test technique derived from [18] by anaphora resolution. As given in Table III there are features which are represented by pronouns. In these datasets, targets are calculated manually which are pronouns and then total pronouns are found out in each dataset. The following table shows examples of the targets manually labeled dataset having pronouns.

Table III represents the influence of the pronouns on target features. The influence of the product features in the canon power dataset which contains 60 targets feature out of 173 pronouns, therefore the influence of the pronouns on target features of canon power is 34.68%. Similarly, the influence of Canon S100 is 40.39%, diaper champ is 43.79%, Hitachi router is 17.74%, iPod is 32.27%, Linksys router is 12.5%, micro MP3 is 16.11%, Nokia is 37.5 and Norton is 38.61% respectively.

IV. Results and Discussion

A. Datasets

We have used manually labelled datasets regarding nine products of customer review that have been described frequently in research of opinion mining and target identification. These datasets used for analysis and assessment of proposed work. The author's website is openly used to avail these datasets, every product features for opinion recording is conveniently labeled via a manual procedure with respect to mentioned annotation strategy as follows. Table IV shows an explanation of desired nine datasets.

- The sentence that consists of positive or negative remarks regarding features of the product then this sentence is considered as opinionated.
- The opinion statements consist of positive or negative suggestions described adjectives.
- The criteria for the product are the product feature that represented by the customer's opinions.

B. Tools and Implementation

This Section shows the achievement matrices and assessment principles that have been utilized throughout the time of research process to assure the validity of the results. The accuracy is calculated by utilizing the following three performance matrices.

$$\text{Precision}=\frac{\text{Relevant Instance} \cap \text{Retrieved Instance}}{\text{Retieved Instance}} \quad (1)$$

$$\text{Recall}=\frac{\text{Relevant Instance} \cap \text{Retrieved Instance}}{\text{Relevant Instance}} \quad (2)$$

$$\text{F-score} = 2.\ \frac{\text{Precision.Recall}}{\text{Precision+Recall}} \quad (3)$$

C. Tools and Implementation

This section described explanation regarding simulation tools utilized in this task. The following state-of-the-art software is applied to experiments and simulation. The part of speech tagging is accomplished via the Stanford part of speech tagger [29]. The parts of the speech tagging software are freeware and broadly described in English language texts. The algorithm used in this thesis depends on the grammatical attributes for evaluation of language elements. Thus, by using this software the actual datasets are changed to POS tagged corpora. The test evaluation and pattern extraction are performed by Text Stat 3.0 and from author's website, it's easily accessible for academic research.

TABLE II. Sample Product Features

Explanation	Dataset								
	Canon power	**Canon S100**	**Diaper Champ**	**Hitachi router**	**iPod**	**Linksys Router**	**Micro MP3**	**Nokia**	**Norton**
Target features	battery design color size camera power	camera memory battery pixel resolution autofocus	Odor price weight bags changing working nursery use	routing time routing table price Size design power	Size battery recording power quality service	loading installation security operating system speed bend width connectivity	size installation sensitivity price battery portability	Size battery life camera bluetooth display design	Performance anti-spam internet-security memory usage installation user- interfaces

TABLE III. Measures of Pronouns as a target

Explanation	Datasets								
	Canon power	**Canon S100**	**Diaper Champ**	**Hitachi router**	**iPod**	**Linksys Router**	**Micro MP3**	**Nokia**	**Norton**
Total pronouns	173	203	290	248	158	376	701	232	259
Targets which are pronouns	60	82	127	44	51	47	113	87	100
Average pronouns	34.68%	40.39%	43.79%	17.74%	32.27%	12.15%	16.11%	37.5%	38.61%

TABLE IV. DATASETS DESCRIPTION

Explanation	Datasets								
	Canon power	Canon S100	Diaper Champ	Hitachi router	iPod	Linksys Router	Micro MP3	Nokia	Norton
Reviews	45	34	49	95	51	56	67	88	87
Total sentences	735	678	1478	234	564	234	543	123	456
Sentences with target features and opinion	345(47%)	289(46%)	654(43%)	234(45%)	567(24%)	453(46%)	325(43%)	456(44%)	454(43%)
Total distinct base noun phrases (BNP)	567	537	987	890	768	908	735	890	556
Total target features	345	453	435	456	456	542	556	456	432
Average($\frac{Total\ BNP}{Targets}$)	1.64	1.18	2.26	1.95	1.68	1.67	1.32	1.95	1.28
Target types	145	178	123	156	189	156	124	187	145
$\frac{\boldsymbol{Target\ types}}{\boldsymbol{Total\ targets}}$	0.42	0.39	0.28	0.34	0.41	0.28	0.22	0.41	0.33

The WordNet.Net Library is developed by Troy Simpson and from author's website, it's available openly. This library facilitates the WordNet dictionary for similarity scoring by a DotNet port. The implementation of the semantic-based relevance scoring algorithm is employed by this library. The WordNet dictionary is a collection of a lengthy lexical database consisting of 117000 synsets. Every synset shows a distinctive idea that is combined with the conceptual-semantic and lexical association [30]. MS Excel is used to generate results and graphs.

D. Results

Initially, the datasets are changed over into a parts of speech tagged datasets, utilizing the Stanford parser [29]. At that point, the proposed algorithm is executed through the model framework with the following setups to extract the candidate features.

The experimental setup depends on a combination of four unique patterns, i.e. linking verb base noun phrases, definite base noun phrases, preposition based noun phrases and subjective base noun phrases with pronouns. This setup is named as pBNP.

In every step, the result of each pattern is contrasted with the manually labeled features to recognize True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN). Accuracy such as precision, recall, and f-score is determined by utilizing the confusion matrix generated by the proposed framework.

To make the results comparable, the same setup is used for both the Likelihood and semantic-based Hybrid Likelihood techniques. The evaluation measures precision (P), recall (R) and f-score (F), which are calculated using the following parameters:

- TP = number of extracted pBNP which are target features.
- FP = number of extracted pBNP which are not target features
- TN =number of non-target features pBNP, which are not extracted
- FN = number of targets features pBNP, which are not extracted

This setup implements the semantic likelihood ratio test with the proposed lexical patterns (pBNP). The cBNP-L uses the candidate features extracted through cBNPs and employs the likelihood ratio test for relevance scoring to extract the opinion targets.

E. Influence of Anaphora Resolution in Opinion Target Extraction

Table V shows the result of nine datasets in term of precision, recall, and f-score to the impact of anaphora resolution.

F. Comparative Results of Proposed Method with the Existing Approaches

Table VI presents average comparative results between the baseline, the semantic-based Hybrid Likelihood Ratio Test techniques and Semantic-based LRT with Anaphora resolution in terms of the average precision, recall and f-score respectively.

Fig. 2 describes the comparative results of proposed semantic based opinion target extraction through anaphora resolution with existing hybrid semantic based likelihood ratio test. As shown in the above graph the score of the proposed technique is higher than the existing semantic based Hybrid Likelihood Ratio Test. Subsequently precision decreases slightly while high increase the recall and improve f-score.

TABLE V. PRECISION, RECALL AND F-SCORE WITH EFFECT TO ANAPHORA

Data Set	Precision (%)	Recall (%)	F-score (%)
Canon Power	90.75	63.16	73.78
Canon S100	87.82	58.14	70.11
Diaper Champ	88.54	64.68	75.08
Hitachi router	86.87	60.45	71.59
Ipod	90.05	64.63	75.57
Linksys Router	55.02	80.03	70.34
Micro MP3	60.08	75.07	69.39
Nokia	81.07	75.08	78.63
Norton	67.08	73.02	72.34

TABLE VI. COMPARATIVE RESULTS

Row Labels	Average of Base Line	Average of Semantic-based LRT	Average Semantic-based LRT with Anaphora resolution
Precision	92.02	89.65	88.04
Recall	53.16	60.47	71.45
F score	66.84	71.62	72.12

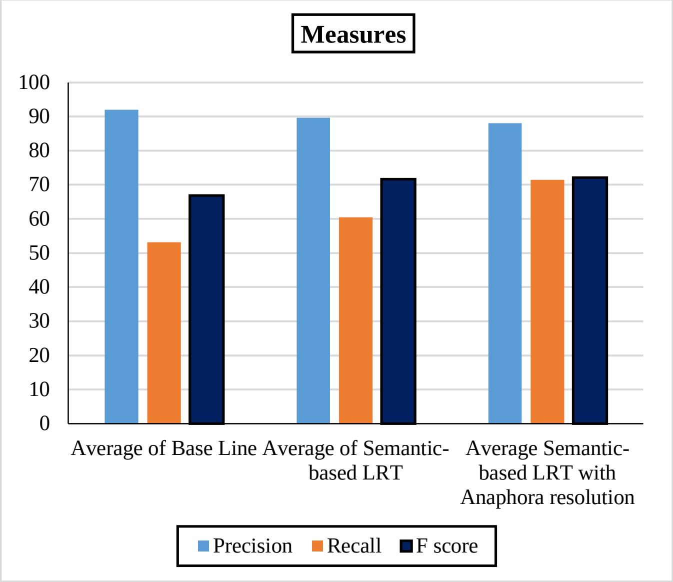


Fig. 2. Comparative results of the proposed method with an existing technique.

V. CONCLUSION

This study describes an impact of anaphora resolution on opinion target identification in text documents. We used nine datasets taken from author website for the evaluation of desired work. The proposed work recognized opinion targets from evaluative expression and slightly enhance its result by employing anaphora resolution. The learning of the current task and drawback of the proposed work discover that there is space for enhancement in the proposed method. Thus, suggested method retrieve domain progressive assessment expressions that can be utilized for identification of target attributes in a cross-domain via a supervised machine learning algorithm. Thus the future task must be given attention in this dimension.

We have demonstrated that by expanding an opinion mining algorithm with anaphora resolution for opinion target extraction, an interesting improvement can be accomplished. Anaphora resolution can also be utilized in other OM algorithms which are used for identification of opinion targets.

REFERENCES

[1] B. Liu and L. Zhang, "A survey of opinion mining and sentiment analysis," in Mining text data, ed: Springer, 2012, pp. 415-463.

[2] E. Breck and C. Cardie, "Opinion Mining and Sentiment Analysis," in The Oxford Handbook of Computational Linguistics 2nd edition, ed, 2017.

[3] K. Khan, B. Baharudin, and A. Khan, "A Review of Unsupervised Approaches of Opinion Target Extraction from Unstructured Reviews," Journal of Applied Sciences, Engineering and Technology, vol. 7, p. 12, 2014.

[4] K. Khan, B. Baharudin, A. Khan, and A. Ullah, "Mining opinion components from unstructured reviews: A review," Journal of King Saud University-Computer and Information Sciences, vol. 26, pp. 258-275, 2014.

[5] R. Mohana, "Anaphora resolution in Hindi: Issues and directions," Indian Journal of Science and Technology, vol. 9, 2016.

[6] E. D. Liddy, "Anaphora in natural language processing and information retrieval," Information Processing & Management, vol. 26, pp. 39-52, 1990.

[7] R. Padmamala, "A Novel Knowledge-engineering based approach for anaphora resolution of Tamil pronouns," in Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM), 2015 International Conference on, 2015, pp. 24-29.

[8] B. C. M. Xian, M. A. Saloot, A. S. M. Ghazali, K. Bouzekri, R. Mahmud, and D. Lukose, "Benchmarking Mi-AR: Malay anaphora resolution," in Optoelectronics and Image Processing (ICOIP), 2016 International Conference on, 2016, pp. 59-69.

[9] K. Bloom, N. Garg, and S. Argamon, "Extracting appraisal expressions," in Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference, 2007, pp. 308-315.

[10] N. Jakob and I. Gurevych, "Using anaphora resolution to improve opinion target identification in movie reviews," in Proceedings of the ACL 2010 Conference Short Papers, 2010, pp. 263-268.

[11] L. Zhuang, F. Jing, and X.-Y. Zhu, "Movie review mining and summarization," in Proceedings of the 15th ACM international conference on Information and knowledge management, 2006, pp. 43-50.

[12] R. Mitkov, The Oxford handbook of computational linguistics: Oxford University Press, 2005.

[13] S. Agarwal, M. Srivastava, P. Agarwal, and R. Sanyal, "Anaphora resolution in Hindi documents," in Natural Language Processing and Knowledge Engineering, 2007. NLP-KE 2007. International Conference on, 2007, pp. 452-458.

[14] T. Liang and D.-S. Wu, "Automatic pronominal anaphora resolution in English texts," International Journal of Computational Linguistics & Chinese Language Processing, Volume 9, Number 1, February 2004: Special Issue on Selected Papers from ROCLING XV, vol. 9, pp. 21-40, 2004.

[15] K. Khan, B. B. Baharudin, and A. Khan, "Mining opinion targets from text documents: A review," Journal of Emerging Technologies in Web Intelligence, vol. 5, pp. 343-353, 2013.

[16] K. S. Park, D. U. An, and Y. S. Lee, "Anaphora Resolution System for Natural Language Requirements Document in Korean," in Information and Computing (ICIC), 2010 Third International Conference on, 2010, pp. 11-14.

[17] V. Mujadia, D. Agarwal, R. Mamidi, and D. M. Sharma, "Paninian grammar based hindi dialogue anaphora resolution," in Asian Language Processing (IALP), 2015 International Conference on, 2015, pp. 53-56.

[18] A. Khan and B. Baharudin, "Pattern and semantic analysis to improve unsupervised techniques for opinion target identification," Kuwait Journal of Science, vol. 43, 2016.

[19] J. Supraja, "A spatial approach to perception identification in editorials enhanced with anaphora resolution," in Nature & Biologically Inspired Computing, 2009. NaBIC 2009. World Congress on, 2009, pp. 421-426.

[20] [M. Z. Asghar and A. Khan, "Sentiment Classification through Semantic Orientation Using SentiWordNet," Life Science Journal, vol. 11, 2014.

[21] K. Dhole and H. Kohli, "Document categorization using semantic relatedness & Anaphora resolution: A discussion," in Research in Computational Intelligence and Communication Networks (ICRCICN), 2015 IEEE International Conference on, 2015, pp. 439-443.

[22] R. Ali, M. Abid, and K. R. Ahmad, "Implementation of the rule-based approach for the resolution of strong personal anaphora in Pashto discourse," in Multitopic Conference, 2008. INMIC 2008. IEEE International, 2008, pp. 501-507.

[23] A. Senapati and U. Garain, "Anaphora Resolution in Bangla using global discourse knowledge," in Asian Language Processing (IALP), 2012 International Conference on, 2012, pp. 49-52.

[24] R. Ali, M. A. Khan, M. Bilal, and I. Rabbi, "Reciprocal anaphora resolution in pashto discourse," in Emerging Technologies, 2008. ICET 2008. 4th International Conference on, 2008, pp. 1-5.

[25] S. Athira, T. Lekshmi, R. Rajeev, E. Sherly, and P. Reghuraj, "Pronominal anaphora resolution using salience score for Malayalam," in Computational Systems and Communications (ICCSC), 2014 First International Conference on, 2014, pp. 47-51.

[26] P. Ning and S. Jun-feng, "The third personal pronoun anaphora resolution in the paroxysmal text of the Chinese web," in 2010 International Conference on Computer Application and System Modeling (ICCASM 2010), 2010.

[27] M. A. Khan and J. A. Nasir, "Distributive anaphora resolution in Urdu discourse," in Emerging Technologies, 2008. ICET 2008. 4th International Conference on, 2008, pp. 38-43.

[28] S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira, "Analysis of representations for domain adaptation," in Advances in neural information processing systems, 2007, pp. 137-144.

[29] K. Toutanova, D. Klein, C. D. Manning, and Y. Singer, "Feature-rich part-of-speech tagging with a cyclic dependency network," in Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1, 2003, pp. 173-180.

[30] K. Khan, B. Baharudin, and A. Khan, "A Review of Unsupervised Approaches of Opinion Target Extraction from Unstructured Reviews," Journal of Applied Sciences, Engineering and Technology, vol. 7, pp. 2400-2410, 2014.

A Novel Multiple Session Payment System

Mohanad Faeq Ali, Nur Azman Abu, Norharyati Harum
Faculty of Information Communication Technology,
Universiti Teknikal Malaysia Melaka,
Hang Tuah Jaya, Durian Tunggal, Melaka, MALAYSIA

***Abstract*—A wireless smartphone can be designed to process a financial payment efficiently. A user can just swipe his/her credit/debit card over the counter and all the processing needed shall be done seamlessly. A smartphone is a popular device to carry around. It is a hassle to carry and keep track on so many physical debit/credit cards in a wallet. An electronic debit/credit card on a smartphone is a more convenient alternative. This research project will embark on an electronic debit/credit card on a smartphone and migrate to an IoT money. A novel session payment system using IoT money has been introduced to minimise debit/credit card risk. The scope of this paper is confined to the security model for an easy payment system based on Internet of Things (IoT). Previously, each IoT money is unique and used once only on one-time payment. The session payment system will ease the burden on protecting the database of the payment system. This paper will extend the use of one-time payment to a multiple session payment system using an IoT money note**.

Keywords*—Easy payment system; internet of things; secure payment system

I. Introduction

Internet of things (IoT) defined as uniquely recognizable object or thing with virtual presentations through internet-like structure. IoT initially proposed in year 1998 [1]. In previous year, concept of IoT became famous through several applications for example smart electrical reading meter, greenhouse condition monitoring, telemedicine communication monitoring, and intelligence transportation. IoT consist of four major components. Components of IoT are for sensing, heterogeneous access, process information, service and applications. An extra component is also needed to cater for security and privacy. Recently, IoT has penetrated subsequent industry applications. IoT helped in terms of cyber-transportation systems (CTS), cyber-physical systems (CPS) and machine-to-machine (M2M) communications [2] [3]. In terms of security and privacy, IoT will encounter severe challenges. The challenges experience from security and privacy because IoT establish the networking with the help of traditional internet connection, mobile data and sensor from network. Next reason of facing security problem because of all 'things' is connected through mobile infrastructure. Lastly, reason facing security problem because of mobile device spreads and communicate faster between one and another. Eventually, latest security and privacy challenges will rise simultaneously at different places in the same time.

IoT technology is potent in the field of e-commerce. IoT has brought improvement in economic growth and provide a competitive element to e-commerce. Even though application of IoT is relatively in early stages but this technology starts becoming established. In order to gain a significant the previous from IoT technology, the current development has moved on to mobile payment systems. A flexible one-time payment note system has been presented [4]. IoT technology application consists of three aspects, namely e-commerce database, payment and logistics. It is crucial to concentrates on the issues for e-commerce's security measures. This one-time note will save the need of protecting the database especially on the credit/debit card information. A more balanced approach shall be presented here for easy and friendly use of the IoT money. This paper will extend the use of one-time payment to a multiple session payment system using an IoT money note.

The IoT system will encompass on specific security issues in E-commerce business [5]. Issues rises from this research on authenticity, integrity and confidentiality from data obtained it IoT payment system should be given attention. Current phase, the autonomous control and ambient intelligence does not belong to original IoT concept. Through advance network technique development, cloud computing and distribution of multi – agent control, shift integration from IoT concept and autonomous control in M2M research exist. Autonomous controls in M2M research serve the purpose to create transformation of M2M in form of CPS. CPS focuses on better interactive application, interaction and real – time distribution in mobile control. Hence, new technology and method must be created to fulfil greater requirements in terms of privacy, reliability and security [6]. The rest of this paper shall be organized as follows. Section II will comes some key issues in cyber security and privacy. A review on online payment systems is given in Section III. Section IV will recapture a proposal on a session of IoT card which can be used once only. The one time card is extended to be on IoT note in Section V for multiple users.an evaluation set of criteria in proposed in Section VI the session IoT card and IoT note against used to evaluate monetary systems. Finally, this paper will give convulsive remarks in Section VII.

II. Key Issues in Cyber Security and Privacy

Internet has evolved from useful tool for research in universities into basic needs. Crime will always happen in the presence of valuable resource obtain through illicit usage of latest technology. The interrelated nature of internet is that resource from internet can be hacked from anywhere around the globe. Cyber security has the duty to overcome hacking issue. Cyber security circles around five keys namely confidentiality, cryptography, authentication, non – repudiation and access. Confidentiality means keep data privately, so authorized users includes machines and human can access into data.

Authentication serves the purpose to verify either data have been tempered with so that data can be sent by authorised owner. Non-repudiation is a key to avoid denial from sender regarding shipment of purchase order from other user. Non - repudiation in some cases considered as unique key, however, in this paper, this key is included as one of cyber security key after authentication process. Access means provide entrance for authorized users to view and tempered with data, computing resources and communications infrastructure.

A survey of information security breaches conducted every year by UK Department for Business, Innovation and Skills with Price Water House Coopers. Result obtained shows increase from 81% in year 2016 to 90% in 2017 regarding the matter cyber breaches experience by big organization. Total of 74% experienced security breach among small sized organization. Forecast from this survey that there will be double – digit growth rate per year. Currently, internet is a need for modern business. Cyber security is a must organization to secure information systems. Nevertheless, as cyber security improved and overcome current issue of breaches, simultaneously, cybercrime transform to more extensive, destructive, sophisticated and comprehension activity.

III. An Online Payment System

The idea of online payment was coming from the previous payment process such as bank checks, travels checks and many orders. Nowadays, there are many online payments such as E-cash, E-purse internet payment system, E-checks, Amazon, ALIBABA and others. However, a typical online payment carries financial security. The previous network communication has been operating E-commerce prior to the arrival of internet of things (IoT) attending security problems such as privacy, identification, authentication, reification, certification, and personalisation. The previous research paper offers an extended design on the one-time payment system. Online payment is a transaction to purchase goods or service paid by human using Internet. Online payment results in lower costs in business because payments made electronically without using physical notes or postage. An online payment helps also to improve customer’s retention because most likely customer would return to an e–commerce site where transaction’s information already been input and kept [7]. In conjunction to an online payment, it is no longer relevant for customer to wait in queue just to make payment because everything is in the fingertip. This section shares various review on various online payments.

Credit Cards: Credit Card is a form of electronic money which can adapt and also used to perform online purchases. Though, people still disagree about simplicity of credit-card transactions because of security concerns, there are still risk where cards and money content were stolen, thus customers fear credit-card fraud from merchants or other parties [8]. Consequently, most of credit card issuers' strong security standards in order to provide fraud protection online.

Europay Mastercard Visa (EMV) is another secure international standard for purchase, fund withdrawal and for authenticating credit and debit card transactions [9]. Master card will manage to use EMV as standard security chip for online payment. At the same time, Visa, Discover, American Express and Europay will also join in and use EMV as global standards.

Virtual Credit Cards: Virtual credit card is an improvement from online credit cards. Virtual credit card is a recent idea provides unique number for users while using same current credit card number. This special number is used to make online transaction for purchases. This new feature allows users to use credit card for online purchase without releasing card number. Users now may give transaction number rather than credit card number to American Express or other merchandise which perform private payment [10].

Debit card: Debit card provides direct cash from personal account to purchase an item. The time duration for fund transfer between account holder and merchant may take 1 to 2 days [10]. The use Debit card is effective in decreasing credit card fraud which will get direct authorization from an account owner.

Smart cards: smart card physically resembles plastic payment card whereby microchip installed as a part on the surface of card. A smart card can carry more information rather than any credit cards that have magnetic strips. Also smart cards can carry anther information such as identification, transportation, banking, health care, and others.in other hand A smart cards can also using for online payment but suppose to used reader to be able to read information of the cards for payment and the secure will be sending data throw internet [10].

E-Cheque: E-Cheque is actually an electronic version of cheque made from paper. E – Cheque contains the same copy of information as cheque made referring to legal framework. The procedure is the same as paper cheque but the advantage is more faster, cheaper and have high security [8]. To use e-cheque, account number is needed together with routing number generate from bank to be keyed in. Financial ponders allow permission to make payment via customer's bank, which either perform electronic funds transfer (EFT) or cheque printing.

Digital Cash: This cash is an example of digital currency. Digital cash allows users to shop online even though the users do not possess debit card. This digital cash procedure is the same as previous practice where people have to reload digital cash account by deposit money to purchase goods or service online. Digital cash is frequently link to another technology called digital wallets [8].

E-Wallets: Electronic wallet is software available in desktop. The users will have to download this software so that user may stores number of credit card and other user’s information. If shop or restaurant accepts e–wallets, the owner of e–wallet will just click and all the formalities were filled up automatically. Currently credit card companies such as MasterCard and Visa offers e–wallet software application [10].

Peer-to-Peer Payments: P2P are type of payments which are growing rapidly because this payment allows two users to perform fund transfer among them [10].

Mobile Payments: Mentioned payments using wireless device such as smartphones supposed to make customer feel convenient, security of payment made electronic increase and transaction fee decrease [11], [12]. Mobile payment system gives ease to business personnel to gain information regarding the customers from the last purchased made. Mobile payment is highly suitable for mobile devices rather than other telecommunication medium due to amazing growth and big penetration to all brands of mobile devices [13], [14]. Mobile payment methods are suitable for offline micropayments as well as for online purchases. This method is a potential attraction for online traders due to an enormous user base of mobile phones. A mobile payment also offers better security and reduces the overall cost for all transactions being made [12]. Nevertheless, mobile payment came across several challenges to obtain significant customer base including inability to perform international transaction and issues related to privacy.

Mobile Wallets: A study regarding consumer usage of mobile wallet has been covered in [16]. A mobile wallet in a smartphone act as leather wallet equipped with digital cards, receipt, coupon and money. Mobile wallet is needed to be installed from online stores in smartphones for the purpose of making purchase either online or offline. Current technology connected the smartphones to QR codes, sound waves, and NFC (Near Field Communication) [15]. These waves and codes basically are solutions which are cloud-based. Mobile wallet is forecasted to give much more convenient payment environment for customers in near future [17].

Touch n Go: Touch ‘n Go is a payment method by using an e-payment prepaid card [18]. This smartcard gives the users a fund anonymous account to make low value cashless payment in easy and comfortable way. Touch ‘n Go is initially designed to pay the toll collection on selected Peninsular Malaysia’s highways. Touch ‘n GO is widely accepted for Common Ticketing Program (CTS) for general public transports located in Klang Region. Later, Contact ‘n Go has been acceptable by car parking operators, in recreational areas and selected retail shops. Considering the convenience of customers, Touch n Go card can be reloaded at a selected of petrol stations, automated teller machines and Automated Reload Kiosks at train stations.

PayPal: PayPal belongs to one of the largest online payment processors worldwide. After growing and create partnerships with E-bay, a big number of online acceptance PayPal as one of accepted methods for payments.

Bitcoins and Other Cryptocurrencies: Bitcoin is a new online currecy created by unknown person in the year 2009 [19]. Bitcoin is also known as cryptocurrency and digital payment system. This cryptocurrency is known recently created by Satoshi Nakamoto. The transaction used is between peer – to – peer whereby the users can perform transaction directly without intermediaries. This virtual currency has anonymity. Bitcoin is known as open – source software starting year 2009. Bitcoin became important needs in marketing covers 90% among all transaction. After Bitcoin is launched, other competitor such as Ethereum, Filecoin, XRP, Gnosis tokens and Tezos emerge causes Bitcoin market share dip below 80% and dive straight until 50% left**.**

Samsung Pay: Recently, a new mobile payment app namely Samsung pay emerged. Samsung Pay is a wallet service provided by Samsung electronics that allows users to perform transaction with other Samsung devices. Samsung pay adapts new secure technology called Magnetic Secure Transmission which allows contactless payments. This contactless payment will be used on payment terminals which support magnetic stripe and normal contactless cards. Samsung Pay supports contactless payments using near-field communications such as NFC and MST. Samsung pay app is available on all Samsung devices, preinstalled or available for download as application update. Users must register to Samsung account with valid credit card. The procurer will verify users fingerprint to authorize any transaction made. Future transactions made will not be necessary to use credit card on any information from credit card. If merchant uses contactless NFC terminals, the user may touch mobile phones to NFC reader to perform transaction [20]. With this method, cashier may input payment details and users will swipe mobile phones at the card-swipe region on the card reader to perform transaction.

IV. A Session IoT Card

A common online electronic payment system via debit/credit card payment system which is enables users to pay for purchases or services online. The system operates on three basic models namely; minimum security model, a third party broker model [21] with a simple encrypted payment system and security electronic transaction model such as SET [22]. Business personnel could misuse customer’s information and make transactions, or owner can temper with the consumer’s site. Information related to consumer can be stolen and misuse by other party. For example, a vendor can make a higher price quotation based on consumer’s previous behaviour. Following are the risk rises based on customer’s view:

a) A consumer can evolve into a competitor who will adapt the prices and strategies learned.

b) A customer could turn up and to be an imposter. They will not produce any bill payment.

c) A consumer has the tendency to become a hacker so that the consumer able to: changing the order requested by customers; changing price; change on available goods; and illegally acquire contact information of customers.

Once a debit/credit card has been used in an online transaction, it becomes vulnerable to be used or abused for another transaction due to anonymity issue [23]. The transaction will be recorded and stored in a database. Since most of the databases are not securely encrypted, they are vulnerable to an open attack such as a ransomware. Other methods including the E-purse and E-check internet payment systems are also vulnerably subjected to the above problem.

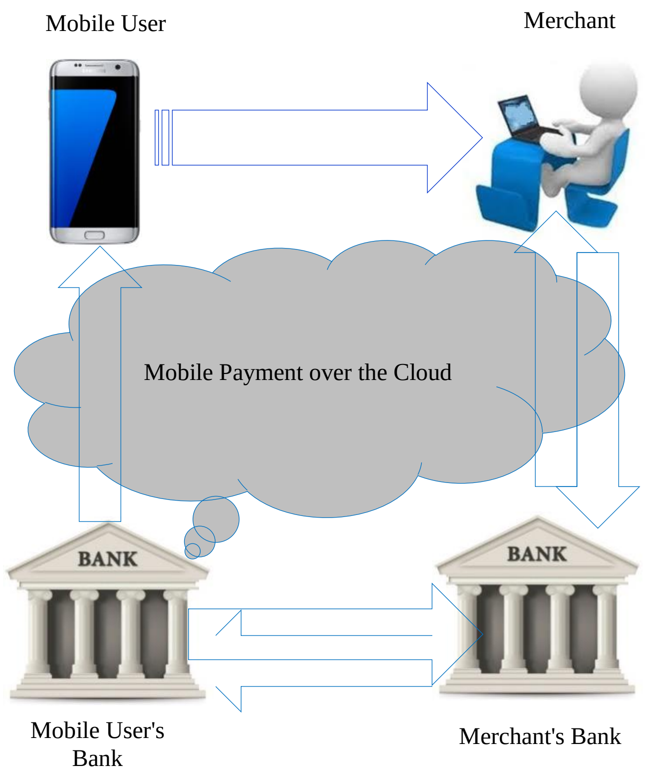


Fig. 1. Electronic cash payment process with E-cash.

A smartphone has become an essential part of life. It is not only an integral part of life but also a source of daily communication. An owner of the smartphone will carry and safeguard the security of the smartphone at any cost all the time. It is more personal to embed a debit/credit card electronically into a smartphone. This paper shall propose an electronic IoT note as part of a credit line a smartphone. A novel IoT payment system shall be introduced to minimise debit/credit card risk.

Fig. 1 illustrates the working of an IoT E-Commerce secure payment mode [24]. This new model will pay special attention to the new card session number. This IoT note will be dynamically changed and updated to the new number once the note is claimed from a user's bank or financial provider. Therefore, it will be a randomly unique number per note which is recognized by an IoT service provider. Each new session card number will also be individually digitally signed by the financial provider [25].

Once an IoT note from a user's smartphone is transferred to the merchant terminal, the payment system will first verify the digital signature of the session number. Once verified, the payment system will check all the transactions being carried out by this note until it is claimed to the merchant's bank. A threshold amount should be set on each IoT session card number. An encrypted update shall be prompted by the financial provider to deliver a new IoT note to the smartphone.

An e-commerce system can be viewed in three different dimensions. The dynamic control used for system upgrade, the real time detection, response and recovery and security coordination between various components.

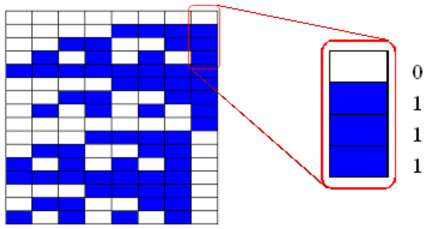


Fig. 2. The top right hand corner represents top right hand hexa value of $0\cdot2^3+1\cdot2^2+1\cdot2^1+1\cdot2^0=7$ in Table I.

TABLE I. A SAMPLE OF AN IOT CARD NUMBER FOR 01 23 45 67 89 AB CD EF 12 34 56 78 9A BC DE F0 WRITTEN IN A STATE ARRAY OF HEXADECIMALS

01	23	45	67
89	AB	CD	EF
12	34	56	78
9A	BC	DE	F0

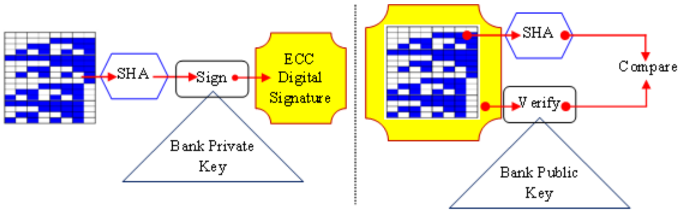


Fig. 3. The Session IoT card number will be signed by its financial provider.

A currently secure session number is 128-bit. It can be viewed as 16 bytes compared to the current 16 digit numbers on a debit credit card. This random Session IoT card number is proposed here as shown in Table I. A sample number is displayed as a state of byte array according the Advanced Encryption Standard (AES) written from left to right along each row of 4 bytes. A direct conversion of binary 2D barcode is generated and shown in Fig. 2. Each hexa has been converted to a column of 4-bit number. This basic 2D barcode can be set an efficient mode of transferring an electronic payment through a smartphone camera.

The research study has proposed new secure technique with a digital signature. Prior to issuing the Session IoT card number, the bank will hash and sign it. The digital signature will be wrapped around the Session IoT card number as shown on the right hand side of Fig. 3. The Session IoT card number will be accompanied by a digital signature. The digital signature must be signed using the private key of the issuing bank as the financial provider. An elliptic curve cryptosystem (ECC) will be light and compact [ECDSA]. A 256-bit ECC would be ideal here to accompany the 128-bit Session IoT card number. Meanwhile a merchant could validate digital signature from using the bank public key and compare to the hashed Session IoT card number.

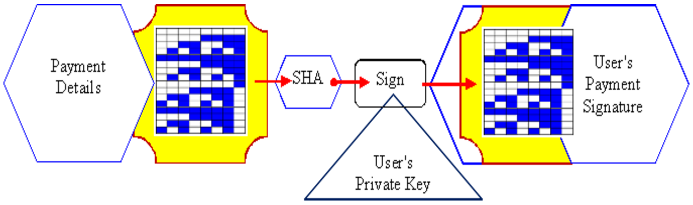


Fig. 4. An online payment on each transaction will be typically signed by the IoT session smart card owner.

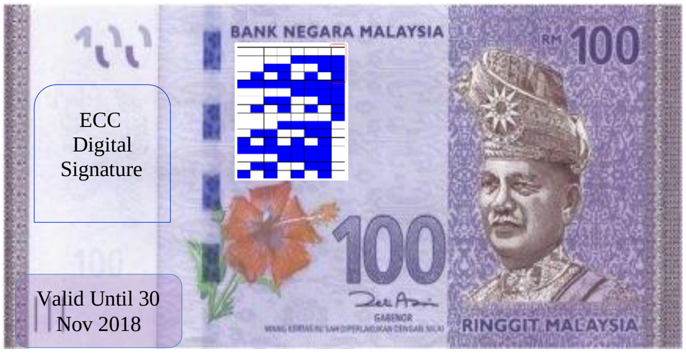


Fig. 5. A nice sample of RM 100 note.

Each payment will also be signed digitally by the user as shown in Fig. 4. Practically, the digital signature will be exercised by a password keyed in by the user. It is imperative the password should not be stored in the smartphone. The password will be used to decrypt user's private key for signature signing. Since each IoT Session number will only carry certain amount, the user cannot spend more than the amount reserved on the number as if it is a currency note. For instances, a note could carry a value of 5, 10, 20, 50 and 100 Malaysian Ringgit. The barcode IoT Session number will also follow the traditional colour of the paper note, i.e. green, red, yellow, turquois and purple, respectively.

The IoT note will also come along with the ECC digital signature as shown in Fig. 5. This note shall be honoured by the first merchant who claims its use once only. This note will also have a validity date on it as written on bottom left corner of the RM 100 in Fig. 5. Typically, it is valid for a month only. A larger value IoT note will have shorter validity period in order to minimise the risk exposure. The user will slide the note to the IoT payment application during a transaction.

The proposed model is based on a smartphone which becomes a mobile intelligent personal terminal for e-commerce businesses. Lightweight PC Tablet act as carrier embedded with RFID reader payment module instead of a physical debit/credit smart card. This mode of payment has the potential to be integrated into an online payment system. This payment application mode is secure and simple. An IoT PDA's payment resolved program which adapts that RFID reader module installed in a smartphone. This IoT program will make the user to avoid pay cumbersome online banking. There is portable handheld personal device and make the whole process completely contactless.

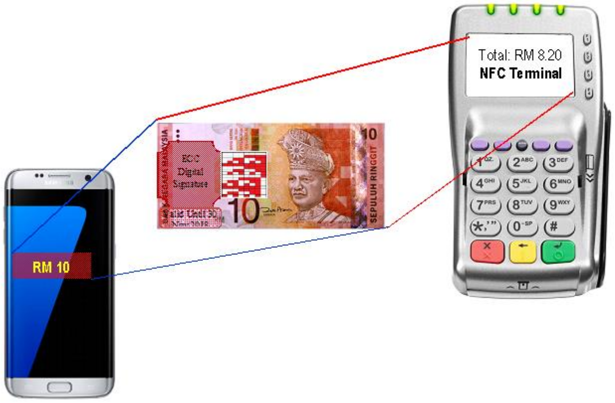


Fig. 6. A user will slide an RM 10 note from his pocket money to an NFC cashier terminal within his smartphone IoT application.

As visualized in Fig. 6, a user may use an RM 10 note from his pocket money to a cashier terminal within his smartphone IoT application to pay for a purchase less than RM 10 for example RM 8.20. The user will sign the transaction for RM 8.20. Thus, the merchant may claim only RM 8.20 from the RM 10 note he/she has received.

This electronic Session IoT card payment will also make use of an online payment tool, for example Alipay and Tenpay using latest IoT RFID contactless technology. In this case, an IoT payment is used during online shopping. A lightweight PC Tablet PC will act as carrier. A payment module installed in the RFID reader where secure and simple smart card payment method achieved by a friendly feature of sliding the note to a merchant iconic application. In IoT handheld payment, all funds were allocated by bank dedicated channel to avoid security risks through open internet. By using the AES algorithm, entire data are encrypted for card users and on data transmission from mobile devices to a clearing centre in order to guarantee maximum security during fund transfer. The bank will maintain a money database to detect double-spending and ensure the validity of this IoT note.

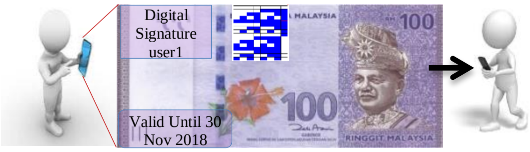


Fig. 7. Transfer money from user1 to user2.

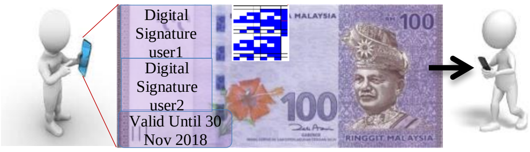


Fig. 8. Transfer money from user2 to user3.

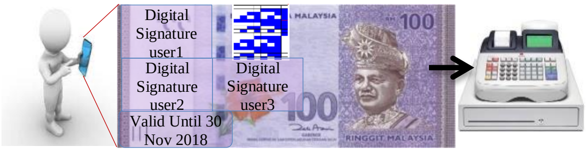


Fig. 9. Transfer money from user3 to merchant.

V. An IoT Note

This paper will propose an easy flow IoT notes as new technique of payment. Traditionally, a user will use his/her debit/credit card to start use IoT session payment [4]. This paper will propose an extension to IoT payment notes to make it transferable between two users and more from one time IoT note. The first user can digitally sign one IoT note and give it to another registered user. The second user can further sign on same note to make a payment to the third or next registered user until the last his/her will give to a merchant to claim the money from the first user's bank account However, the secure will use between first user to second user. After that second user to third user until the last his/her will give to a merchant base on ECC algorithm [26].

In the first case, the IoT notes in digitally signed by the IoT financial provider. Once the IoT note has been issued, the money is already taken out of the first owner account. When the first owner make a payment or pass the IoT note to a merchant or the new owner that shows in Fig. 7, the first owner will encrypt and digitally sing the IoT note to the new. The new owner can then check on the authentic of digital signature by the bank and first owner. The same process will be done by the second owner of the IoT note. When the second owner want to make a payment or transfer the IoT note to a merchant or the third owner that shows in Fig. 8, the second owner will encrypt and digitally sign the IoT note to the new owner. These IoT notes are expected to be in small dominance. The new third owner will check the bank digital signature, the first owner digital signature and the second owner digital signature following the block chain mode. The third user will signed the IoT notes to make transfer or payment to a merchant that shows in Fig. 9.

VI. Evaluation Criteria of Money

This research project will evaluate current online payment systems with IoT card and IoT note. The evaluate membership factors consist of Claimability, Transferability, Recognition, Anonymity, Denomination and Validity Date. Only an IoT note is expected to cover all evaluate factors.

A. Claimable Money

Second evaluation measures an ability of the owner or carrier of electronic money to claim it from the financial provider or bank. The first type of a payment transaction will be between a user to a merchant and then the merchant will be able to claim the money from a bank as shown in the third column of Table II. The second type of claimability will be between users and merchant that means the first user will give the money to second user without any claims from a bank. Also second user will give the same electronic money that he or she received from first user to another user without any claim from the bank. Finally, the end user will give same money to a merchant without claim from the bank. Finally, the merchant will be claimed from the bank.

B. Transferable Money

The third evaluation classifies whether electronic money is transferable or not. The first kind of money transferable will be between users. The second kind of electronic money transfer will be between users without claim from the bank. After that end user will make a payment to a merchant and the merchant will be able to claim the money from the bank as shown in the fourth column of Table II. Non-transferable electronic money can only be used for one payment only.

C. Recognition of Valid Money

The fourth evaluation measure will classify the type of recognition given to the electronic money as a formal money or informal money by the central bank. The first type of recognition of money is accepted by the banking world as in the fifth column of Table I (Yes). Otherwise, the second type of recognition in the case when electronic money is rejected by banking world in the same column of Table II (No).

D. Anonymity

The next evaluation measure is to classify whether the electronic money is attached to the owner or carrier of the money. The money in a financial account belongs to the account holder. The financial provider can check to whom given money belongs to at a given moment. The carrier of the money may claim the ownership of the money without any reporting to the money issuer or financial provider that shows in Table II. Anonymity is an important element of privacy.

E. Denomination

Another element of money is denomination. Paper money always comes along in certain denomination. It is particularly crucial to introduce a fixed denomination on new electronic money in order to be successful and popular in a new electronic payment system. The last evaluation measure check on a fixed stable denomination in each electronic money rather than having an open amount depending on the transaction amount that shows in Table II.

F. Velocity of Money

This paper will also introduce and evaluate electronic monetary payments in terms of their velocity of money. It measures the frequency of use within the monetary payment system. The First type of payment will be used between users to merchants as shown in the second column of Table II. The electronic money can only be used once only as a payment money. The second type of payment will be using between the first user to second user and also from second user to the end user. Finally, the end user will be used to make a payment to a merchant and the merchant will claim from the bank as shown in Table II as multiple payments.

Velocity of money refers to the frequent usage of same currency to purchase goods produced and service available domestically within duration of time (money transfer between an owner to another). Alternatively, it can refer to the frequent of average unit currency utilized for any transaction of changing hands for both purchasing of goods and financial assets.

In this paper, the **velocity of IoT note** shall be used to further evaluate the effectiveness of an IoT note. Indirectly, the frequency of use of an IOT note will be used in measuring speed the money transfer from one holder to the next. In other words, number of times per unit money is transferred or expenses to purchase service and goods per unit note before it is claimed to the bank or financial provider.

TABLE II. COMPARISON BETWEEN TYPE OF MONEY WITH IOT NOTES USING EVALUATION IOT NOTES

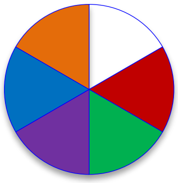
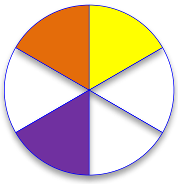
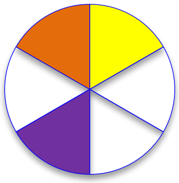
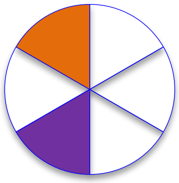
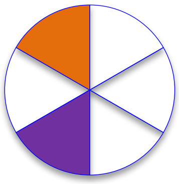
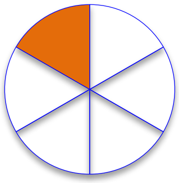
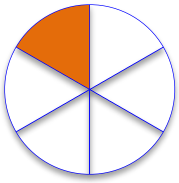
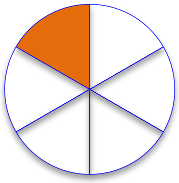
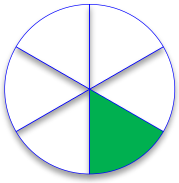
Type of Money	Claimable Money	Transferable Money	Recognition	Anonymity	Denomination	Validity Date	Velocity of Money
Currency	Yes	Yes	Yes	Yes	Yes	No	Multiples
Credit Cards	Yes	No	Yes	No	No	Yes	Once
Virtual Credit Cards	Yes	No	Yes	No	No	Yes	Once
Debit Cards:	Yes	No	Yes	No	No	No	Once
e-Checks	Yes	No	Yes	No	No	No	Once
Digital Cash	Yes	No	No	No	No	No	Once
e-Wallets	Yes	No	No	No	No	No	Once
Mobile Wallets	Yes	No	No	No	No	No	Once
Touch n Go	No	No	No	Yes	No	No	Once
Pay pal	Yes	Yes	No	No	No	No	Multiples
Bitcoins	No	Yes	No	Yes/ No	Yes/ No	No	Multiples
Samsung Pay	Yes	Yes	No	No	No	No	Once
Session IoT card	Yes	No	Yes	Yes	Yes	Yes	Once
IoT notes	Yes	Yes	Yes	Yes	Yes	Yes	Multiples

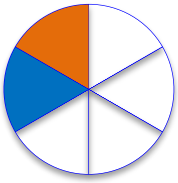
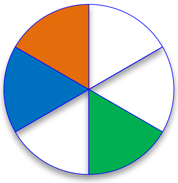
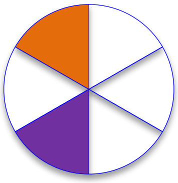
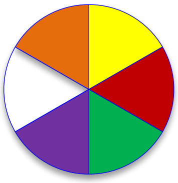
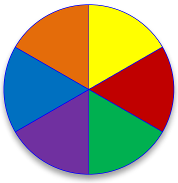
Several online payment systems have been reviewed. A colour scheme as a membership criterion has been given in Tables III and IV.

TABLE III. A MEMBERSHIP EVALUATION OF MONETARY SYSTEMS

Colour	🟧	🟦	🟪	🟩	🟥	🟨
Note	Claimable	Transferable	Recognition	Anonymity	Denomination	Validity Date

TABLE IV. EVALUATION SCORES ON VARIOUS PAYMENT SYSTEMS

Type of money	Velocity of payment	Coverage
Currency	Multiples	
Credit Cards	Once	
Virtual Credit Cards	Once	
Debit Cards:	Once	
e-Checks	Once	
Digital Cash	Once	
e-Wallets	Once	
Mobile Wallets	Once	
Touch n Go	Once	

Pay pal	Multiples	
Bitcoins	Multiples	
Samsung pay	Once	
Proposed Session IoT card	Once	
Proposed IoT note	Multiples	

VII. Conclusion

A traditional banking system is presumably secure and stable. An ease of use has attracted the banking system to online and even mobile. A trade-of between an ease of use on online banking system and a traditional security protocol has to be made. An online banking has been operating outside a secure line at a user's end.

A user may send and make payment directly to a merchant with minimum security protocols. Recently, there are cases of hackers start to attack online banking accounts. The transaction details may be hacked during the process or even later at the merchant database. There is a need to have a new online payment system with minimum information details which can be related back to the original user or account owner.

A one-time note will save the need of protecting the database especially on the credit/debit card information. A more balanced approach has been presented here for easy and friendly use of the IoT money. This paper has extended the use of one-time payment to a multiple session payment system using an IoT money note.

Acknowledgement

The authors would appreciate UTeM Zamalah Scheme. This research study is supported by Universiti Teknikal Malaysia Melak (UTeM), to continue first author's study under UTeM Zamalah Scheme.

References

[1] M. T. Rose, L. H. Stein, N. S. Borenstein, C. M. D. Lowery and E. Stefferud, Computerized Payment System for Purchasing Goods and Services on the Internet, U.S. Patent, Washington, DC: U.S. Patent and Trademark Office, No. 5, pp. 757-917, 1998.

[2] M. N. Al-Mhiqani, R. Ahmad, W. Yassin, A. Hassan, Z. Z. Abidin, N. S. Ali and K. H. AbdulKareem, Cyber-Security Incidents : A Review Cases in Cyber-Physical Systems, International Journal of Advanced Computer Science and Applications, vol. 9, no. 1, 2018, pp. 499-508.

[3] J. Wan, H. Yan, H. Suo and F. Li, Advances in Cyber-physical Systems Research, KSII Transactions on Internet and Information Systems (TIIS), vol. 5, no.11, pp. 1891-1908, 2011.

[4] M. F. Ali, N. A. Abu and N. Harum, A Novel Session Payment System via Internet of Things (IOT), International Journal of Applied Engineering Research, vol. 12, no. 23, pp. 13444-13450, 2017.

[5] T. A. Kraft and R. Kakar, E-Commerce Security, Proceedings of the Conference on Information Systems Applied Research, Washington DC, 2009, pp. 1-11.

[6] M. Chen, J. Wan and F. Li, Machine-to-machine Communications: Architectures, Standards and Applications, KSII Transactions on Internet and Information Systems, vol. 6, no.2, pp. 672-685, 2012.

[7] R. Ding and J. Wright, Payment Card Interchange Fees and Price Discrimination, Journal of Industrial Economics, vol. 65, no. 1, 2017, pp. 39-72.

[8] P. Zhang, Y. He and K. P. Chow, Fraud Track on Secure Electronic Check System, International Journal of Digital Crime and Forensics (IJDCF), vol. 1, no. 2, 2018, pp.137-144.

[9] N. El-Madhoun, E. Bertin and G. Pujolle, An Overview of the EMV Protocol and Its Security Vulnerabilities, IEEE Fourth International Conference on Mobile and Secure Services (MobiSecServ), pp. 1-5, 2018.

[10] E. Turban, J. Outland, D. King, J. K. Lee, T.P. Liang and D. C. Turban, Business-to-Business E-Commerce, Springer, Electronic Commerce, Cham, 2018, pp. 123-166.

[11] J. Urban, Mobile Payments: Consumer Benefits and New Privacy Concerns, SocArXiv, 2016.

[12] C. J. Hoofnagle, J. M. Urban and S. Li, Mobile Payments: Consumer Benefits and New Privacy Concerns, University of California, Berkeley, School of Law, pp.1-20, 2012.

[13] Z. Bezovski, The Future of the Mobile Payment as Electronic Payment System, European Journal of Business and Management, vol. 8, no. 8, pp. 127-132, 2016.

[14] B. U. I. Khan, R. F. Olanrewaju, A. M. Baba, A. A. Langoo and S. Assad, A Compendious Study of Online Payment Systems: Past Developments, Present Impact, and Future Considerations, International Journal Of Advanced Computer Science and Applications, vol. 8, no. 5, 2017, pp. 256-271.

[15] S. Ghosh, J. Goswami, A. Kumar and A. Majumder, Issues in NFC as a Form of Contactless Communication: A Comprehensive Survey, IEEE International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM), pp. 245-252, May 2015.

[16] N. Doan, Consumer Adoption in Mobile Wallet: A Study of Consumers in Finland, vol. 2, no. 6, pp. 816-838, 2014.

[17] T. Husson, The Future of Mobile Wallets Lies beyond Payments, Forrester Research, 2015, pp.127-132

[18] A .S. I. Almselati, R. A. O. K. Rahmat and O. Jaafar, An Overview of Urban Transport in Malaysia, Social Science, vol. 6, no.1, pp. 24-33, 2011.

[19] R.K. Webster, Challenges in Compensating Employees in Cryptocurrencies, Mitchell Hamline Law Journal of Public Policy and Practice, vol. 39, no.1, 2018, pp. 157-182.

[20] K. Cao and A. K. Jain, Hacking Mobile Phones using 2D Printed Fingerprints, MSU Technical report, MSU-CSE, pp. 2-16, 2016.

[21] J. Russell, N. Beitner, O. Dewdney, R. Underwood and W. Jordan, E-commerce Payment System, U.S. Patent Application, no. 09, pp. 810-836, 2002.

[22] R. H. Weber, Internet of Things–New Security and Privacy Challenges, Computer Law and Security Review, vol. 26, no.1, pp. 23-30, 2010.

[23] J. Siegal, S. Rowell and T. Hintz, Method and System for Providing Online Authentication Utilizing Biometric Data, Open Invention Network LLC, U.S. Patent, vol. 9, pp. 146-911, 2018.

[24] G. G. Si, X. Zhao, J. Wang, X. Long and T. Hu, A Novel Mutual Authentication Scheme for Internet of Things, Proceedings of IEEE International Conference on Modelling, Identification and Control (ICMIC), pp. 563-566, June 2011.

[25] A. Abdollahi and M. Afzali, A Single Sign-on based Integrated Model for E-banking Services through Cloud Computing, International Journal of Advances in Computer Science and Technology, vol. 3, no.1, pp. 34-38, 2014.

[26] J. W. Bos, J. A. Halderman, N. Heninger, J. Moore, M. Naehrig and E. Wustrow, Elliptic Curve Cryptography in Practice, International Conference on Financial Cryptography and Data Security, Springer, pp. 157-175, March 2014.

Real-Time Concept Feedback in Lectures for Botho University Students

Alpheus Wanano Mogwe
Library & Information Studies | Faculty of Computing
University of Botswana | Botho University
Gaborone, Botswana

Abstract—**This is a mixed methodology study which focused on developing a real-time concept feedback system for Botho University students. The study takes advantage of the tablets distributed freely by the institution to ameliorate the problem of lack of understanding of module concepts during lecture lessons. The system addresses issues of providing real-time feedback as the lecture is ongoing without disturbing other students, thus upholding effective class participation and interaction without the need of voicing own concerns loud to the lecturer, and in turn the lecturer is able to view the students' interactions and address them. The real-time concept feedback system was used to test student comprehension of concepts, improve participation, engagement and attendance. The study identified many factors affecting students' participation and interaction in a traditional class which inhibits understanding of concepts; hence, the development of the application to address such. It concluded that real-time concept feedback systems are vital in addressing students understanding in lecture sessions, thus upholding the importance of ICTs in education.**

Keywords—Real-time feedback systems; interactive technology; e-learning; information technology; understanding of concepts

I. Introduction

With Information and Communication Technologies (ICTs) having strong repercussions in the education and learning sector around the world, Botho University has become the first institution in Botswana to provide free tablets for learning and bridge the digital gap between its students. In a knowledge based economy, education is a requirement and in this modern era it comes with need for knowledge and interaction with ICTs as envisioned through the efforts of Botho University of providing a free tablet to each and every Botho student, a gesture which goes a long to facilitate the teaching and learning process. This is so, because ICTs are exceptionally a commanding tool for information and knowledge diffusion – the primary characteristic of the whole education course. Therefore, ICTs play a pedagogic duty complementing the traditional delivery of education and learning.

With these tablets, the e-learning environment has been improved and brought to higher standards as all the students in the campus own a tablet, are able to access online resources including online databases, repositories and books. This has made learning easier, and led to this mixed methodology research to measure students' understanding in a class set up using their own tablets whilst interacting with a mobile web application developed to address their concerns.

The study identified a major problem of lack of understanding of concepts during class sessions during an ongoing lecture. This is compounded by lack of interaction from other students who would have not understood the lecture concepts but carry on for some various reasons such as being shy to participate in the midst of others or the traditional mode of delivery itself [1] where the lecturer would call for volunteers or handpick own choice of students. Moreover, lack of immediate feedback whenever they fail to understand concepts, as feedback is a key component to successful learning [2], [3] attributes to the same problem. This leads to lack of understanding which eventually leads to failure. Thus, the study sought to understand how real time concept feedback in lectures could be incorporated in the learning environment of Botho University using the already available tablets devices to ameliorate the lack of understanding which happens during lecture sessions whilst being inclusive of all students irrespective of them being fast or slow learners in their comfort zones-seats. The research question of the study was "How can understanding of concepts in a class be improved in real time?"

The study focused on the following objectives to drive the main purpose identified above:

- To analyse students understanding of concepts in a lecture session and how they communicate their understanding or lack of understanding.
- To establish students participation and interaction in a lecture session and how it affects understanding of concepts in a lecture.
- To ascertain students behaviour when they fail to understand concepts taught during a class or lecture session and how that impacts on the rest of the lecture and module coverage.
- To identify real-time concept feedback systems with intent on understanding their operations, strengths and weaknesses for aid in development of a real time concept feedback system.

Through these objectives, a clear understanding of various characteristics related to students, learning, teaching and understanding of concepts in a lecture session was articulated to fully address the main problem of lack of understanding in a class and how real time concept feedback in lectures could be incorporated in the learning environment. This is important before considering how real time concept feedback systems could actually revolutionize the learning environment as education is not a field which easily lends itself to experimentation but works on a codified manner. Therefore, a further analysis of this subject area was supplemented through secondary research involving various literatures on the same.

II. Literature Review

Education delivery is a foundation of teaching and learning. It is more than coming into a class and start talking, but an intricate situation involving the lecturer (instructor) and the learners (students). One factor attributing to the complexity of teaching and learning is the different paces of learning in each and every student and charging through a syllabus disregarding those paces and students diversity in learning is a peril recipe for students' failure both in terms of assessment and learning outcomes [1], [4]. And this is due to understanding of the concepts being taught in a class or lecture session. Due to the diversity of students' paces in learning, their understanding tends to vary as others would be fast learners whilst others are slow or moderate learners, which could be a misled point to the lecturer or instructor to believe majority understand or do not understand [1].

Understanding of this concepts becomes key for them to navigate through the whole lecture session which either leads to participation, interactivity and a sense of belonging or vice versa of all the mentioned. Therefore, real-time feedback is a key to achieving the positive of that. Without receiving regular and real time feedback, it becomes difficult for them to understand and or comprehend materials being taught [5]. Various authors have attributed that real-time feedback is an important aspect in the learning environment as it helps the instructors to be more responsive to learners and assist them in a timely manner [2] not only to notice shortfalls of the students after the assessment component [4].

A. *Personalizing Learning for Real-Time Feedback*

One way to achieve real-time feedback is being able to personalize learning for each and every student to address their pace diversity. The moment a lecturer steps into a class environment and start the lecture session, he or she is inundated with a lot of questions, and those are mostly related to (1) Are the students understanding the lecture (2) Are the concepts clearly modularized for clear understanding (3) Is the lecture relatable. All these questions are only answered through students' feedback and in traditional delivery mode, no emphasize is per individual student but generalization is employed. Thus, a majority affirmative response paves the way for the lecturer to continue teaching [1], [4] whilst ignoring the silent pleas of those who would have not understood but shied away to voice their concerns due to various reasons.

This in turn affects the lecture or the overall module or subject area, though timely feedback may be incorporated later to understand how they are doing in the subject. Timely feedback becomes a problem as it wouldn't be carried often thus the gap left to be addressed through the use of real time feedback. With students having tablets in Botho University, real-time feedback systems become a favorable option to personalize learning for students. With real time feedback systems, such concerns are addressed as students interact with systems individually thus able to voice out concerns without being muzzled by the voice of many. According to Cialdini (2008), anonymity is one aspect which students want to maintain as a way of not embarrassing themselves and it gives them confidence to introspect own level of understanding without fear or any form of victimization learning [2].

B. *Real-Time Feedback Systems Enhances Students Self-Regulation Skills*

Sitzmann et al. [5] stipulates that real-time feedback systems enhance students' self-regulation skills. Students are in a better position to learn how to evaluate their own learning, thus being able to gauge if they are on the right track and have a good understanding of the concepts being taught, thus influencing their behaviour [6], [7]. Unlike, having to wait for consultation after class, which many students hardly do, utilising real-time feedback systems gives students a sense of control over their learning during the lecture and empowers them to silently query the lecturer whenever they get lost in understanding of the concepts.

In addition to evaluating their own understanding, this also helps to evaluate the lecturers' teaching and delivery mode in real time, and both the students and lecturer can quickly make amends or appropriate changes whenever they were off track in either understanding or delivering the concept. On the course of enhancing the self-regulation skills, the relationship between the lecturer and the students is strengthened for the benefit of the students to understand the learning material instead of having lackluster assessments or disappointing results at the end [1]. Thus, with real-time feedback systems a clearer idea of student understanding is captured well in time and this improves their self-regulation skills.

C. *Real-Time Feedback Systems Increase Students Classroom Presence*

Real-time feedback systems increase students' classroom presence, interaction and participation [8], [9]. With students noting that their needs are being addressed, their attendance also improves and builds the confidence of interaction and participation in the class [10]. Authors in [11], [12] notes the importance of personalised response systems and diverse learning styles in education and how this affects their engagements in class, attendance, motivation and participation in a class [13]. Through this online feedback, information needed to improve certain areas would be addressed whilst assisting in improving learners' motivation and learners' ability to reflect on own learning. Moreover, increased classroom presence, interaction and participation would lead to increased student performance. Studies have indicated that real-time feedback systems also improve student performance although they fail to identify the causes of such improvements [2], [14].

III. Methodology

The research study followed a mixed methodology approach. Understanding of concepts during teaching and learning is one area which hasn't been well explored within the Botho University environment, and the author sought to understand how this could be addressed and how real-time concept feedback in lectures for Botho University Students could be utilised to ameliorate the understanding of concepts in class sessions. The research was conducted within the Botho

University environment with questionnaires distributed and interviews done on selected individuals.

The mixed methodology approach was selected as the best approach to understand fully and in-depth on this area and offer conclusive analysis and suggested solutions to the identified problem of students failing to understand learning concepts during class and how this situation can be mitigated through the use of real time concept feedback systems in lectures. Exploring this untapped area would increase awareness of shortfalls of timely feedback versus real-time feedback whilst determining the cause nature of the problem and possible ways which can be utilised to solve the problem [15]. A mixed methodology approach helps understand the situation from all angles for better conclusions [16]. This then, forms the basis of the conclusive research, with alternative solutions [17]-[19].

Population for the study and sampling procedure are also important. The population of this study was made up of student and lecturers of Botho University. A population of study is important to validate the target audience, and Burns and Grove [20] defines a population as all elements inclusive of objects, individuals and events that meet the sample criteria for inclusion in a study. A questionnaire was distributed to 30 students. The questionnaire administration and data gathering followed the Exponential discriminative snowball sampling technique.

IV. Results

The purpose of the study was to assess how understanding of concepts in a class can be improved in real-time, as lack of understanding of concepts during ongoing lecture sessions was identified. It sought to understand how real time concept feedback systems in lectures could be incorporated in the learning environment of Botho University using the already available tablets devices to ameliorate the lack of understanding which happens during lecture sessions. To do that, it was imperative to carry a questionnaire data gathering to understand the students and their understanding capabilities and behaviours and conduct interviews on selected individuals. This questionnaire was administered following the exponential discriminative snowball sampling technique, where a subject identified would attempt the questionnaire and refer multiple referrals to do the same too. A selection of the subjects who met the aim and objectives of the study would then follow, and interviews would be conducted on few selected to iron out inconsistencies. The following is the results of the study summarized as per the study objectives.

A. *Objective 1: To analyse students understanding of concepts in a lecture session and how they communicate their understanding or lack of understanding.*

1) Understanding of Concepts: This question sought to understand if students easily understood concepts in a lecture session.

From Fig. 1 (refer to appendix) it can be noted that 20% of the respondents indicate they always understand concepts when taught, whilst 37% indicate that sometimes they do understand the concepts, 27% indicate that they rarely understand concepts with 13% saying they never understand and 3% noting that they don't know if they understand concepts or not. This indicates a similar pattern identified during various studies that understanding of concepts by students varies [1], [4] and to comprehend materials taught becomes difficult if understanding of concepts is lacking [5].

2) Relating Concepts Understanding to Learning Outcomes: The question sought to understand if respondents would relate understanding of concepts during lecture session to learning outcomes of the module.

Respondents at 30% indicated that they are able to relate concepts to learning outcomes whilst the other 30% noted that sometimes they do relate the two. 33% noted that they rarely relate learning concepts to learning outcomes and 7% said they never relate the two. For those who never relate the two, the issue could be tied to lack of comprehension of the concepts, thus making it difficult to follow on the concepts taught and relating [5]. Refer to Fig. 2 in appendix.

3) Do you communicate lack of Understanding: the question sought to understand I students do voice out their concerns when they do not understand any concepts during a lesson.

The results indicate that 33% communicate their lack of understanding of concepts with 10% noting that sometimes they do. 40% of respondents noted that they rarely communicate their lack of understanding and 17% never do so. This shows that only few students are able to communicate whenever they understand or do not understand, whilst majority do not communicate (refer to Fig. 3 in appendix) and this could be attributed to the traditional mode of delivery which is rigid as attributed by Caldwell [1].

4) What Holds Communication Amongst Learners

Fig. 4 in appendix indicates ccommunication is vital to get feedback, and 30% of respondents noted that they are free-spirit respondents who always communicate their lack of understanding, whilst other respondents noted that they are various factors which holds them from communicating their lack of understanding ranging from uncertainty of whether to communicate at 20%, others are naturally quite at 7% and shyness was blamed for lack of communication at 43%. According to Cialdin (2008), many factors contribute to students remaining silent on communicating whenever they do not understand, similar conclusion reached by other authors [1], [2] and [4].

B. *Objective 2: To establish students' participation and interaction in a lecture session and how it affects understanding of concepts in a lecture.*

1) Do You Normally Participate in a Class?

Respondents as indicated in Fig. 5 (refer to appendix) they do participate in a class at 43%, with 17% saying sometimes they participate and 37% noting that they rarely participate and 3% saying don't participate in a class. Lack of participation has been found to be a major setback of making students to easily understand concepts and this could be attributed to the traditional model itself [1], [3] and lack of technology to promote inclusiveness policy for all students where anonymity would be upheld [2].

2) Do You Interact with Other Learners and Your Lecturer?

Many learners indicate at 50% that they do interact with other learners and the lecturer in a class, with 10% noting that sometimes they do participate and 40% saying they rarely participate. Interaction is a vital tool in accessing feedback and lack of it impacts negatively in understanding of concepts and getting relevant feedback [5] as indicated in Fig. 6 in appendix.

3) Does Participation and Interaction Influence you Understanding?

Respondents agree at 40% that participation and interaction influences understanding of concepts in a class, with 30% saying sometimes this two attributes could influence understanding of the concepts whilst 13% say this two rarely influence. 17% of the respondents noted that they don't know if these two attributes influence understanding of concepts (refer to Fig. 7 for full results) Self-regulation is one aspect which has been touted to be improved through the use of real-time feedback systems, and this is attributed to participation and interaction [6], [7]. From Fig. 7 it shows that lack of such systems impacts negatively to student self-regulation.

C. Objective 3: To ascertain students behaviour when they fail to understand concepts taught during a class or lecture session and how that impacts on the rest of the lecture and module coverage.

1) What happens when you fail to understand lecture concepts & outcomes? The question sought to understand the behaviour of students in relation to lack of understanding concepts during a lecture session.

It was noted that the traditional mode of delivery has come as a muzzling approach to slow learners which leads to them not to communicate or participate when not understanding [1] [4], thus this in a way influences the behaviour exhibited by student when they fail to understand concepts, as indicated in Fig. 8. At 23%, respondents noted that they get motivated to further ask questions whenever they fail to understand concepts, whilst 30% indicated the opposite. 23% of the respondents noted that they ask their fellow students whilst 13% ask the lecturer and 10% noted that they just don't do anything. All the behaviours exhibited here are linked to the diversity of paces in learning, as more vocal and fast learners are able to exhibit behaviours of asking in a class and not giving up [1].

2) How that does impacts your whole lecture and module? Upon understanding how students behave whenever they don't understand concepts, it was also relevant to appreciate how their actions on the previous question impact their whole lecture session and consequently the module area.

Lack of understanding can have a negative impact on a student during the lecture, or an overall negative impact on the module itself. This is only possible to be noticed at the timely feedback intervals which sometimes fall after assessments [1] [4] thus not being real time and becomes difficult to amend and correct. 37% of the respondents as indicated in Fig. 9, noted that they try by all means to understand whilst 30% never gives up on understanding the concepts. 20% of the respondents noted that they give up, with 13% noting that sometimes they do give up. It is within the groups that try to understand including those who give up that concentration needs to be channeled to and introduce other forms which could help bridge the understanding gap, and real-time concept feedback systems become handy for utilisation.

V. Conclusion

The study, as indicated by the literature review and the students themselves, concluded that understanding of concepts is clearly a problem amongst students. This lack of understanding of concepts during lecture sessions is further compounded by the traditional way of delivery which constrains students and hampers them to communicate their lack of understanding. Participation and interaction during classes have also been found to be low in many students, which also contribute to lack of understanding. The students' behaviour is also a contributing factor as students easily give up on following whenever they are not understanding due to the systems in place which centralizes mostly onto the traditional approach of delivery, thus need of timely diagnostic feedback concept systems which students have come to appreciate [21], [22].

The researcher concluded that a real-time concept feedback system is essential in the Botho University environment as the technology is available for utilisation. Real-time feedback is paramount in learning, as students can easily track their learning and provide feedback in real-time for more clarifications when they fail to understand. The researcher concluded by developing the real-time concept feedback application, but due to time constraints, on writing of the findings it has not been

- Piloted to test the students perceptions on the utilisation of the product.
- Piloted to test the lecturers on the utilisation of the same.
- It has not been fully tested and evaluated how it measures to address the gap identified during the problem statement.

The researcher concludes that more work needs to be done on this area and alternative ways found to be utilised to address the issue of understanding in class, with the aid of using real-time concept feedback systems both for the benefit of the lecturer and the students.

Acknowledgment

The researcher thanks all the Botho University community and participants of the study.

References

[1] Caldwell, J. (2007). Clickers in the Large Classroom: Current Research and Best-Practice Tips. Life Sciences Education, 6, 9-20.

[2] Hedgcock, W.H and Rouwenhorst, R.M. (2014). Clicking their way to success: using student response systems as a tool for feedback. Journal for Advancement of Marketing Education, Volume 22, Issue 2, Fall 2014

[3] Martin, Florence , James Klein, and Howard Sullivan (2007), "The Impact of Instructional Elements in Computer -Based Instruction," British Journal of Educational Technology , 38(4), 623-36.Cialdini, Robert B.(2008),Influence: Science and Practice,New York: Pearson

[4] Sun, L., "The Use of a Real Time Online Class Response System to Enhance Classroom Learning," ASEE Engineering Design Graphics Division - 69th Midyear Conference, Normal, IL, October 12-14, 2014.

[5] Sitzmann, Traci, Katherine Ely, Kenneth G. Brown,and Kristina N. Bauer (2010), “Self-Assessment of Knowledge: A Cognitive Learning or Affective Measure?”Academy of Management Learning & Education, 9(2),169-191.

[6] Edens, K. M. (2009). The interaction of pedagogical approach, gender, self-regulation, and goal orientation using student response system technology.Journal of Research on Technology in Education, 41(2), 161- 177.

[7] Carver, Charles S.,and Michael F.Scheier(2000), “Scaling Back Goals and Recalibration of the Affect System are Processes in Normal Adaptive Self-Regulation: Understanding ‘Response Shift’ Phenomena,”Social Science & Medicine , 50,1715-1722.

[8] Ivo, N., Mitko, Z., Dragan, B. and Thanos, H., “Designing a Mobile Clicker System for Educational Purposes”. 2013. The 9th International Scientific Conference.

[9] Stowell, Jeffrey R., and Jason M. Nelson (2007),“Benefits of Electronic Audience Response Systems on Student Participation, Learning, and Emotion ,”Teaching of Psychology, 34,253-58.

[10] Siau, K., Sheng, H., and Nah, F. (2006). Use of a Classroom Response System to Enhance Classroom Interactivity. IEEE Transactions on Education, 49(3), 398-403.

[11] Gauci, S. A., Dantas, A. M., Williams, D.A., & Kemm, R.E. (2009). Promoting student-centered learning in lectures with a personal response system. Advances in Physiological Education, 33, 60-71

[12] Bajraktarevic, N., Hall, W., & Fullick, P. (2003). Incorporating learning styles in hypermedia environment: Empirical evaluation. In P. de Bra, H. C. Davis, J. Kay, & M. Schraefel (Eds.), Proceedings of the workshop on adaptive hypermedia and adaptive web-based systems (pp. 41-52). Nottingham, UK: Eindhoven University.

[13] Tietz, Wendy (2005), “Using Student Response Systems to More Fully Engage Your Students,” paper presented at the Colloquium on change in Accounting Education, Long Beach, California, October 26-29.

[14] Fies, C. & Marshall, J. (2006). Classroom Response Systems: A Review of the Literature, Journal of Science Education and Technology, 15 (1), 101-109.

[15] Saunders, M., Lewis, P. & Thornhill, A. (2012) “Research Methods for Business Students” 6th edition, Pearson Education Limited

[16] Brown, R.B. (2006) “Doing Your Dissertation in Business and Management: The Reality of Research and Writing” Sage Publications, p.43

[17] Singh, K. (2007) “Quantitative Social Research Methods” SAGE Publications, p.64

[18] Sandhusen, R.L. (2000) “Marketing” Barrons

[19] Nargundkar, R. (2008) “Marketing Research: Text and Cases” 3rd edition, p.38

[20] Burns, N., & Grove, S. K. (1993). The practice of nursing research. Boston: Elsevier Health Sciences

[21] Lin, J.-W., Lai, Y.-C., & Chuang, Y.-S. (2013). Timely Diagnostic Feedback for Database Concept Learning. Educational Technology & Society. 16. (2), 228–242. 228

[22] Parvez, Shahida M..A pedagogical framework for integrating individual learning style into an intelligent tutoring system. ProQuest Dissertations Publishing, 2008.

APPENDIX

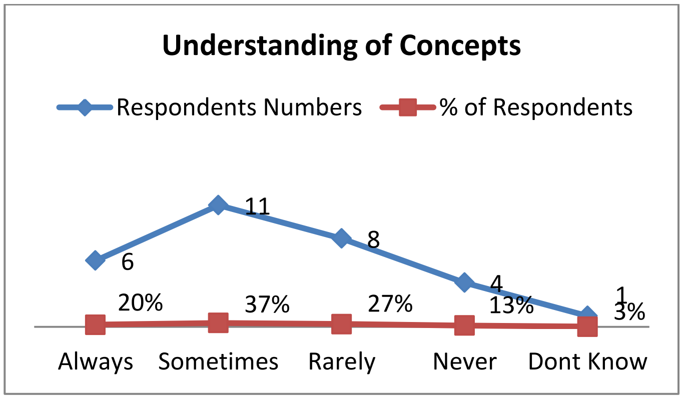


Fig. 1. Understanding of concepts.

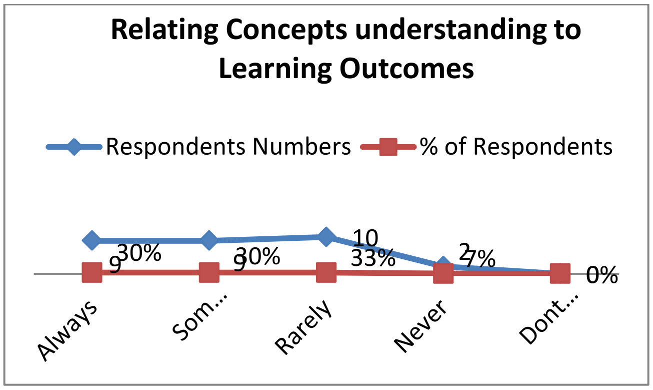


Fig. 2. Relating concepts understanding to learning outcomes.

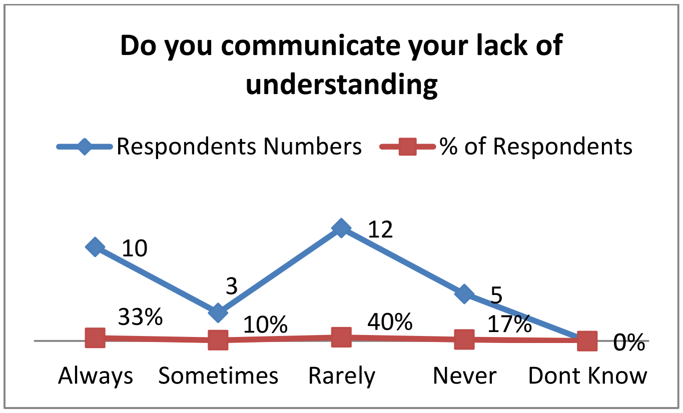


Fig. 3. Communication of lack of understanding.

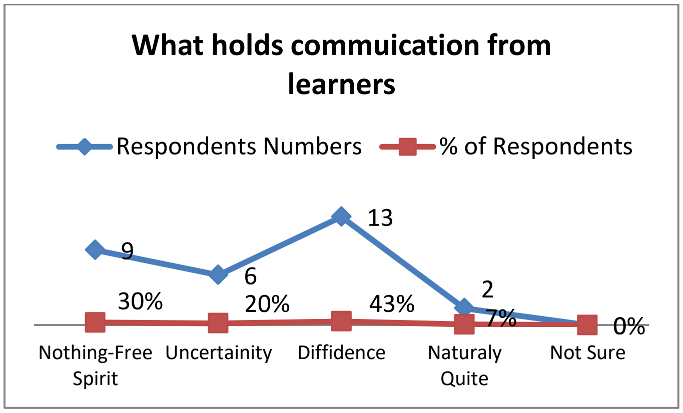


Fig. 4. Factors holding learners from communicating.

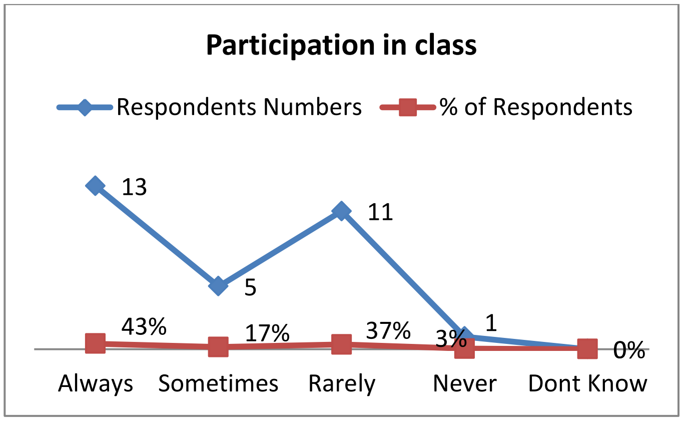


Fig. 5. Participation in class.

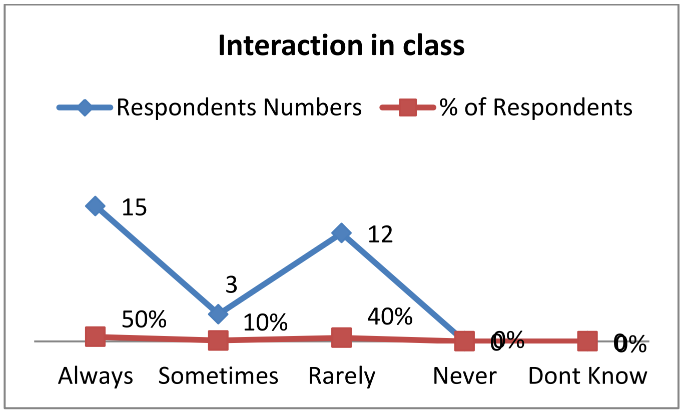


Fig. 6. Interaction in class.

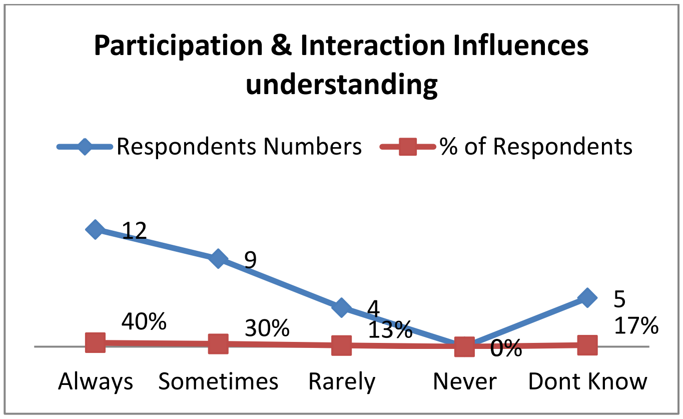


Fig. 7. Participation & interaction influences understanding.

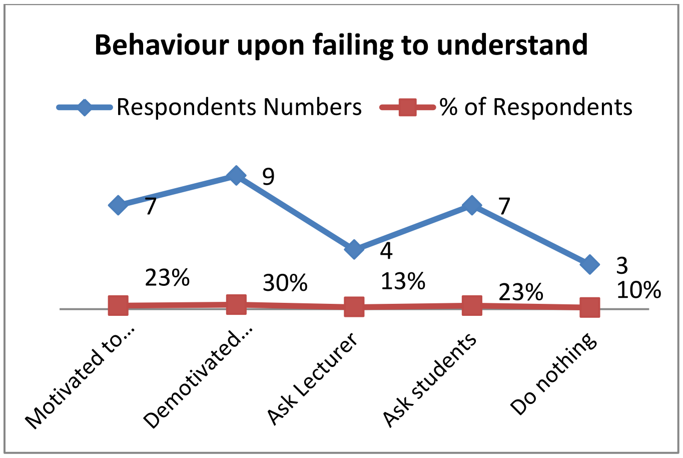


Fig. 8. Behaviour upon failing to understand.

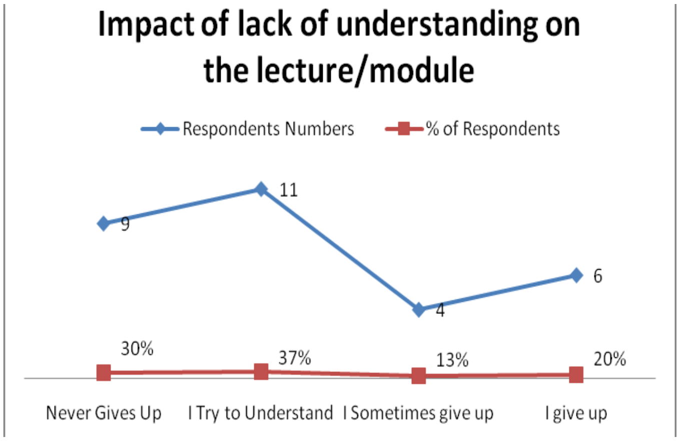


Fig. 9. Lack of understanding Impact on lecture/module.

Swarm Optimization based Radio Resource Allocation for Dense Devices D2D Communication

O. Hayat
1.NUML H-9 Islamabad, Pakistan.
Wireless Communication Center
(WCC), Faculty of Electrical
Engineering, Universiti Teknologi
Malaysia

R. Ngah
Wireless Communication Center
(WCC), Faculty of Electrical
Engineering, Universiti Teknologi
Malaysia

Siti Z. Mohd Hashim
Big Data Centre,
Faculty of Computing,
Universiti Teknologi Malaysia

***Abstract*—In Device to Device (D2D) communication two or more devices communicate directly with each other in the in-band cellular network. It enhances the spectral efficiency due to cellular radio resources (RR) are shared among the cellular users and D2D users. If the RR sharing is not legitimate properly, it causes interference and inefficient use. Therefore, management of RR between cellular users and D2D users is required to control the interference and inefficient use of RR. In D2D enabled cellular network, D2D users have a good signal to noise ratio (SNR) compared with cellular users due to the short distances and dedicated path. Using this advantage, an efficient RR allocation algorithm based on swarm optimization is proposed in this paper, that allows utmost spatial reuse in multi-users and OFDMA networks. The algorithm determines the required RR on the request of D2D users following the indicator variable. It enhances the capacity (Bit/Hz), overall system throughput and spectral efficiency with respect to sub-carriers in OFDMA networks. The performance of the proposed algorithm is evaluated via MATLAB simulations.**

Keywords*—Device to device (D2D) communication; radio resources (RR) allocation; OFDMA networks; sub-channels and sub-carriers; cellular users and D2D users

I. Introduction

The Device to Device (D2D) communication in fourth-generation long-term evolution (4G LTE) focuses on public safety, but the potential advancements that can be given by D2D operation are not completely exploited yet [1]. D2D communication as an underlay to cellular system is viewed as one of the key advances for improving the performance of upcoming cellular systems. In 5G systems, it is anticipated that D2D operation will be locally coordinated as a component without bounds the 5G system. The fundamental potential gains by D2D including, capacity and throughput, low latency, availability and reliability and proximity services. All these gains can be achieved only using efficiently resources allocation and utilization. Collectively it is called radio resources (RR) allocation for D2D communication. In a cellular system, multiple devices exist with multiple services and operators. When many devices qualify for D2D, then who will provide the resources to accomplish D2D communication. It includes data channel, control channel and other cellular services without affecting the cellular users. The RR allocation in OFDMA cellular network has three scenarios i) cellular users to D2D users ii) D2D users to cellular users iii) D2D users to D2D users as presented in Fig. 1.

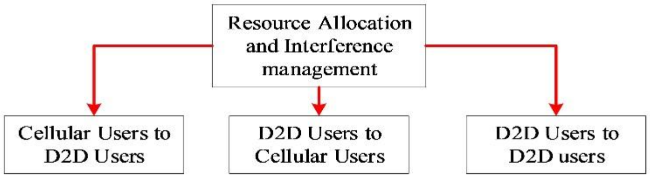

Resource Allocation and Interference management
Cellular Users to D2D Users
D2D Users to Cellular Users
D2D Users to D2D users

Fig. 1. Scenarios for radio resources allocation among D2D users and cellular users in the cellular system.

The fundamental thought of D2D communication is that suitable selected devices reuse the cellular resources to set up direct communication links [2]. Given conditions are that the D2D communication does not put an adverse effect on cellular users like interference and cellular users have right to use the resources freely. Despite its awesome potential in coverage and capacity, it has some challenges particularly RR allocation. The essential thought is to reuse cellular resources by enabling adjacent wireless devices to build up direct communication links. This idea does not just enhance the proficiency of spectrum utilization, yet additionally has an extraordinary potential for upgrading the system performance articulated in terms of system capacity, throughput, spectral efficiency, and end-to-end delays. There are two approaches for RR allocation: half duplex and full duplex. Conventionally an user equipment is equipped with a single antenna, therefore two orthogonal time stages are needed for individual transmission and reception in half-duplex. In first time stage, all users equipment should keep silent and listen from the base station on the downlink channel. In second time stage each device request for resources as cellular users or D2D users. Although this approach will not cause interference between cellular users and D2D users while degrading the RR reuse gain. To overcome this deficiency, full duplex OFDMA is an alternative and allows multi-users to use the same RR simultaneously [3].

To organize the system controlled D2D communication as an underlay to the cellular system, a network planner faces few difficulties, which mostly arise because of the absence of consistent channel information at the base stations. Efficient feedback is significant to get channel information. The channel information for cellular users at the serving base station can be acquired efficiently. Conversely, such information is not accessible for D2D channels. The reason is the division of the control plane from user/data plane because of the system controlled D2D communication. A quick outcome of this division is that D2D users can't specifically use pilot signals communicated by the base station in contrast to cellular users for estimation of D2D channels. Additionally, local transmission of the individual pilot signal by every D2D users is not possible and would not tackle the issue because of pilot contamination. Since techniques for overwhelming pilot contamination in D2D scenarios experience the ill effects of the requirement for expanded feedback and control overhead. Various formulations have been proposed for RR allocation, for example, proportional and max-min fairness, inelastic traffic, weighted fair queuing and convex optimization techniques [4].

D2D and RR allocation both are state of art and future research challenges. The emphasis is on D2D situations, for example, situations with normally low mobility where data offloading, improvement of network capacity, reduced latency and enhance data rates play a leading role. The attention will be on in-band underlay D2D communication, in which D2D utilizes similar resources of the spectrum from the cellular network. It is sensible to expect that RR allocation to D2D users must be accomplished in a distributed manner under entirely restricted channel information. In addition, it is of most extreme significant that immediate transmissions among devices are coordinated to guarantee that they don't detrimentally affect the performance of cellular users. Such coordination must include a cautious power allocation of D2D users to available RR, essentially utilized as downlink or uplink. This issue, which is hard to understand even in a centrally controlled system, is additionally provoked in a D2D setting by the requirement for distributed arrangements. Therefore, RR allocation model for multi-devices in OFDMA system is proposed for high data rate, energy efficiency and interference avoidance between cellular users and D2D users. With the D2D pair establishment, RR can be allocated to that pair for communication. After discovery, as discovered device receives a request for connection, RR is thus allocated to discoverer devices only. It allows the D2D pair to transmit and receive data over the same allocated channel. Swarm optimization is applied for RR allocation to minimize the interference between cellular users and D2D users. It enhances the system capacity, throughput, and frequency efficiency.

Rest of the paper is organized as follow: Section 2 explains the Radio resource allocation techniques for D2D users and the radio resource allocation model and results are discussed in Section 3. In the end, the paper is concluded in Section 4.

II. RR Allocation Techniques for D2D Users

There are two types of RR allocation techniques in in-band D2D communication: underlay and overlay as described in Fig. 2. The expansion of the D2D layer as an underlay to cellular systems postures new difficulties in term of interference control compare with ordinary cellular networks. RR allocation for D2D in underlay cellular network is proposed [5] based on joint scheduling. It controls the power to avoid the interference and maintain the QoS of D2D link, but the problem is accommodation of maximum users is quite difficult. RR allocation in mobility structure for underlay D2D is presented in [6] in which, RR are apportioned based on distance. When the distance is increased, channel allocation becomes problematic between D2D pair. A distance limit model for RR allocation is proposed in [7], in which RR are allocated cellular and D2D link based max-flow algorithm. It enhances the sum rate but creates interference. These difficulties originate from the reuse of radio resources among cellular users and D2D users, which make intra-cell interference.

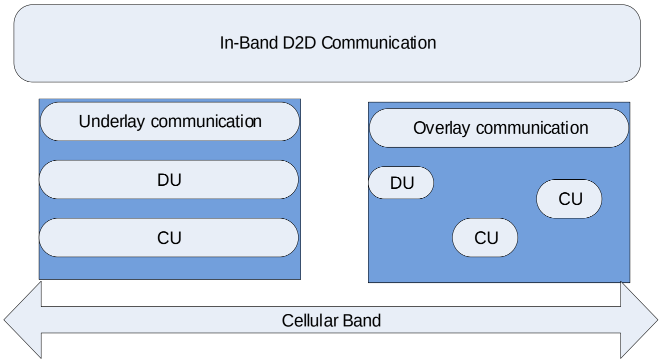


Fig. 2. In-band D2D radio resource distribution as an underlay and overlay.

Consequently, to exploit the advantages of D2D communication and accomplish an enhanced network performance over standard cellular networks, cautious RR allocation that considers both cellular users and D2D users is fundamental. RR allocation procedures for D2D underlay communication can be ordered relying upon the optimization metric [8]. RR allocation figures out which particular frequency and time resources ought to be allotted to each D2D and cellular links. RR allocation algorithms can be comprehensively grouped by the level of system control, centralized versus distributed, and the level of coordination between cells, single cell versus multi-cell [9]. Every cellular user in OFDMA is allocated to the sub-carriers and every sub-carrier is allocated by the network [10]. To facilitate an essential unit of RR allocation in OFDMA, sub-carriers are characterized as a sub-channel. Contingent upon how the sub-carriers are allocated to build each sub-channel.

The RR allocation techniques are grouped into a random type, comb type and block type as is appeared in Fig. 3. To avoid the wastage of RR, random type RR allocation is considered in this research. In a random type RR allocation, each sub-channel is comprised of a set of sub-carriers allocated randomly over the whole spectrum. If random type sub-channels are utilized, then interference is incorporated to accomplish the adversity gain. For this situation, all pilots situated over the entire bandwidth might be utilized for channel estimation between cellular users and D2D users. This sort of sub-channels tends to normal out the channel quality over the entire band [11]. Therefore, it can oblige high mobility, anyhow, when the quality of each sub-carrier consistently differs from one frame to the next. In a D2D enabled cellular network, besides, it is helpful for decreasing the co-channel interference by haphazardly allocating sub-carriers such that the probability of sub-carrier interference among D2D users and cellular users decreases. In random type RR allocation, to avoid the interference between co-cells different random type allocation is performed as presented in Fig. 3 random type(a) and random type(b). In this research in-band underlay, RR resource allocation technique is considered. The RR allocation is generic and can be pragmatic to many systems [12] for example, multi-cast, ad-hoc and Wi-Fi network. Therefore, some successful solution is required for D2D communication enabled network in which optimization is required to minimize with delay and interference.

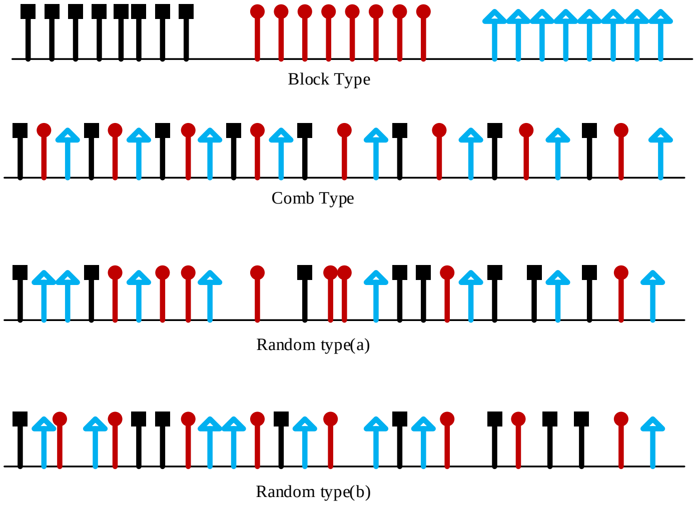


Fig. 3. In-band D2D radio resource distribution as an underlay and overlay.

III. Resources Allocation Model and Results

RR allocation depends on sub-channel and sub-carrier, whether they are centralized or localized. In D2D enabled the network utilizes the localized sub-channel and needs efficient RR allocation between D2D users and cellular users. The sub-channelization is classified into two classes adaptive and diversity sub-channel. There are three types of downlink diversity sub-channel according to usage partial, full and optional (partial or full) usage. It is totally depending on whether every sub-channel is built by scattered sub-carriers throughout the entire band or not. Similarly, in uplink, there is two types of diversity sub-channel partial and optional usage. Adaptive sub-channels are utilized in both downlink and up-link and in all kinds of the sub-channel comprises of 48 sub-carriers [13]. In this work OFDMA based LTE system is assumed. The basic difference between OFDM and OFDMA network is RR allocation on time domain in OFDM while frequency and time domain both in OFDMA to user equipment [14]. In OFDMA system, spatial RR are centrally allocated to D2D pair by the base station individually. In the RR allocation, the frequency is divided into sub-channel and time into slots. OFDMA makes grids of the channel which involve a sub-channel per slot. OFDMA system with a sub-carrier and sub-carrier selection is very important because of constraints between base station power and user equipment support as presented in result Fig. 4. Results elaborate that more sub-carriers lead more power and capacity from system model shown in Fig. 5.

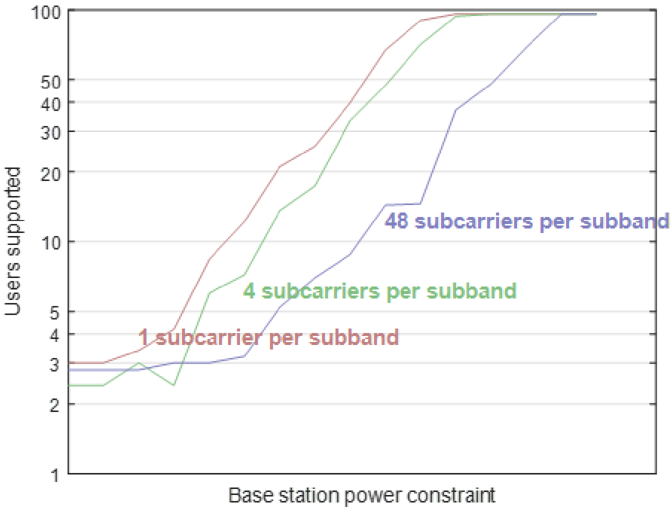


Fig. 4. User supported versus base station power constraints for different sub-carriers.

$$y = \boldsymbol{h}_{C1-B1}\mathbf{s} + n \quad (1)$$

where $\boldsymbol{h}_{C1-B1}$ is channel matrix of MIMO system patterned by $M_R \times M_T$.

$$\boldsymbol{h}_{C1-B1} = \begin{bmatrix} h_{1,1} & h_{1,2} & \cdots & h_{1,M_T} \\ h_{2,1} & h_{2,2} & \cdots & h_{2,M_T} \\ \vdots & \vdots & \ddots & \vdots \\ h_{M_R,1} & h_{M_R,2} & \ldots & h_{M_R,M_T} \end{bmatrix} \quad (2)$$

Let C-1 is cellular users and D-2 to D-5 are D2D users in a cell and these devices are transceiver devices as presented in Fig. 5. RR allocation is followed the following condition:

$$x_D^c = \begin{cases} 1 & if\ D2D\ pair\ reuse\ the\ resources\ of\ C1 \\ 0\ otherwise \end{cases} \quad (3)$$

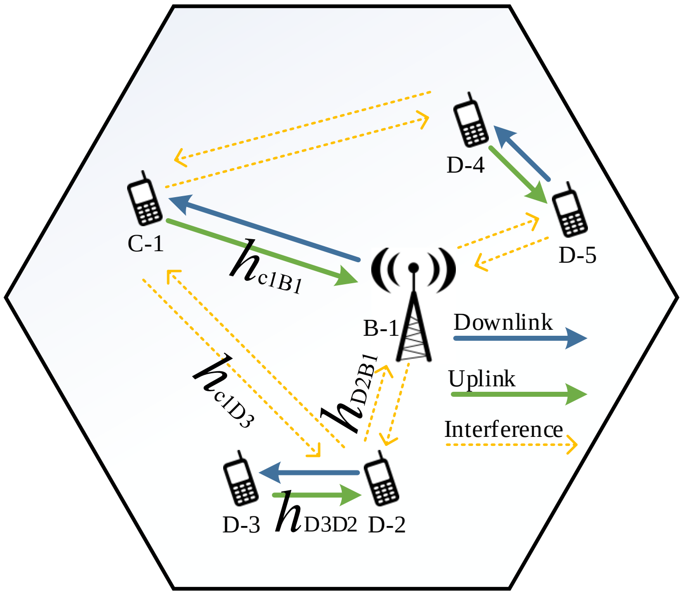


Fig. 5. RR system model in OFDMA single cell network.

x_D^c is indicator variable of cellular resources are allocated to D2D users. SINR of C-1 at base station B-1 is give as

$$\gamma_{C-1} = \frac{P_C \boldsymbol{h_{C1-B1}}}{N_0 + \sum_1^d x_D^c P_D \boldsymbol{h_{D2B1}^c}} \tag{4}$$

Where P_C and P_D are transmitted power of C-1 and D2D of D3D2 respectively. $\boldsymbol{h_{C1-B1}}$ is channel gain from cellular C-1 to base station B-1 and $\boldsymbol{h_{D2B1}^c}$ is channel gain of D-2 to B-1 when it is using cellular resources.

SINR at D2D of D3D2

$$\gamma_D = \frac{\sum_1^d x_D^c P_D \boldsymbol{h_{D3D2}^c}}{N_0 + \sum_1^d x_D^c P_C \boldsymbol{h_{C1D3}}} \tag{5}$$

$\boldsymbol{h_{D3D2}^c}$ *is* the channel gain between D-3 and D-2 using RR of cellular. The throughput can be calculated both for cellular users and D2D users using Shannon capacity model $R_c = log_2(1+\gamma_{c1})$ and $R_D = log_2(1+\gamma_D)$. There are many parameters effect on RR allocation of cellular users to D2D users to maintain the quality of service. These parameters are power allocations and interference management of cellular users and D2D users. This is optimization problem to maximize the system throughput by controlling transmission power and interference as presented in Fig. 6. Mathematically this optimization problem can be solved as

$$max_{x_D^c \geq 0, P_{D\geq 0}} \sum_1^d R_D + \sum_1^c R_c \tag{6}$$

Condition that $R_c \geq R_{c,min}$, $\sum_1^d x_D^c \leq 1$ and $\sum_1^c x_D^c \leq 1$. $R_{c,min}$ *is* the minimum required data rate of the cellular users. From the (4) and (5) it can be observed that D2D users should reuse RR of cellular users of largest gain $\boldsymbol{h_c^d}$. Therefore, D2D users have best SINR. Cellular users should allow D2D users to channel reuse which have smallest gain $\boldsymbol{h_{DB}}$. Therefore, the SINR of the cellular users at the base station is maximized [14]. The distance between D2D users is much smaller than the cellular users and base station practically. Therefore, data rate of D2D users generally much greater than the cellular users. To achieve best data rate, each D2D users should be matched with cellular users using

$$x_D^c = \begin{cases} 1 \; if \; C1 = arg\; max_c \; h_{D2D}^c \\ 0 \; otherwise \end{cases} \tag{7}$$

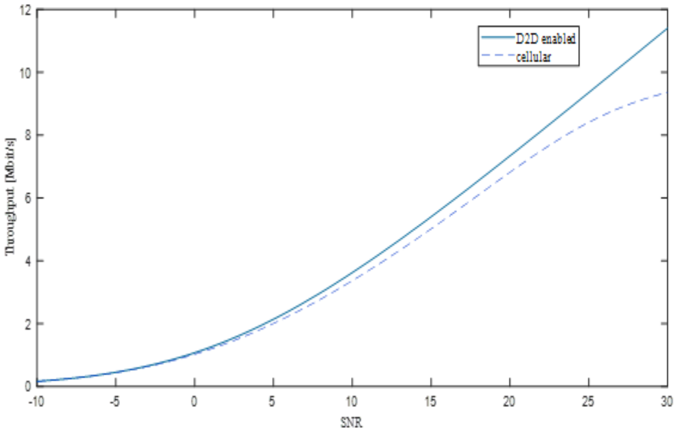


Fig. 6. Network throughput and SNR relation in D2D enabled the cellular system.

This means that D2D users should reuse the RR of cellular users if it has good data rate on that RR. It enhances the system capacity B/Hz as is proofed in Fig. 7. Another important parameter effect on RR allocation is transmitted power of D2D users. Quality of service constraints of cellular users who shares the similar RR with D2D users is not violated. From (4) and (5):

$$R_{c,\ min} \leq log_2(1+\gamma_{c1}) \tag{8}$$

$$R_{c,\ min} \leq log_2\left(1+\frac{P_C h_{C1-B1}}{N_0+P_D h_{D2B1}}\right) \tag{9}$$

$$P_D \leq \frac{1}{h_{D2B1}}\left(\frac{P_C h_{C1-B1}}{2^{R_{c,\ min-1}}} - N_0\right) = P_{D,\ max} \tag{10}$$

From the above equations quality of service of cellular users is not disrupted as long as D2D users lies in the [0, $P_{D,max}$]. Data rate is also increasing function with power and (6) verify the optimal transmit power of D2D users. As subcarriers increases frequency efficiency increases with visible difference between only cellular users and shared with D2D users as is explained in Fig. 8. D2D RR allocation algorithm follow is described as in Fig. 9.

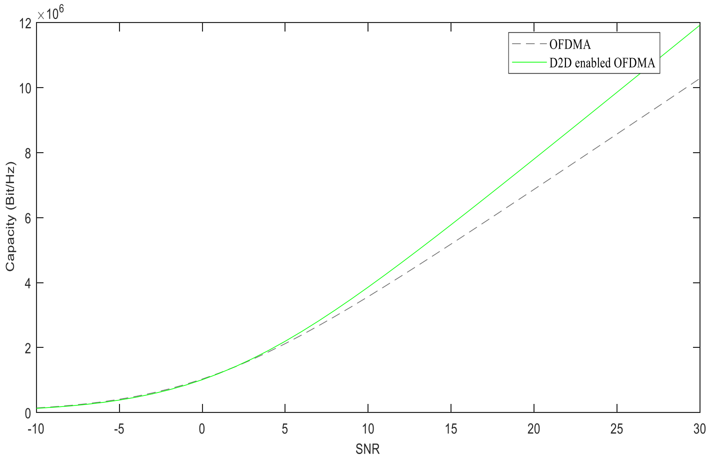


Fig. 7. Show the D2D enabled network capacity is better than simple cellular network due to random radio resources allocation.

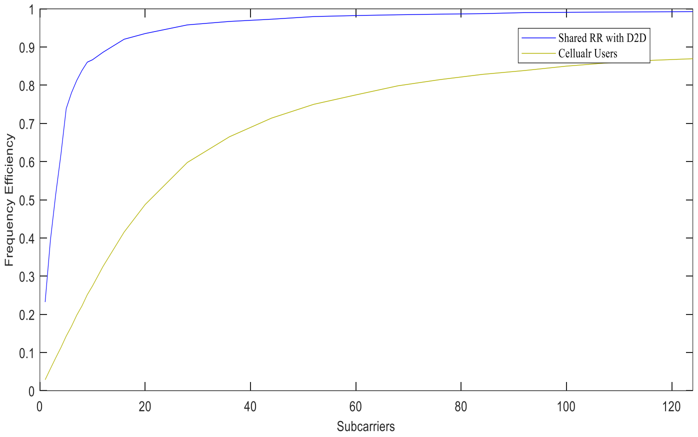
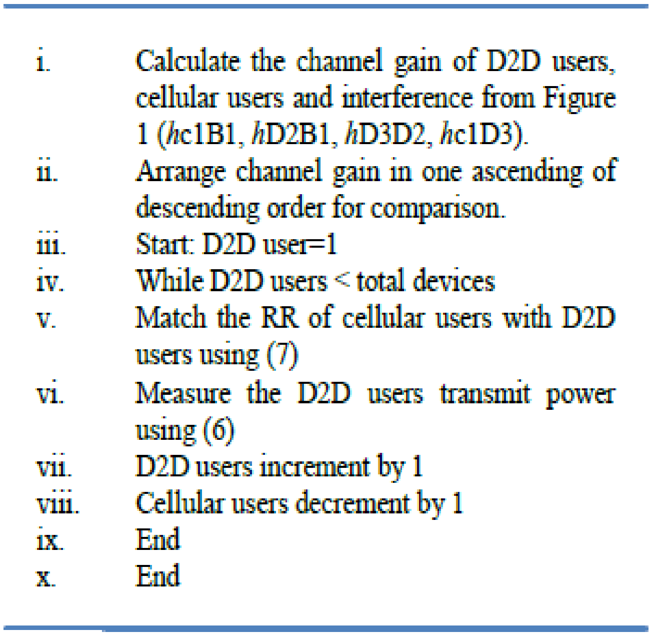


Fig. 8. Explain the efficiency increase with shared radio resources with cellular users and D2D users.



```
i.     Calculate the channel gain of D2D users, cellular users and interference from Figure 1 (hc1B1, hD2B1, hD3D2, hc1D3).
ii.    Arrange channel gain in one ascending of descending order for comparison.
iii.   Start: D2D user=1
iv.    While D2D users < total devices
v.     Match the RR of cellular users with D2D users using (7)
vi.    Measure the D2D users transmit power using (6)
vii.   D2D users increment by 1
viii.  Cellular users decrement by 1
ix.    End
x.     End
```

Fig. 9. Radio resources allocation algorithm flow.

IV. Conclusion and Future Work

After device discovery in D2D communication, resource allocation is the major issue to avoid the interference. In in-band underlay cellular network, swarm-based radio resource allocation technique is proposed which provide the random based resource allocation between cellular users and D2D users. In this proposed model OFDMA network uplink and downlink subchannel are allocated as subcarriers. It enhances the system capacity, frequency efficiency, and throughput. Further, this work can be extended for scheduling between cellular users and D2D users.

Acknowledgment

The authors would like to express their gratitude to the Ministry of Higher Education (MOHE) in Malaysia and Universiti Teknologi Malaysia (UTM) for providing the financial support for this research through the HICOE grant (R. J130000.7823.4F965 4J215). The grant is managed by Research Management Center (RMC) at UTM.

References

[1] O. Hayat, R. Ngah, and Y. Zahedi, "Device discovery for D2D communication in in-band cellular networks using sphere decoder like (SDL) algorithm," (in English), Eurasip Journal on Wireless Communications and Networking, journal article vol. 2018, no. 1, p. 74, Apr 3 2018.

[2] O. Hayat, R. Ngah, and Y. Zahedi, "Cooperative Device-to-Device Discovery Model for Multiuser and OFDMA Network Base Neighbour Discovery in In-Band 5G Cellular Networks," (in English), Wireless Personal Communications, journal article vol. 97, no. 3, pp. 4681-4695, Dec 2017.

[3] G. P. Zhang, K. Yang, and H. H. Chen, "Socially Aware Cluster Formation and Radio Resource Allocation in D2d Networks," (in English), Ieee Wireless Communications, vol. 23, no. 4, pp. 68-73, Aug 2016.

[4] A. Abdelhadi and H. Shajaiah, "Optimal Resource Allocation for Cellular Networks with MATLAB Instructions," arXiv preprint arXiv:1612.07862, 2016.

[5] P. Phunchongharn, E. Hossain, and D. I. Kim, "Resource Allocation for Device-to-Device Communications Underlaying Lte-Advanced Networks," (in English), Ieee Wireless Communications, vol. 20, no. 4, pp. 91-100, Aug 2013.

[6] W. Lin, C. Yu, and X. Zhang, "RAM: Resource allocation in mobility for device-to-device communications," in International Conference on Big Data Computing and Communications, 2015, pp. 491-502: Springer.

[7] G. Wang, L. F. Xiong, and C. Z. Yuan, "Resource Allocation for Device-to-Device Communications Based on Guard Area Underlying Cellular Networks," (in English), Chinese Journal of Electronics, vol. 26, no. 6, pp. 1297-1301, Nov 2017.

[8] A. Asadi, Q. Wang, and V. Mancuso, "A Survey on Device-to-Device Communication in Cellular Networks," (in English), Ieee Communications Surveys and Tutorials, vol. 16, no. 4, pp. 1801-1819, 2014.

[9] M. Dohler, T. Nakamura, A. Osseiran, J. F. Monserrat, and P. Marsch, 5G Mobile and Wireless Communications Technology. Cambridge University Press, 2016.

[10] S. C. Yong, K. Jaekwon, Y. Y. Won, and G. K. Chung, "MIMO-OFDM wireless communications with MATLAB," in Singapore: John Wiley & Sons (Asia) Pte Ltd, 2010, p. 544.

[11] K. J. Zou et al., "Proximity Discovery for Device-to-Device Communications over a Cellular Network," (in English), Ieee Communications Magazine, vol. 52, no. 6, pp. 98-107, Jun 2014.

[12] A. Abdel-Hadi and C. Clancy, "A utility proportional fairness approach for resource allocation in 4G-LTE," in Computing, Networking and Communications (ICNC), 2014 International Conference on, 2014, pp. 1034-1040: IEEE.

[13] D. Astely, E. Dahlman, G. Fodor, S. Parkvall, and J. Sachs, "LTE release 12 and beyond [accepted from open call]," IEEE Communications Magazine, vol. 51, no. 7, pp. 154-160, 2013.

[14] A. Radwan, "Resource Allocation for Device-to-Device Communications Reusing Uplink in Cellular Networks," 멀티미디어학회논문지, vol. 18, no. 12, pp. 1468-1474, 2015.

An Efficient Link Prediction Technique in Social Networks based on Node Neighborhoods

Gypsy Nandi
Assam Don Bosco University
Guwahati, Assam, India

Anjan Das
St. Anthony's College
Shillong, Meghalaya, India

***Abstract*—The unparalleled accomplishment of social networking sites, such as *Facebook*, *LinkedIn* and *Twitter* has modernized and transformed the way people communicate to each other. Nowadays, a huge amount of information is being shared by online users through these social networking sites. Various online friendship sites such as *Facebook* and *Orkut*, allow online friends to share their thoughts or opinions, comment on others' timeline or photos, and most importantly, meet new online friends who were known to them before. However, the question remains as to how to quickly propagate one's online network by including more and more new friends. For this, one of the easy methods used is list of *'Suggested Friends'* provided by these online social networking sites. For suggestion of friends, prediction of links for each online user is needed to be made based on studying the structural properties of the network. Link prediction is one of the key research directions in social network analysis which has attracted much attention in recent years. This paper discusses about a novel efficient link prediction technique *LinkGyp* and many other commonly used existing prediction techniques for suggestion of friends to online users of a social network and also carries out experimental evaluations to make a comparative analysis among each technique. Our results on three real social network datasets show that the novel *LinkGyp* link prediction technique yields more accurate results than several existing link prediction techniques.**

***Keywords*—Link prediction; online social networks; common neighbors; Jaccard's coefficient; Adamic/Adar; preferential attachment; FriendLink**

I. INTRODUCTION

Online Social Networks (OSNs) have become a means for millions of online users to express and share their opinions with other users. These OSNs provide an excellent opportunity for allowing interactions and exchange of thoughts, opinions and ideas among the online users in a group or community. Such networks can be represented as graphs, where a node or a vertex corresponds to a user present in the graph and an edge corresponds to any form of association between the nodes or users, such as friendship ties. Also, these OSNs are dynamic and this raises a question as to: how does the graph structure of these networks change over time? Based on this question, this paper studies about the likeliness of any two nodes of a graph to be associated in the near future, considering that presently there is no connection in the current snapshot of the OSN graph being studied. This problem, commonly called the link prediction problem, is a research area being studied by many researchers in this field to generate faster and more appropriate result with special consideration to scalability and dynamic nature of the graph. Fig. 1 gives a basic idea about how link prediction is done by studying the structural links of a network. In this figure, five nodes have been considered at time t and future predictions are being made at time t+1. By studying the existing links, two predictions are made which are marked as dashed lines in the figure.

Liben-Nowell and Kleinberg [1] were the first to study the link prediction problem and propose a prediction model for the same. Their model mainly studies the linkage structure of a social network and discusses several link prediction methods for inferring new links. In [2], Hasan et al. studied several classification models for possible link prediction in co-authorship domain that aimed to provide a comparison of several features using different feature analysis algorithms. Zheleva et al. [3] studied binary classification algorithm that mainly studies friend or family networks for link prediction. They have mainly worked on the predictive power of overlaying friendship and family ties on three real world social networks. In [4], Tylenda et al. studied time-aware and time-agnostic maximum entropy methods in which time-based weighting of edges were used. Chen et al. [5] made a detailed study and comparison of four algorithms related to link prediction, namely, Friend of a Friend (FOAF), SONAR, Content-plus-Link (CplusL) and content matching algorithms. Schifanella et al. [6] considered a sampling link prediction algorithm that can help users find friends with similar topical interests as well as facilitate the formation of topical communities. They also introduced a null model to show that a part of the similarity between online users is due to the correlations between user activity and user degree centrality in the OSN. In [7], Papadimi-triou et al. found the FriendLink Algorithm for fast and accurate link prediction in OSNs which outperforms many other related algorithms in terms of accuracy and time-complexity. Bayesian network has been also considered as a consistent model to understand the relations between future links to be predicted in networks [19], [20]. Recently, negative link prediction in social networks has attracted the attention of many researchers and considerable research work is being carried out to find efficient techniques for the same [15], [16], [18]. Such techniques aim to perform link prediction across multiple signed networks. Recent work has also focused on noise-filtering technique to predict links in complex networks [17].

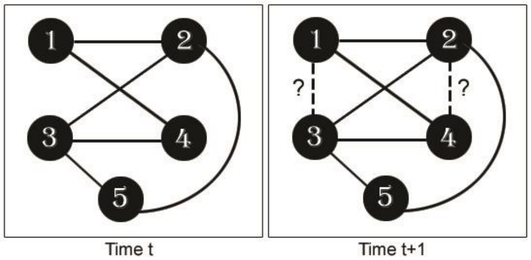


Fig. 1. Example of prediction of links in a given network.

The rest of the paper is organized as follows. In Section 2, a discussion on five standard link prediction techniques, namely, the *Common Neighbors*, *Jaccard's coefficient, Adamic/Adar, Preferential Attachment*, and *FriendLink* are given. Section 3 discusses about a novel link prediction technique *LinkGyp* which aims to provide better result than the above mentioned existing link prediction techniques. Section 4 illustrates the experimental results by comparing the various predictions of links made by the existing techniques with the novel technique. A conclusion of the paper and discussion on the scope for future work is given in Section 5.

II. Discussion on Standard Link Prediction Techniques

In this section, we at first discuss five standard link prediction techniques used in social networks and give a comparative analysis of the same. All these five techniques use the local-based features of a graph. There are, however, many global-based approaches also, which utilize the entire path structure in a network, but such approaches are computationally expensive for even a decent-sized social networks.

As seen in Fig. 1, vertices 1-5 indicates that there are five nodes or users in the network and the edges represent the existing links between each two nodes. In all the techniques explained below, an OSN is considered which is modeled as an directed or undirected graph *G=(V, E)*, where *V* denotes a set of vertices and *E* denotes a set of edges between two vertices in the network. Given below are some simple local similarity approaches based on node neighborhoods.

A. Common Neighbors

The technique of finding common neighbors for link prediction is considered as the most basic and significant method for prediction of links among nodes in the network. This approach was initially applied in the context of collaboration networks by Newman [8]. The basic idea of this technique is to find out the number of common nodes or neighbors or friends between two non-neighbors or non-friends. Now, the higher the number of common neighbors, the more likely is the chances of those two non-neighbors of being linked in the near future.

Using this concept, a link prediction score can be calculated between any two nodes p and q, where p and q are non-neighbors or non-friends at a given time t. The probability that these two nodes p and q will be linked in the near future is based on the score value given below:

$$score(p,q) = Neighbors(p) \cap Neighbors(q)$$

Considering Fig. 1, studying the network at time *t*, a prediction of future links (at time *t+1*) can be made between nodes 1 and 3 as well as 2 and 4. This is so because the number of common neighbors or the score value in both the cases is 2 (which are higher compared to the rest of the non-neighbors in the network). The technique of common neighbors is very simple and easy to analyze; yet this technique is very effective and it has been experimentally evaluated that it often outperforms several other complicated techniques used for link prediction. Algorithm 1 explains the link prediction technique based on the concept of common neighbors scores. In this algorithm, the social network *G* and the top *'n'* link predictions are taken as input and the top *'n'* nodes based on the score value is displayed as output. The score is calculated for each non-neighbors of a node and top-*n* predictions for the node are made based on the descending order of score values.

```
Algorithm 1: Common-Neighbors(G, n)

1: for each vertex v do
  1.1: $N_2 \notin$ Neighbors(v)
  1.2: for each vertex j $\in N_2$
  1.2.1: score = |Neighbors(v) $\cap$ Neighbors(j)|
  1.2.2: Store value of j and corresponding
         value of score
  1.3: end for
  1.4: Sort values of j in descending order of
       score
2: end for
3:  Display top n values of j
```

B. Jaccard's Coefficient

Jaccard's coefficient [9] is another simple technique of link prediction which is similar to common neighbors' technique discussed above as this technique also relies on the number of common neighbors between two nodes. In case of *Jaccard's coefficient*, the probability that two nodes *p* and *q* will be linked in the near future is based on the score value given below:

$$score(p,q) = \frac{|\,Neighbors(p) \cap Neighbors(q)\,|}{|\,Neighbors(p) \cup Neighbors(q)\,|}$$

As can been seen from the score calculation mentioned above, in case of *Jaccard's coefficient*, the number of common neighbors is simply divided by the number of total neighbors. For instance, in Fig. 1, the score of nodes 1 and 3 at time t, using *Jaccard's coefficient* of link estimation is 0.67. Similarly, score of nodes 1 and 5 at time t is 0.33. Hence, there is a higher chance of nodes 1 and 3 being linked in the near future compared to nodes 1 and 5. Hence, the score value between two nodes will always remain between 0 and 1; 0 when there is no single common neighbor between two nodes and 1 when the two nodes being compared is the same node. Algorithm 2 explains the link prediction technique based on the concept of *Jaccard Coefficient* scores. In this algorithm, the social network G and the top *'n'* link predictions are taken as input and the top *'n'* nodes base on the score value is displayed as output.

```
Algorithm 2: Jaccard-Coefficient(G, n)

1: for each vertex v do
   1.1: $N_2 \notin$ Neighbors(v)
   1.2: for each vertex j $\in N_2$
     1.2.1:score=|Neighbors(v) $\cap$ Neighbors(j)|/
           |Neighbors(v) $\cup$ Neighbors(j)|
     1.2.2: Store value of j and corresponding
            value of score
   1.3: end for
   1.4: Sort values of j in descending
        order of score
2: end for
3: Display top n values of j
```

C. Adamic/Adar

Adamic and Adar [10] have found another approach to predict links between two nodes in a network. In *Adamic/Adar* technique, all the common neighbors of two non-friends or non-neighbors are taken into consideration, and how many connections each of these common neighbors have are also considered. Thus, the probability that two nodes p and q will be linked in the near future is based on the score value given below:

$$score(p,q)=\sum_{x\in N(p)\cap N(q)}\frac{1}{\log|\,Neighbors(x)\,|}$$

Here x represents the set of common neighbors of nodes *p* and *q* and in the score calculation the number of neighbors of x is taken into consideration. In such a case, if a neighbor of *p* and *q* has two links or friends, a weight of 1/log(2) = 1.4 is considered. And again, if a neighbor of *p* and *q* has five links, a weight of 1/log(5) = 0.62 is considered. Hence, the more links a neighbor has, a better score value is obtained. Algorithm 3 explains the link prediction technique based on the concept of *Adamic/Adar* scores. In this algorithm, the social network G and the top *'n'* link predictions are taken as input and the top *'n'* nodes based on the score values is displayed as output.

```
Algorithm 3: Adamic-Adar(G, n)

1: for each vertex v do
   1.1: $N_2 \notin$ Neighbors(v)
   1.2: for each vertex j $\in N_2$
     1.2.1: Initialize score to 0
     1.2.2: for each k $\in$ (Neighbors(v) $\cap$
            Neighbors(j))
        1.3.2.1: score = score + 1 / (log|
                 Neighbors(k))
     1.2.3: end for
     1.2.4: Store value of k and corresponding
            value of score
   1.3: end for
   1.4: Sort values of k in descending order
        of score
2: end for
3: Display top n values of k
```

D. Preferential Attachment

The technique of *Preferential Attachment* for predicting links in a network is based on the concept that two non-neighbors or non-friends have higher chances of being connected by a link in the future if the product of their number of individual neighbors is high. This results in the calculation of score value as given below:

$$score(p,q) = Neighbors(p)\,\mathbf{.}\,Neighbors(q)$$

The term *preferential attachment* refers to the observation that in networks that grow over time, the likelihood that an edge is added to a node with n neighbors is proportional to n [11]. Experiments conducted by researchers have revealed that co-authorship is correlated with the product of the neighborhood sizes and the similar concept is applied for link prediction in social networks. Algorithm 4 explains the link prediction technique based on the concept of *Preferential Attachment* scores. In this algorithm, the social network *G* and the top *'n'* link predictions are taken as input and the top *'n'* nodes based on the score values is displayed as output.

```
Algorithm 4: Preferential-Attachment(G, n)

1: for each vertex v do
   1.1: $N_2 \notin$ Neighbors(v)
   1.2: for each vertex j $\in N_2$
     1.2.1: score = |Neighbors(v) *
            Neighbors(j)|
     1.2.2: Store value of j and corresponding
            value of score
   1.3: end for
   1.4: Sort values of j in descending order
        of score
2: end for
3: Display top n values of j
```

E. FriendLink

The *FriendLink* [7] approach of link prediction studies user's neighborhood by making use of paths of greater length. Here, two users connected with many unique pathways have a higher likelihood to know each other. Algorithm 5 explains the *FriendLink* approach for link prediction which takes as input the social graph *G*, the adjacency matrix *A* of graph *G*, number of nodes *'n'* present in the graph, and the maximum length of paths *'l'* explored in *G*. The algorithm provides as output the similarity matrix between two nodes in *G*. Based on the weights of similarity matrix friends can be recommended for a target node.

In the main program of Algorithm 5, the adjacency matrix of the graph is modified so that instead of holding the traditional values 0 or 1, the matrix is filled up with values 0 or vj, where vj is a node to which node vi is connected. In the function *ComputePaths*(), matrix multiplication of this modified adjacency matrix is performed with itself to produce all paths from node v_i to node v_j. Lastly, in the function *ComputeSimilarity*(), the similarity value between two nodes is measured to estimate the strength of connections between two non-linked nodes.

```
Algorithm 5: FriendLink(G, A, n, l) [7]

Main Program
1: for $v_i$ = 1 to n do
   1.1:  for $v_j$ = 1 to n do
     1.1.1: if A($v_i$,$v_j$) = 1 then
              A($v_i$,$v_j$) = $v_j$
            else A($v_i$,$v_j$) = 0
     1.1.2: end if
    1.2: end for
2: end for
3: for i = 2 to l
   3.1:  CombinePaths()
   3.2:  ComputeSimilarity(i)
4: end for
End Main Program

Function CombinePaths()
5: for $v_i$ = 1 to n do
   5.1:  for $v_j$ = 1 to n do
       5.1.1:  for k = 1 to n do
         5.1.1.1: if A($v_i$,k) <> 0 and A(k,$v_j$)<>
                  0 then
                  A($v_i$,$v_j$)=concatenate(A($v_i$,k),
                  A(k,$v_j$))
         5.1.1.2: end if
       5.1.2: end for
   5.2: end for
6: end for
7: return A($v_i$,$v_j$)
End Function

Function ComputeSimilarity()
8: for $v_i$ = 1 to n do
   8.1:  for $v_j$ = 1 to n do
     8.1.1:  denominator = 1
     8.1.2:  for k = 2 to i do
        8.1.2.1: denominator=denominator*(n-k)
     8.1.3:  end for
     8.1.4:  sim($v_i$,$v_j$) = sim($v_i$,$v_j$) + $\frac{1}{i-1}$ *
             $\frac{|paths^{i}_{vi..vj}|}{denominator}$
   8.2: end for
 9:  end for
 10: return sim($v_i$,$v_j$)
 End Function
```

There are also several other methods of link prediction which are based on the ensemble of all paths [22] such as *Katz* [13], *Hitting Time* and *SimRank* [14]. There are also other higher level approaches for link prediction such as clustering and low-rank approximation, which can be combined with the above mentioned link prediction techniques to give a more accurate output. The authors of [23] have used *Maximal Entropy Random Walk (MERW)* for link prediction, which emphasizes the centrality of nodes of the network. Other link prediction techniques consider temporal information to accurate predicts among non-edged nodes. Several other techniques focus different other issues such as giving weightage to more influential nodes, considering a subgraph based on the closed knit group in the graph, and so on. However, the primary focus on link prediction circles around which technique can give better accurate results along with better efficiency.

III. Novel *LinkGyp* Link Prediction Technique

In this section we first give a brief outline of our novel approach, named *LinkGyp* and then analyze the steps of the proposed algorithm.

A. Outline of the LinkGyp Technique

The *LinkGyp* prediction technique is a new approach proposed for prediction of links keeping in mind the scalability issue needed to be taken care of for huge-sized social networks. The basic idea of this technique is to initially take into consideration only those non-neighbors of a node whose product of their individual neighbors are among the top in descending order of list. A list is generated that includes the highly potential ‘could be friends but currently non-friends’ of a node and their corresponding scores. Using this list, a smaller sized graph for the node is now considered that is dependent on the number of top recommendations to be made. This results in a truncated graph where not all non-neighbors of a node are to be considered for a node. In fact, for a large-sized graph that involves huge number of non-friends for a node, the ultimate consideration of number of potential non-friends gets limitized to a great extent.

Once the smaller sized-graph is selected, the selection of top-n nodes results in a much faster execution by considering the *Adamic*/*Adar* approach where the simple counting of common features is refined by weighting rarer features more heavily [7]. As explained before, the *Adamic*/*Adar* method computes the similarity between two nodes p and q by means of a common feature of the two, say x. The similarity measure is then $\sum_{x} 1/|\log(\text{frequency}(x)|$ where, frequency(x) refers to frequency of occurrence of the common features between nodes *p* and *q*. The result obtained is the top-*n* prediction of links for each node of the graph.

B. The LinkGyp Algorithm

Algorithm 6 explains a novel link prediction technique *LinkGyp* that aims to provide better results than the above mentioned local similarity approaches of link prediction. In this algorithm, the social network graph *G=(V,E)* and the value of ‘n’ are taken as input. Here, *‘n’* represents the number of link predictions to be made for each node. Steps 1.1 to 1.4 concentrate on calculating scores for two non-edged nodes based on the product of the size of their individual neighbors.

Based on the descending order of their scores, the top *2n* non-edged nodes are considered for a node *v*. The reason behind choosing *2n* as the threshold value for selection of the subgraph is that more than *2n* lead to a bigger sized subgraph and less than *2n* may lead to consideration of very less nodes. Hence the choice of *2n* is considered due to performance considerations and it represents a performance-quality-tradeoff. For a reasonably-sized *‘n’*, experiments have been conducted with different multiples of *‘n’* and it has been found that *2n* is the optimum consideration for choosing the top *2n* non-edged nodes. However, if ‘n’ is too small, then the subgraph will also contain limited information and may lead to lower quality results. Again, if *‘n*’ is too large, it may lead to very small or no improvement in the result. Hence giving a right input value of n (5 ≤ *n* ≤ 50) will lead to better and more accurate results. The rest of the other non-edged nodes are discarded and further steps are carried out only for the top *2n* resulting nodes. These few steps are carried out keeping in mind the ground truth in economics that “the rich get richer”. Also, it results in a smaller choice of nodes with least computational complexity.

Next, a similar approach to *Adamic*/*Adar* explained above is followed to find the *‘n’* best predicted non-edged nodes for vertex ‘v’ from the set of only *2n* number of nodes selected in steps 1.1 to 1.4. The reason behind choosing the *Adamic*/*Adar* approach is that this technique considers the case that an affair owned by less objects, compared to owned by more objects, has greater effect on link prediction. In this way, the scores are calculated for each of the two non-edged nodes and ultimately the output for top-*n* predicted nodes is displayed for each unique user v based on the descending order of score.

The idea behind using this algorithm is mainly the scalability issue while dealing with dense social networks. As mentioned before, from steps 1.6 to 1.10, the estimation of links is done for a very small subgraph consisting of only *2n* nodes, where *‘n’* is the number of prediction of links to be made. Prior to step 1.5, the calculation of score1 is simple and does not involve studying in-depth the entire social network. It is only basically studying how many neighbors two non-edged nodes have. Hence, this algorithm proves to be an efficient method of prediction of links in social networks.

C. Complexity Analysis of the LinkGyp Algorithm

Online social networks are usually largely populated with information. Link prediction algorithms based on global based features, such as *Katz index* or Random Walk with Restart, are computationally too expensive for large graphs as it involves the inversion of matrix for link prediction. However, the standard existing link prediction algorithms discussed above are based on local based features, and comparatively have less time complexity than global based feature algorithms.

If we specifically consider the time complexity of our proposed *LinkGyp* technique, it is mainly O(2n), where *‘n’* is the number of link predictions to be made per node. This is much effective in terms of complexity analysis as the value of *‘n’* will be significantly much smaller compared to the total number of nodes ‘g’ for the entire graph. However, most of the other discussed link prediction techniques (such as, *Jaccard Coefficient*, *Adamic/Adar*, etc.) consider the entire nodes of the graph for prediction of links for a particular node that in real-time would be in terms of thousands, lakhs or even more. Hence, the complexity of *Jaccard coefficient* and *Adamic/Adar* techniques is O(g), where the value of *‘g’* is significantly greater than *‘n’*. For *Friendlink* algorithm, the time complexity is O(g x al), where *‘a’* is the average nodes degree in a graph and *‘l’* refers to the path lengths. Thus, the basic idea of *LinkGyp* algorithm is that the estimation of links is done for a very small subgraph consisting of only *2n* nodes, which gives better results as far as complexity of time is to be considered. The next section discusses the experimental results which prove that the above discussed novel link prediction algorithm gives a considerably better output compared to several basic existing link prediction techniques.

```
Algorithm 6: LinkGyp(G, n)

1: for each vertex v do
    1.1: cnt1 = |Neighbors(v)|
    1.2: for each j ∉ Neighbors(v) do
      1.2.1: cnt2 = |Neighbors(j)|
      1.2.2: score1 = cnt1*cnt2
      1.2.3: Store value of j and corresponding
             value of score1
    1.3: end for
    1.4: Sort values of j in descending order
         of score1 in arr1
    1.5: for i = 1 to 2n do
      1.5.1: Initialize score2 to 0
      1.5.2: Initialize cnt3 to 0
      1.5.3: for each k ∈ arr1 do
        1.5.3.1:  Initialize score2 to 0
        1.5.3.2: for each z ∈ (Neighbors(v) ∩
                 Neighbors(k))
          1.5.2.2.1:  cnt3=cnt3+|Neighbors(z)|
        1.5.3.3:  end for
      1.5.4: score2 = score2 + (1/log(cnt3))
      1.5.5: Store value of z and corresponding
             value of score2
      1.5.6: end for
    1.6: Sort values of z in descending order
         of score2
    1.7: end for
    1.8: Display top n values of z
2:  end for
```

IV. Experimental Evaluation

For conducting the experiments, three publicly available real-world datasets have been used that contains friendship network between users of social networking websites, namely the *facebook* dataset [24], the *hamsterster* dataset [12] and the *brightkite* location-based social networking website [21]. Table I gives few statistical information of all the three datasets.

TABLE I. Statistics of the Various Datasets

Dataset	*facebook*	*hamsterster*	*brightkite*
#Nodes	63731	1858	55228
#Edges	817035	12534	214078
Average Degree	25.640	13.492	7.353
Maximal Degree	1098	272	272
Average Path Length	2.832	3.453	2.76

a) The *facebook* dataset is an undirected network containing 63,731 nodes and 817035 edges that describes friendship data of *facebook* users. A node represents a user and an edge represents a friendship between two users. Fig. 2 illustrates the graphical view of the *facebook* dataset (nodes having degree less than six have not been considered) As can be seen from the figure, the yellow colored, bigger-sized nodes have highest degrees, the blue colored moderate-sized nodes have average degrees and red colored small-sized nodes have fewer degrees.

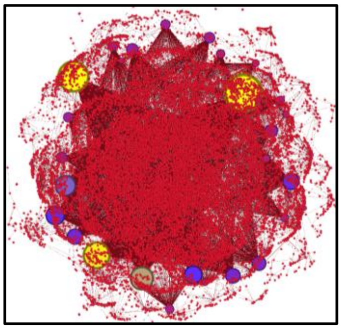


Fig. 2. The *facebook* dataset represented as a graph having different sized and colored nodes based on degree.

b) The *hamsterster* friendship dataset contains 1858 distinct nodes and 12534 edges which indicates the ties or friendship among all users in the network. The entire dataset has been represented in a graph as shown in Fig. 3 in which the yellow colored, bigger-sized nodes have highest degrees, the blue colored moderate-sized nodes have average degrees and red colored small-sized nodes have fewer degrees.

c) The *brightkite* friendship dataset contains undirected user-user friendship relations that have been gathered from a former widely used location-based social network. This dataset contains 55,228 distinct nodes and 214,078 edges that indicate friendship ties between two users. Fig. 4 illustrates the graphical form of a slice of the *brightkite* dataset (nodes having degree of one have not been considered) in which, the yellow colored, bigger-sized nodes have highest degrees, the blue colored moderate-sized nodes have average degrees and red colored small-sized nodes have fewer degrees.

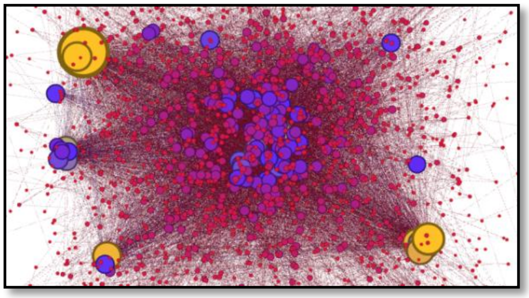


Fig. 3. The *hamsterster* dataset represented as a graph having different sized and colored nodes based on degree.

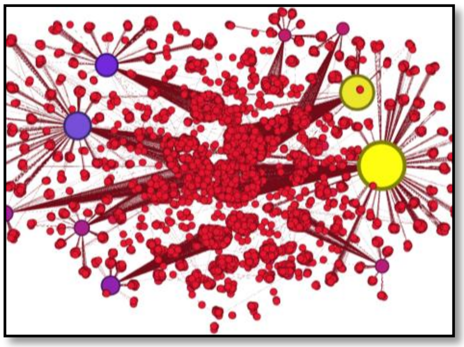


Fig. 4. The *brightkite* dataset represented as a graph having different sized and colored nodes based on degree.

Several experiments were conducted for the link prediction techniques mentioned above on all the three datasets. These experiments mainly aim at illustrating the performance comparison of the above mentioned link prediction techniques when compared to the random method generation of links for predicting future associations among nodes.

To conduct all the experiments, each of the entire dataset was divided into training and testing datasets consisting of 60% and 40% records respectively. Care was taken to include at least all core nodes in the training data set ('core' is the set containing nodes they have a direct link to minimum 10 other nodes). Tables II to VII illustrate the number of common predictions made between each two techniques for the two datasets which basically demonstrates which techniques are similar to each other in generation of link prediction results.

From the results obtained from Tables II to VII, it can be concluded that the random generation technique yields the least common predictions compared to the other six link prediction techniques. It can also be considered from all these tables that *Common Neighbors* and *Jaccard's Coefficient* predicts more similar friend suggestions for future links compared to the rest of the link prediction techniques.

TABLE II. THE NUMBER OF COMMON PREDICTIONS MADE ON THE FACEBOOK DATASET OUT OF 10000 (1000 USERS X 10) PREDICTIONS

	Common Neighbors	**Jaccard's Coefficient**	**Adamic/ Adar**	**Preferential Attachment**	**FriendLink**	**LinkGyp**	**Random Generation**
Common Neighbors	10000	8275	2504	2486	8175	3742	167
Jaccard's Coefficient		10000	1859	2144	6879	3301	173
Adamic/ Adar			10000	2570	4269	4957	204
Preferential Attachment				10000	3578	8146	196
FriendLink					10000	7686	185
LinkGyp						10000	168
Random Generation							10000

TABLE III. THE NUMBER OF COMMON PREDICTIONS MADE ON THE FACEBOOK DATASET OUT OF 20000 (1000 USERS X 20) PREDICTIONS

	Common Neighbors	**Jaccard's Coefficient**	**Adamic/ Adar**	**Preferential Attachment**	**FriendLink**	**LinkGyp**	**Random Generation**
Common Neighbors	20000	17751	6687	4591	15239	6157	427
Jaccard's Coefficient		20000	5826	4044	13486	6312	394
Adamic/ Adar			20000	3746	10017	7916	267
Preferential Attachment				20000	9810	16984	205
FriendLink					20000	14012	196
LinkGyp						20000	278
Random Generation							20000

TABLE IV. THE NUMBER OF COMMON PREDICTIONS MADE ON THE *HAMSTERSTER* DATASET OUT OF 18580 (1858 USERS X 10) PREDICTIONS

	Common Neighbors	Jaccard's Coefficient	Adamic/ Adar	Preferential Attachment	FriendLink	LinkGyp	Random Generation
Common Neighbors	18580	14833	5758	4218	16891	2983	208
Jaccard's Coefficient		18580	3384	2991	13567	1260	85
Adamic/ Adar			18580	3710	13001	3258	131
Preferential Attachment				18580	11763	13728	169
FriendLink					18580	12023	148
LinkGyp						18580	133
Random Generation							18580

TABLE V. THE NUMBER OF COMMON PREDICTIONS MADE ON THE *HAMSTERSTER* DATASET OUT OF 37160 (1858 USERS X 20) PREDICTIONS

	Common Neighbors	Jaccard's Coefficient	Adamic/ Adar	Preferential Attachment	FriendLink	LinkGyp	Random Generation
Common Neighbors	31760	27940	13572	5589	27438	6413	444
Jaccard's Coefficient		31760	9154	3093	24890	3018	360
Adamic/ Adar			31760	5122	22769	7200	285
Preferential Attachment				31760	20026	26728	294
FriendLink					31760	24374	251
LinkGyp						31760	263
Random Generation							31760

TABLE VI. THE NUMBER OF COMMON PREDICTIONS MADE ON THE *BRIGHTKITE* DATASET OUT OF 20000 (10000 X 10) PREDICTIONS

	Common Neighbors	Jaccard's Coefficient	Adamic/ Adar	Preferential Attachment	FriendLink	LinkGyp	Random Generation
Common Neighbors	20000	13076	9394	3186	10782	4280	176
Jaccard's Coefficient		20000	6702	1090	8201	2132	154
Adamic/ Adar			20000	2080	7658	2824	168
Preferential Attachment				20000	7105	162s62	186
FriendLink					20000	15396	154
LinkGyp						20000	202
Random Generation							20000

TABLE VII. THE NUMBER OF COMMON PREDICTIONS MADE ON THE BRIGHTKITE DATASET OUT OF 40000 (10000 X 20) PREDICTIONS

	Common Neighbors	Jaccard's Coefficient	Adamic/ Adar	Preferential Attachment	FriendLink	LinkGyp	Random Generation
Common Neighbors	40000	31750	19206	6236	28106	8872	184
Jaccard's Coefficient		40000	17544	3844	25871	7608	256
Adamic/ Adar			40000	5818	24712	12512	280
Preferential Attachment				40000	23108	26792	358
FriendLink					40000	27115	298
LinkGyp						40000	407
Random Generation							40000

Experiments were also conducted to find the number of correct predictions made on the testing datasets so as to find which techniques yield better results. A result of this is depicted in the Fig. 5-7 which again compare the above mentioned link prediction techniques against the novel *LinkGyp* link prediction technique for all the three datasets. Considerations were made for values of *'n'* as 10 and 20, where *'n'* is the number of predictions to be made for a particular node. Link predictions, in turn, were made for 1000 random distinct nodes present in the *facebook* dataset, and for each of the 1858 distinct nodes present in the *hamsterster* dataset, as well as for 2000 distinct nodes present in the *brightkite* dataset. Experiments reveal that the novel *LinkGyp* technique yields more accurate results followed by the *FiendLink* (considering lengths of path 2), Preferential Attachment and Adamic/Adar techniques. However, the random generation technique of link prediction, which randomly chooses the 'n' non-friends of a node, fails to come at par with all the other five prediction techniques.

The *hamsterster* dataset consists of densely-edged connections compared to the other dataset taken into consideration, namely the *facebook* dataset and the *brightkite* dataset which consists of comparatively sparsely-edged connections. Hence, it can be concluded that the novel *LinkGyp* link prediction technique can be considered as an efficient technique for link prediction keeping in mind the performance, scalability and execution time while dealing with social networks that comprise of thousands, lakhs or even more unique users and this technique is suitable for both densely-edged and sparsely-edged connections.

In summary, the results displayed in Fig. 5-7 indicate that results might slightly differ based on the scalability and sparseness of the dataset we are working upon. However, our novel *LinkGyp* technique outperforms other mentioned link prediction techniques in terms of accuracy.

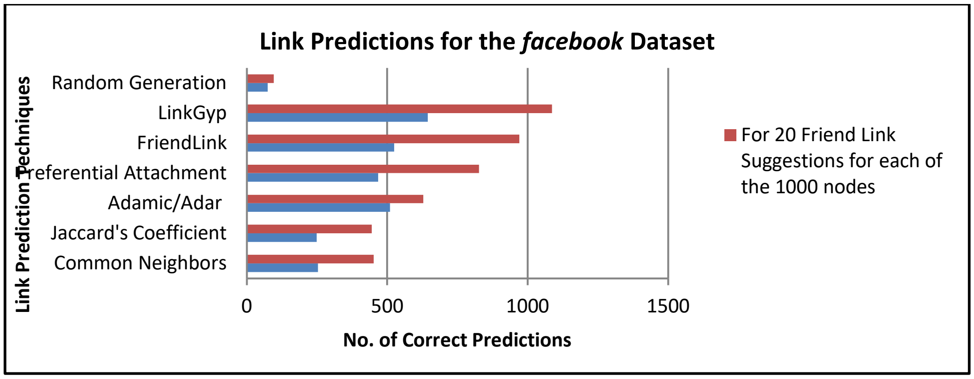


Fig. 5. Number of correct link predictions made on the *facebook* dataset.

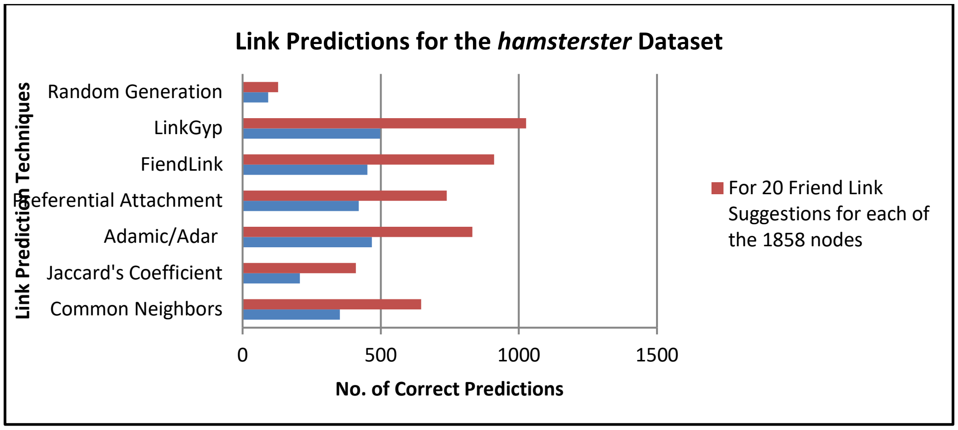


Fig. 6. Number of correct link predictions made on the *hamsterster* dataset.

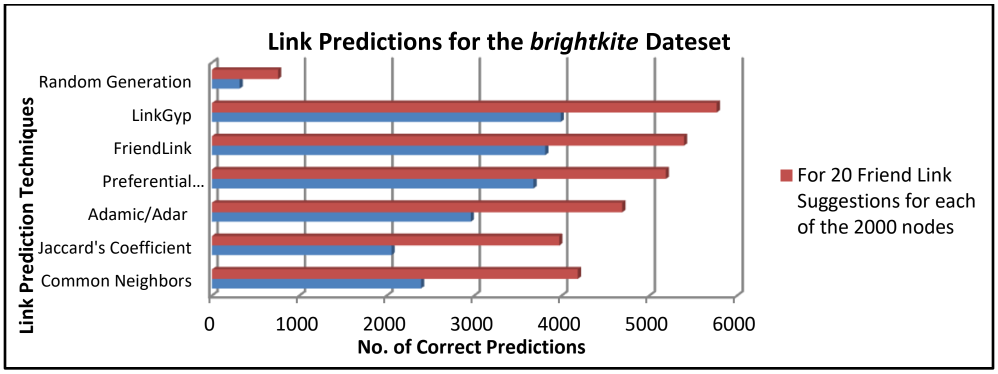


Fig. 7. Number of correct link predictions made on the brightkite dataset.

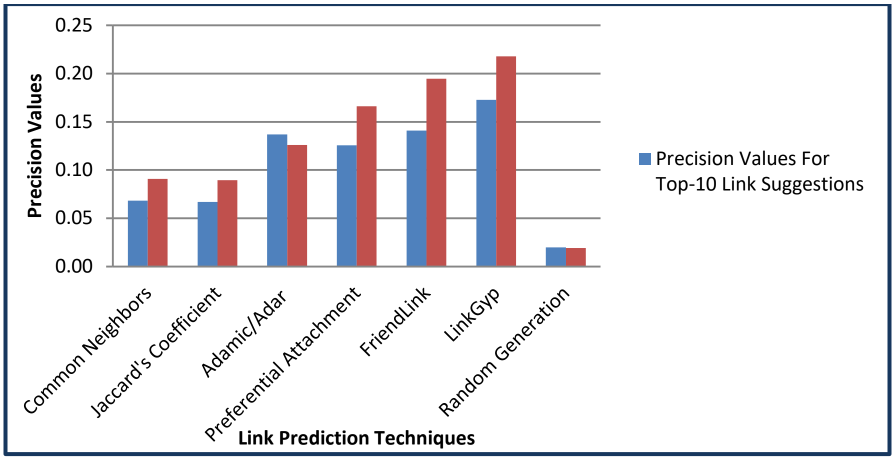


Fig. 8. Precision values of various link prediction techniques for the facebook dataset.

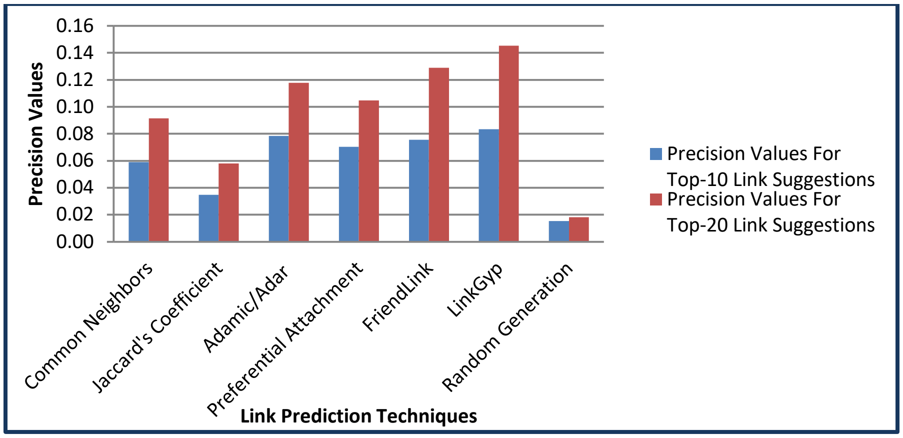


Fig. 9. Precision values of various link prediction techniques for the *hamsterster* dataset.

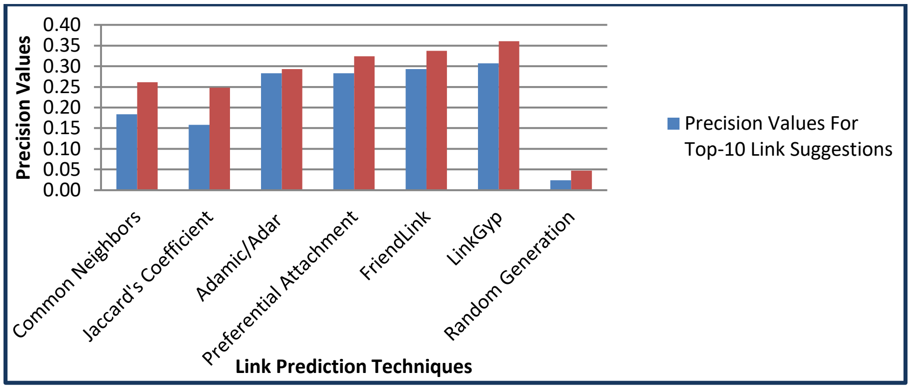


Fig. 10. Precision values of various link prediction techniques for the *brightkite* dataset.

Usually, for quantifying the accuracy of link prediction techniques, two standard metrics are commonly used: area under the receiver operating characteristic curve (AUC) and Precision [22]. Precision for a link prediction algorithm is calculated by considering the ratio of correct links selected to the total number of links selected. For example, if prediction of two new links has been made for a particular user, out of which one is correct and the other is incorrect, prediction value will be 0.5. This indicates that higher the precision value, higher will be the prediction accuracy. In this paper, we have used precision as the metric for evaluation of all the link prediction techniques and the results for the three different datasets are given in Fig. 8-10. The results of each of these figures below take into consideration the predictions made for 1000 random distinct nodes present in the *facebook* dataset, and for each of the 1858 distinct nodes present in the *hamsterster* dataset, as well as for 2000 distinct nodes present in the *brightkite* dataset.

V. Conclusions and Future Work

The tremendous growth in the use of online social networking sites has forced the researchers to carry out in-depth studies in social network mining. The link prediction technique in social networks is one such important research area that is in constant focus and is being studied and analyzed for better results. Our proposed work in this paper related to the proposed technique can be summarized as follows:

- This paper initially discusses the five basic standard techniques of link prediction and then gives a comparative analysis of these techniques using experimental results for the same. It can be concluded that these techniques will remain the simplest and basic techniques for studying and analyzing the concept of link prediction for OSNs and can assist a researcher in this field to get a preliminary idea about the same.

- The paper also discusses a new technique of link prediction namely '*LinkGyp'* that aims to provide a significantly better result in terms of more correct link predictions among non-linked nodes. We performed several extensive experiments on three different real-time datasets (Facebook, brightkite, and hamsterster) to arrive at a common result which proves that the *'LinkGyp'* technique can prove more efficient in prediction of links in social networks compared to several existing approaches.

- Considering link prediction to be one of the key research areas in social network mining, we have made an attempt to further improve the efficiency of link prediction with relate to number of correct predictions as well as run-time complexity.

- Finally, we can conclude that the proposed *'LinkGyp'* technique can be considered as the base model for link prediction technique to further carry out experiments on link predictions for complex networks.

As a future work, we plan to study other features of nodes along with their structural properties for generating better and more accurate results for link prediction in social networks. Also, further directions of study are needed to be carried out to improve the algorithm in order to deal with edges having negative weights (signed networks). The proposed algorithm can also be further enhanced to study the cold-start issue and link prediction for signed networks. If all these mentioned issues can also be considered while developing the link prediction techniques, it will provide new insight for modeling prediction of links in social networks.

References

[1] David Liben-Nowell , Jon Kleinberg, *The link prediction problem for social networks,* Proceedings of the twelfth international conference on Information and knowledge management, (2003), New Orleans, LA, USA [doi 10.1145/956863.956972]

[2] David Liben-Nowell , Jon Kleinberg, *The link prediction problem for social networks,* Proceedings of the twelfth international conference on Information and knowledge management, (2003), New Orleans, LA, USA [doi 10.1145/956863.956972]

[3] M. Hasan, V. Chaoji, S. Salem, and M. Zaki. *Link prediction using supervised learning*. In Proceedings of the Workshop on Link Discovery: Issues, Approaches and Applications, (2005)

[4] Elena Zheleva , Lise Getoor , Jennifer Golbeck , Ugur Kuter, *Using friendship ties and family circles for link prediction,* Proceedings of the Second international conference on Advances in social network mining and analysis, p.97-113, (2008), Las Vegas, NV, USA

[5] Jilin Chen , Werner Geyer , Casey Dugan , Michael Muller , Ido Guy, *Make new friends, but keep the old: recommending people on social networking sites,* Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, (2009), Boston, MA, USA [doi>10.1145/1518701.1518735]

[6] Rossano Schifanella, Alain Barrat , Ciro Cattuto , Benjamin Markines , Filippo Menczer, *Folks in Folksonomies: social link prediction from shared metadata,* Proceedings of the third ACM international conference on Web search and data mining, (2010), New York, New York, USA [doi 10.1145/1718487.1718521]

[7] Papadimitriou, P. Symeonidis, and Y. Manolopoulos, *"Fast and accurate link prediction in social networking systems",* The Journal of Systems and Software 85, (2012), pp. 2119–2132

[8] M. E. J. Newman, *Clustering and preferential attachment in growing networks.* Physical Review Letters E, (2001)

[9] Gerard Salton and Michael J. McGill, *Introduction to Modern Information Retrieval.* McGraw-Hill, (1983).

[10] Lada A. Adamic and Eytan Adar. *Friends and neighbors on the web, Social Networks,* 25(3):211, (2003).

[11] Jérôme Kunegis, Marcel Blattner, Christine Moser, *"Preferential attachment in online networks: measurement and explanations",* in Proceedings of the 5th Annual ACM Web Science Conference, pp. 205-214, (2013), Paris, France [doi>10.1145/2464464.2464514]

[12] Hamsterster friendships network dataset - KONECT, Available at http://konect.uni-koblenz.de/networks/petster-friendships-hamster, accessed June 2015.

[13] Leo Katz. *A new status index derived from sociometric analysis.* Psychometrika, 18(1), pp. 39-43, (1953).

[14] Glen Jeh and Jennifer Widom. *SimRank: A measure of structural-context similarity*. In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, July (2002).

[15] J. Tang, S. Chang, C. Aggarwal, and H. Liu, *"Negative link prediction in social media",* In WSDM 2015 - Proceedings of the 8th ACM International Conference on Web Search and Data Mining, pp. 87-96, ACM (2015).

[16] J. Tang, C. Yi, C. Aggarwal, and H. Liu, *"A Survey of Signed Network Mining in Social Media"*, arXiv preprint arXiv:1511.07569 (2015).

[17] B. Ouyang, L. Jiang and Z. Teng, *"A Noise-Filtering Method for Link Prediction in Complex Networks",* DOI: 10.1371/journal.pone.0146925 (2016).

[18] F. Liu, B. Liu, C. Sun, M. Liu and X. Wang, *"Deep Belief Network-Based Approaches for Link Prediction in Signed Social Networks",* Entropy 17.4, pp. 2140-2169; doi:10.3390/e17042140 (2015).

[19] S. H. Shalforoushan and M. Jalali, *"Link prediction in social networks using Bayesian networks",* in IEEE International Conference on Artificial Intelligence and Signal Processing, pp. 246-250 (2015).

[20] B. Zhang, S. Choudhury, M. A. Hasan, X. Ning, K. Agarwal, S. Purohit, and P. P. Cabrera *"Trust from the past: Bayesian Personalized Ranking based Link Prediction in Knowledge Graphs",* in SDM Workshop on Mining Networks and Graphs - MNG 2016, arXiv: 1601.03778, (2016).

[21] E. Cho, S. A. Myers, J. Leskovec. *Friendship and Mobility: Friendship and Mobility: User Movement in Location-Based Social Networks,* in ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), 2011.

[22] L Lu and T Zhou, *Link prediction in complex networks: A survey*. Physica A: Statistical Mechanics and its Applications, 390: 1150–1170, 2011.

[23] R.-H. Li, J. X. Yu, and J. Liu, *"Link prediction: the power of maximal entropy random walk,"* in CIKM, 2011

[24] Facebook friendships network dataset - *http://konect.uni-koblenz.de/networks/facebook-wosn-links,* accessed June 2016

Heterogeneous Buffer Size Impact on UDP Performance for Real-Time Video Streaming Application

Sarfraz Ahmed Soomro, M. Mujtaba Shaikh, Nasreen Nizamani, Ehsan Ali Buriro, Khalil M. Zuhaib
Department of Electronic Engineering,
QUCEST, Larkana, Sindh, Pakistan

***Abstract*—Communication specifically in real-time (RTC) is a terminology which insinuates any live media transmission that occurs inside time limits. In this paper, heterogeneous buffer sizes in random are utilized on different routers and for different ranges to examine their effect on the performance of network for user datagram protocol's (UDP) video streaming application. It appeared through numerical results that packet switches heterogeneous buffer sizes as a rule influence the general performance of the network. By thinking about bigger range of buffer sizes, throughput improves but End-to-End delay also increases which is customarily not commendable for RTC applications. On the contrary, throughput decreases on account of considering low range of buffer sizes; however, End-to-End delay additionally diminishes. In this manner, the middle of the road scope of buffer sizes range from 30 to 20, recommended for ideal throughput and an adequate lower End-to-End delay.**

Keywords—Real-time communication; buffer size; user datagram protocol; video streaming

I. Introduction

Video streaming is an important application of real-time communication (RTC) which is a live transmission of anything. Any important factor for an RTC application is its timeliness. Buffer is a type of memory usually used for controlling the congestion which occurs in networks by holding the packets of data for a certain period and such period is known as threshold. The threshold is very important factor in buffers, used to hold the data. If the time exceeds (delay) the specified threshold limit, the data would be lost. As timeliness is an important parameter for RTC applications like video streaming, the data packets must reach the destination in due course of time otherwise the data will be considered as lost [1]. The delay usually occurs due to the transmission, processing, queueing, and propagation which is usually called as End-to-End latency or delay. For the superior functionality of the network, there should be high throughput and low latency. User Datagram Protocol (UDP) is an important protocol suite for internet. It has been investigated by many researchers that UDP is a paramount protocol for RTC applications like live video streaming. In this research, the main focus is to investigate the influence of heterogeneous size of buffer on the performance of a video streaming application in real-time. Many people have worked on buffer size of the routers. Subsequently, it was proposed in [2] that as a rule-of-thumb, buffer size cannot be used. In [3], it was exhibited that traffic's round-trip times and loss synchronization are used to find out the size of buffer. In [4], real-time video streaming application was used to assess the networks for their buffer size. It was concluded that real-time video streaming losses are due to high End-to-End delay and jitter. It additionally demonstrated that TCP activity of traffic and also non-TCP type may hurtfully affect each other. In [5], the buffer size of the routers was changed and showed the results of throughput and packet loss of TCP and UDP for both live and non-live applications under two essential overseeing procedures for a waiting queue, Drop Tail and Random Early Detection (RED). The obtained numerical results proved that efficiency and throughput regarding TCP application enhances with a development in size of buffer for RED and packet loss is reducing. In [6], authors considered a heterogeneous wireless network for applications of multimedia type and attempted to choose UDP a magnificent component for transport layer and all other layers to bypass TCP issues. Most of the researchers have not considered the heterogeneous buffer size of the routers. Apart from the others work, in this paper, End-to-End latency or delay and throughput have been examined for heterogeneous buffer sizes of the network routers under live video streaming application.

In rest of the paper under Section II, the proposed network model is discussed. Section III describes the performance indicators which includes average latency, throughput, and buffer size of the router. Numerical results of the paper are explained in Section IV. Finally, the conclusion is presented in Section V.

II. Network Model

The model of proposed network is appeared in Fig. 1 where OMNET++ using NED (Network Descriptive) language[1] has been utilized to build up the network system. This network comprises of nine (09) routers and twenty-four (24) hosts to analyze the throughput, a benchmark metric and latency or End-to-End delay for the video streaming type traffic.

Video streaming is one of the innovation, utilized for sound and video broadcasting in real-time over the web. In live time, live substance, for instance, a political level-headed discussion, any games, or a talk show can be transmitted [7] for which the record is played out while parts of the document are being perceived and decoded just in this mode as there is no necessity

[1] http://inet.omnetpp.org/doc/INET/inet-manual-draft.pdf
https://github.com/inetmanet/inetmanet

for the document to be downloaded in full. Parameters of video streaming for simulation purpose are appeared in Table I. All parameters are fixed except buffer size of router which is altered each time in a specific range when simulation is carried out.

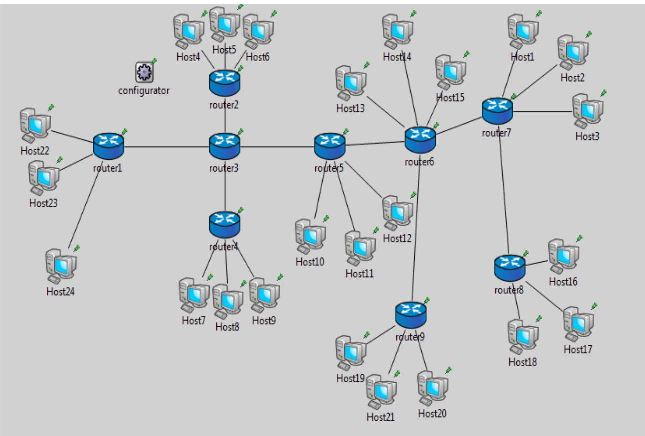


Fig. 1. Network model with point to point connection having 09 routers and 24 hosts.

Various simulations are carried out to assess the network performance in terms of mentioned indicators in Section III, utilizing the considered parameters shown in Table I. Simulations have been performed for video streaming application to examine the outcomes by varying the buffer size randomly of all intermediate range of considered routers in a specified range. That specified range is also changed for other simulations. This is called the heterogeneous buffer size approach. It should be remembered that in homogeneous approach, buffer size of all middle range considered routers is same. The created network comprises 09 routers with 24 hosts and at first middle routers are assigned different buffer sizes in a specific range to inspect the metrics which have been taken into account. Later that specific range is changed to further explore the effect of heterogeneous size of buffer on the metrics of performance. In spite of the fact that the size of buffer and range is altered, a network topology is fixed.

TABLE I. NETWORK SIMULATION PARAMETERS

Parameter	**VALUE**
Total number of nodes	24 (12 receiver, 12 transmitter)
Size of Video	1 GB
Buffer size (No. of frames)	100, 75, 40, 20
Data rate	2 Mb/s
Length of a packet	10000 bytes
Simulation time	10000 Secs (Over all traffic time)
Total number of routers	9
Frame capacity	4475
Start time	1 Second
Connection type	Point-to-point
Propagation delay	100 ms
Transmit interval	5 Seconds
Bit Error Rate (BER)	0.000001

III. INDICATORS OF PERFORMANCE

To maintain the high performance of the network, many parameters of the network are analyzed. By taking up the following metrics of performance, the capability of transport protocol can be assessed for the video streaming application.

A. *Average Latency or End-to-End delay*

This parameter is also known as one-way delay (OWD) and portrayed as what measure of total time is required for the movement of packets from source to destination. Round-trip time in IP networks is a different term than OWD. It includes various types of delays such as processing, propagation, transmission, and queueing [8] and is not simply the half of the round-trip time. Buffers end up being speedier in light of the way that packets for transmission purpose ought to be secured in these for a significant long time and is evaluated in second [9].

B. *Average Throughput*

This is another important parameter to measure the performance of the network. It is defined as the aggregate payload over the whole session separated by the total amount of time. The total amount of time is ascertained by taking the distinction in timestamps between the first and last packet. Its unit is in bits per second (bps).

$$Throughput = \frac{Average\ Payload\ (Total\ bits\ transmitted\ )}{Total\ Duration\ Observed} \tag{1}$$

C. *Buffer Size of Router*

Routers also known as packet switches have the buffers to handle the data during the time of congestion in a network which occurs due to diverse rates of transmission happen amongst packet switches and network transport. Thus, the sizing of packet switches (router) buffers is an important, pivotal and open research topic for the researchers to be addressed [3], [10].

IV. NUMERICAL RESULTS AND DISCUSSION

The network has been built using network descriptive (NED) language in OMNET++[1] and contains total 24 nodes with nodes from 13 to 24 as the sending nodes whereas 1 to 12 are the receiving nodes.

The middle routers 3, 4, 5, 6, 7, and 9 have been considered for varying the buffer size. The application which has been considered to analyze the latency or End-to-End delay and the network throughput is the video streaming. It should be noted that for homogeneous network, the buffer size of all the mid-way routers is fixed whereas in the heterogeneous approach, the buffer size of all intermediate numbered routers mentioned above are randomly varied. Later, the buffer sizes are taken into account in different ranges and finally, evaluate the impact of these heterogeneous buffer sizes on the mentioned indicators, network throughput and latency for UDP performance of transport protocol.

A. *Interpretation of network throughput and latency or End-to-End delay with hetrogeneous packet switch buffer sizes*

The standard indicators known as network throughput and latency or End-to-End delay are the important metrics to analyze the performance of transport UDP protocol suite for applications like video streaming in real-time. Initially, random buffer sizes between the range of 50 to 20 has been considered and later random buffer sizes in the ranges of 30 to 20 and 20 to 10 respectively have been taken into account to examine the numerical results.

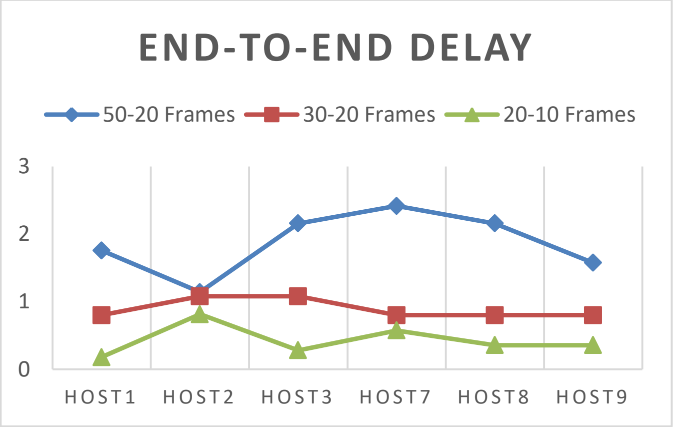


Fig. 2. Heterogeneous router buffer size vs End-to-End delay for video streaming application.

The numerical results are shown for the latency or End-to-End delay and the network throughput in Fig. 2 and 3, respectively. It is evident from the results that latency is quite high and attains the maximum value of 2.41 seconds when the range of 50-20 has been taken as shown in Fig. 2 but at the same time, throughput is also high and gets the maximum value of 93 bits/second. Although throughput is improved in this range but this range would not be considered for video streaming application because of high latency. Further, when the buffer range has been decreased from 50-20 to 30-20 and 20-10, the latency has diminished to its minimum value of 0.18 seconds for 20-10 range but also the throughput has decreased much to its lower value of 40, especially in the range of 20-10 which is not tolerable for the video streaming application.

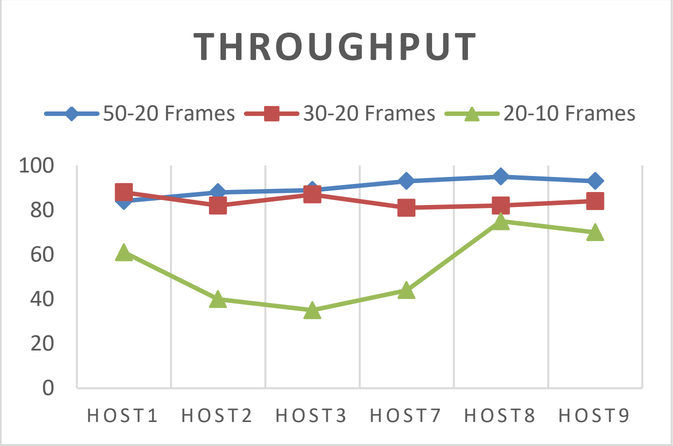


Fig. 3. Heterogeneous router buffer size vs Throughput for video streaming application.

Thus, it is obvious that with the increase in buffer size, the throughput improves but one-way delay is increasing. On the other side, if buffer size range is decreased to its lower range, delay reduces but network throughput is also declining. Thus, both high and low buffer sizes are not acceptable for the best performance of UDP protocol's video streaming application. Hence, the optimized values can be obtained when random buffer sizes are considered in the mid-range for excellent performance of UDP video streaming application where buffer size is in the range of 30-20 frames as clear from Fig. 2 and Fig. 3 results.

V. Conclusion

In this paper, heterogeneous buffer sizes in random were utilized on different routers and for different ranges to examine their effect on the performance of network for UDP protocol with video streaming application. It appeared through numerical results that packet switches heterogeneous buffer sizes as a rule influenced the general performance of the network. By thinking about bigger range of buffer sizes, throughput improved but End-to-End delay also increased which is customarily not commendable for RTC application. Be that as it may, throughput decreased on account of considering low range of buffer sizes, however, latency or delay (End-to-End) additionally diminished. In this manner, the middle of the road scope of buffer sizes range from 30 to 20 was recommended for ideal throughput and an adequate lower End-to-End delay. This work can be further extended by considering other real-time applications like voice over internet protocol (VOIP). Further, user datagram (UDP) applications maybe compared with transport control protocol (TCP).

References

[1] Li Tang; Hui Zhang; Jun Li; Yanda Li, "End-to-End Delay Behavior in the Internet," Modeling, Analysis, and Simulation of Computer and Telecommunication Systems, 2006, MASCOTS 2006. 14th IEEE International Symposium.

[2] Damon Wischik, Nick McKeown, "Buffer sizes for core routers," ACM SIGCOMM Computer Communication, 2005, vol., no. 35(3), pp.75-78.

[3] Dhamdhere, A.; Jiang, H.; Dovrolis, C., "Buffer sizing for congested Internet links," IEEE Proceedings of INFOCOM, 2005, 24th Annual Joint Conference of the IEEE Computer and Communications Societies.

[4] Theagarajan, G.H.P., Ravichandran, S., Sivaraman, V., "An Experimental Study of Router Buffer Sizing for Mixed TCP and Real-Time Traffic," ICON '06, 2006, 14th IEEE International Conference on Networks, Vol.1, pp.1-6.

[5] Ali, S.H., Nasir, S.A, Qazi, S, "Impact of router buffer size on TCP/UDP performance," Computer, Control & Communication, IC4 2013, 3rd IEEE International Conference.

[6] Ouldooz B.K., Mahmood F., "Adaptive end-to-end QoS for multimedia over heterogeneous wireless networks," Computers & Electrical Engineering, 2010, Vol. 36(1), pp.45-55.

[7] Masugi, M., Takuma, T., Matsuda, M., "Quality of Service assessment of video streams over IP networks based on monitoring transport and application layer processes at user clients", IEEE Proceedings of Communications, 2005, Vol. 152(3).

[8] Rohal, P., Dahiya, R., Dahiya, P., "Study and Analysis of Throughput, Delay and Packet Delivery Ratio in MANET for Topology Based Routing Protocols (AODV, DSR and DSDV)," International Journal for Advance Research in Engineering and Technology, 2013, Vol., no. 1(2), pp.54-58.

[9] Cranley, N., Perry, P., Murphy, L., "Dynamic Content-Based Adaptation of Streamed Multimedia," Journal of Network and Computer Applications, 2007, Vol. 30(3), pp. 983-1006.

[10] Prasad, R.S., Dovrolis, C., Thottan, M., "Router Buffer Sizing for TCP Traffic and the Role of the Output/Input Capacity Ratio," IEEE/ACM Transactions on Networking, 2009, Vol. 17(5) pp. 1645 - 1658.

FPGA based Synthesize of PSO Algorithm and its Area-Performance Analysis

Bharat Lal Harijan, Farrukh Shaikh, Burhan Aslam Arain
Institute of Information and Communication Technologies
Mehran University of Engineering and Technology,
Jamshoro, Sindh, Pakistan

Tayab Din Memon
Associate Professor
Mehran University of Engineering and Technology,
Jamshoro, Sindh, Pakistan

Imtiaz Hussain Kalwar
Associate Professor
DHA Suffa University, Karachi, Sindh,
Pakistan

***Abstract*—Digital filters are the most significant part of signal processing that are used in enormous applications such as speech recognition, acoustic, adaptive equalization, and noise and interference reduction. It would be of great benefit to implement adaptive FIR filter because of self-optimization property, linearity and frequency stability. Designing FIR filter involves multi-modal optimization problems whereas conservative gradient optimization technique is not useful to design the filter. Hence, Particle Swarm Optimization (PSO) algorithm is more flexible and optimization technique based on population of particles in search space and alternative approach for linear phase FIR filter design. PSO improves the solution characteristic by giving a novel method for updating swarm's position and velocity vector. Set of optimized filter coefficients will be generated by PSO algorithm. In this paper, PSO based FIR Low pass filter is efficiently designed in MATLAB and further Xilinx System Generator tool is used to efficiently design, synthesize and implement FIR filter in FPGA using SPARTEN 3E kit. For an example specifications, output of PSO algorithm is obtained that is set of optimized coefficients whose response is approximating to the ideal response. Hence, functional verification of the proposed algorithm has been performed and the error between obtained filter and ideal filter is minimized successfully. This work demonstrates the effectiveness of the PSO algorithms in parallel processing environment as compared to the Remez Exchange algorithm.**

Keywords—Particle swarm optimization (PSO); Remez Exchange Algorithm; FPGA implementation; FIR filter

I. Introduction

Digital filters enable us to pass some frequencies unaltered, while totally blocking others. Generally digital filters consist of two types; finite Impulse response (FIR) and infinite Impulse Response (IIR). An exactly linear phase response can be generated by FIR Filter and no any phase distortion or noise present in the output signal which is required in wide verity of telecommunication applications i.e. echo cancellation, noise and interface reduction, speech or image encoding. Different techniques are accessible for design the FIR filter. Window method is most frequently used tool but this method is not much capable to efficiently control the of frequency response in several bands of frequency [1]. Remez Exchange Algorithm or Parks–McClellan (PM) algorithm is ordinary method for designing FIR filter but this method has some limitation of high pass band ripples and computational complexity [2], [3]. It is good to design filter using optimization algorithm because of less mean squire error between desired response and actual response [3]. Optimization is not new techniques while numerous efforts have been already made for optimum design. Like Genetic Algorithm [3], Particle Swam Optimization algorithm [4], Differential Evolution [5], Artificial Bee Colony [6] are implemented for filer design. These methods showed themselves fairly effective by providing better control of performance constraints in addition to high stopband attenuation. Genetic Algorithms gives the effective result for local optimum but not successful in fining global optimum, PSO technique is able to solve problem [7]. Software based PSO algorithm increases the run time because of iterative process, additional processing time and storage is needed for FIR filter implementation [8]. PSO gives the better solutions over GA, processing time of one iteration of PSO algorithm gives higher process speed for optimization problems rather than genetic algorithm [8]. Implementation of digital filers based on FPGA which is flexible, low power, low cast and area sufficient provide better performance and superior to traditional approach [9].

Designing FIR low pass filter using traditional methods require more coefficients if sharp cutoff or no phase distortion is required and actual response $H(\Omega)$ is not more approximating to desired frequency response $Hd(\Omega)$ within a given specification in magnitude and phase [11], [12]. In recent past, one of the alternatives to this approach reported is short word length DSP systems [13], [14] in which sigma-delta modulation is a key element. However, in this research paper we have attempted to present the PSO based FIR filter designed in MATLAB; output of PSO is set of optimized coefficients whose response is approximating to the ideal response. Main objective is to efficient design, synthesize and functional verification of the optimized and original FIR low pass filter using Xilinx System Generator and implement in FPGA through hardware co-simulation, and to perform comparative analysis between both. This work will culminate with development of single-bit ternary PSO algorithm.

II. Fir Filter Design in Matlab using PSO Algorithm

In this section, PSO based FIR Low pass filer design and its implementation in MATLAB is discussed.

A. FIR Low Pass Filter Design

FIR filters are non-recursive filters and only depends upon past input information never on past output information [10]. Designed filter frequency response is given as:

$$H_d(e^{jw}) = \sum_{n=0}^{N} h[n]\, e^{-jwn} \qquad (1)$$

where h[n] shows filter's impulse response, N is order is filter with N+1 length. Ideal response of Low Pass filter is defined as;

$$H_i(e^{jw}) = \begin{cases} 1, & for\ 0 \leq w \leq w_c \\ x, & otherwise \end{cases} \qquad (2)$$

where w_c shows cutoff frequency of LP filter. Here obtained filter is designed by Remez Exchange Algorithm, this algorithm provides so many ripples in stop band, for sharp cutoff more coefficients are required. PSO algorithm is used to overcome this problem by minimizing the error between Remez and Ideal filter. The error equation is;

$$E(w) = [H_i(e^{jw}) - H_d(e^{jw})] \qquad (3)$$

$H_i(e^{jw})$ is the ideal frequency response and $H_d(e^{jw})$ is frequency response of the approximate filter.

B. Particle Swarm Optimization (PSO) Algorithm

PSO is the optimization technique used to determine the search space for specified problem to find the setting or constraint that essential to maximize the specific object [15]. This global optimization technique was, first introduced by J. Kennedy and R. C. Eberhart in 1995, based on common behavior of fish schooling or bird flocking [16]. PSO algorithm can solve optimization-based problems, in this research PSO is used to optimize the FIR filter coefficients to minimize the error. PSO algorithm is iterative process and initialized with population consist of N particles and every particle initialized to random position. For each iteration the error fitness function is used to measure the fitness value of each particle *i* in the search space. Then velocity vector is calculated which influenced by the particles individual experience as well as the experience of its neighbors. Velocity vector is further used to update the particles position which defines the filter coefficients. The velocity update equation for particle is given as:

$$v_i^{t+1} = w.v_i^t + c_1.r_1(p_i^t - x_i^t) + c_2.r_2(p_g^t - x_i^t) \qquad (4)$$

And position updating equation is:

$$x_i^{t+1} = x_i^t + v_i^{t+1} \qquad (5)$$

Superscripts t and t+1 represent the index of preceding and subsequent iterations, w is inertia coefficient, r_1 and r_2 are considered as uniformly distributed random numbers, c_1 and c_2 is cognitive acceleration term and social acceleration term and p_i and p_g are particle best position and swarm best position.

C. Designing Steps

Step I. In the very first step, specifies parameters that are required for designing FIR LP filter; Frequency of Sampling = 1kHz, W_{asps}=0.25, W_{stop}=0.3, Passband ripples= 0.1 and Stopband ripples = 0.01, filter order = 10 (Total no. of coefficients = 11).

Step II. Initialize Swarm size (Particles) = 250, w = 0.65, $c_1 = c_2$ = 2.05, Dimensions (No. of coefficients) D = 11, and maximum iteration *itmax* =100.

Step III. Create the initial particle vectors by utilizing above parameters and calculate initial value of error fitness function for the entire population by using (3).

Step VI. Error fitness vector is being used to calculate the minimum error value and calculate *pbest* (individual best) and *gbest (*group best) from entire swarm.

Step V. Update velocity and the position (filter coefficients) according to (4) & (5), which is to be considered as particle initial vector, error fitness is calculated form updated parameters also *pbest* and *gbest* is calculated accordingly.

Step VI. If values of vector *pbest* and *gbest* considered in Step V are improved than those calculated in Step IV, replaced the vector and no change otherwise.

Step VII. Repeat continuously from Step IV to Step VI till convergence conditions is meet (error fitness value equals minimum error fitness or reaching ***itmax***).

In Fig. 1, frequency response of PSO, Ideal and Remez algorithm-based FIR LP filters, for swam size N = 250 and *itmx*=100 is shown. By increasing swarm size, ripples in stop band are reduced at great extent and with increasing the iteration PSO algorithm gives the sharp cutoff at the cost of more chip area and performance degradation.

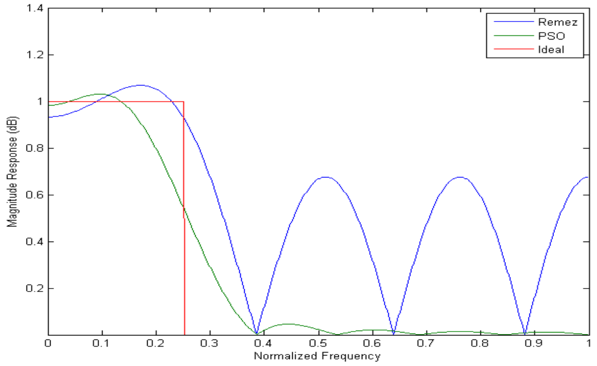


Fig. 1. Frequency response of PSO, ideal and Remez algorithm-based FIR low pass filter.

TABLE I. ORIGINAL COEFFICIENTS AND PSO BASED OPTIMIZED COEFFICIENTS

H(n)	Original Coefficients	PSO Coefficients
H(1)	0.0539051351555210	-0.0299519908221535
H(2)	-0.267487017557134	-0.0573261235259919
H(3)	0.100504853383030	-0.0536428075595944
H(4)	0.198866656248658	-0.00336950466085290
H(5)	0.247575005806795	0.0847531392346484
H(6)	0.265721847550549	0.179660912255896
H(7)	0.247575005806795	0.243609681976491
H(8)	0.198866656248658	0.251158067844883
H(9)	0.100504853383030	0.201996549974828
H(10)	-0.267487017557134	0.121085694673855
H(11)	0.0539051351555210	0.0444379358512675

III. FIR LP FILTER DESIGN UISNG XILINX SYSTEM GENERATOR

Xilinx System Generator is a programing tool used to develop efficient DSP algorithm and implement on FPGA. Due to reprogrammable capability of FPGA, implemented filter coefficients can be changed easily as per requirement [17]. System generator block set is available in MATLAB Simulink and it is high level programing tool for developing high performance DSP systems in FPGA [18-19]. System Generator enables the user to integrate with Simulink and it can easily generate synthesizable VHDL and Verilog code.

In this work, initially FIR Low pass filter is designed using FDA tool for the specification given outlined in designing steps. Further, PSO algorithm is applied on obtained coefficients and output of PSO is optimized coefficients whose response is approximate to ideal response. Xilinx system generator 14.7 is used for efficient direct form-I FIR low pass filter design and implemented in Spartan 3E FPGA kit through co-simulation.

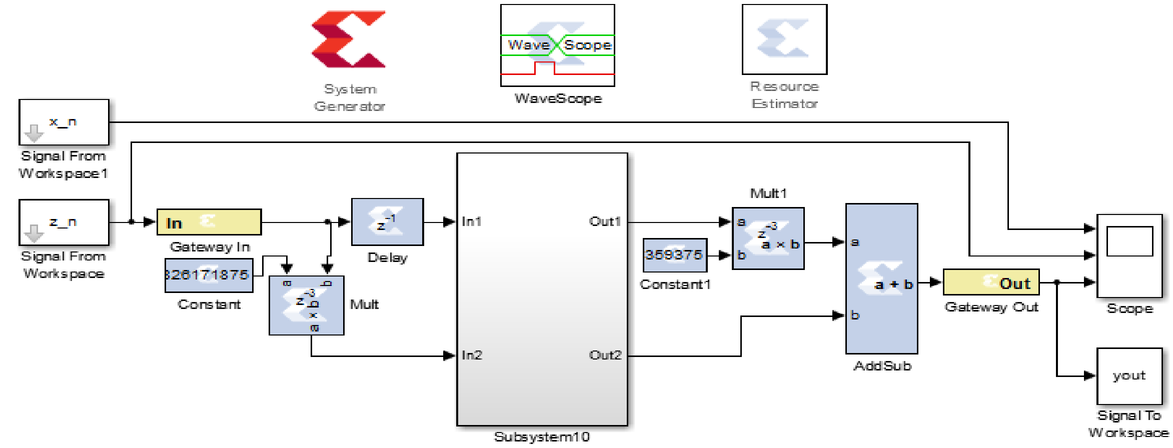


Fig. 2. Simulation model of direct form I FIR low pass filter.

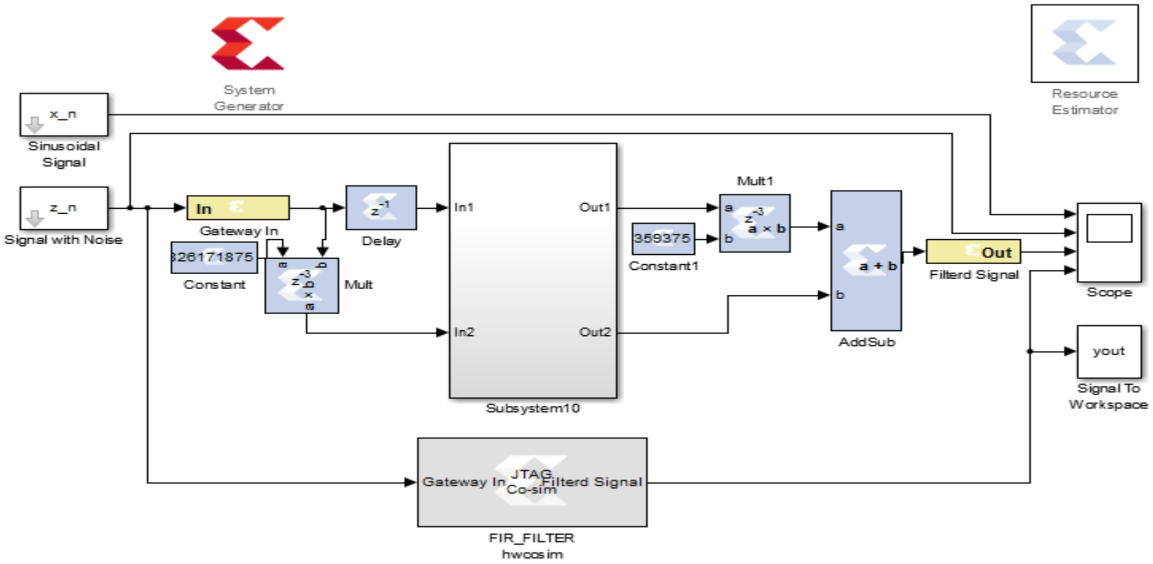


Fig. 3. Hardware co-simulation model of direct form I FIR low pass filter.

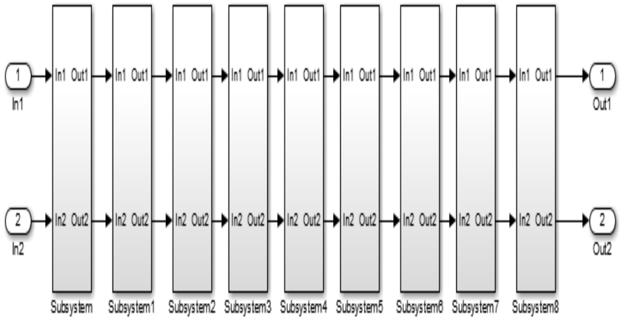


Fig. 4. Internal structure of direct Form I FIR filter model.

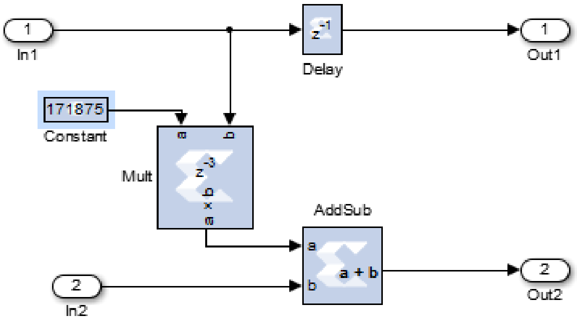


Fig. 5. Subsystem internal structure.

Simulation and Hardware Co-simulation model is shown in Fig. 2 and 3. JTAG cable shown in Fig. 2 is used for communication between Xilinx System Generator and Spartan 3E FPGA kit. System generator block set generate the of JTAG block of compatible signal for Spartan 3E kit. Resource Estimator is used to calculate the resources used by the device. It is used only when hardware is connected. Fig. 4 and 5 shows subsystem and internal structure of FIR Filter.

IV. Simulaiton Results and Discussion

Sinusoidal test signal of 125Hz frequency is generated in MATALB workspace as shown in Fig. 6 and White Gaussian Noise is added to original signal with Signal to Noise Ratio (SNR=1). Noisy signal is used as input of FIR filter. In first place, we used FIR filter model as shown in Fig. 2 with original filter coefficients, output signal of original filter is shown in Fig. 7 which contains more noise present in input signal. In Fig. 8, better output response is obtained, while we have used same model as shown in Fig. 2 but PSO optimized coefficients are employed as shown in Table I. This output signal is also taken to workspace in order to draw the spectrum as shown in Fig. 9, 10 and 11. The mean square error is computed using (3), and obtained Error = 0.3447933 when filter designed by Remez Exchange Algorithm and Error = 0.06442020 while filter designed by PSO algorithm. Area utilization is also observed using Spartan 3E kit, Table II shows area utilization of FIR filter using Spartan 3E FPGA kit.

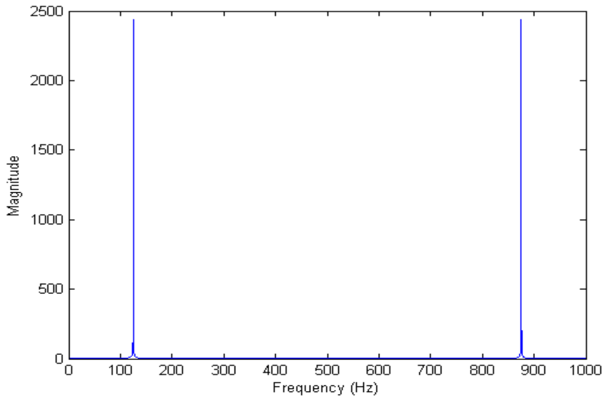


Fig. 6. Frequency spectrum of input signal.

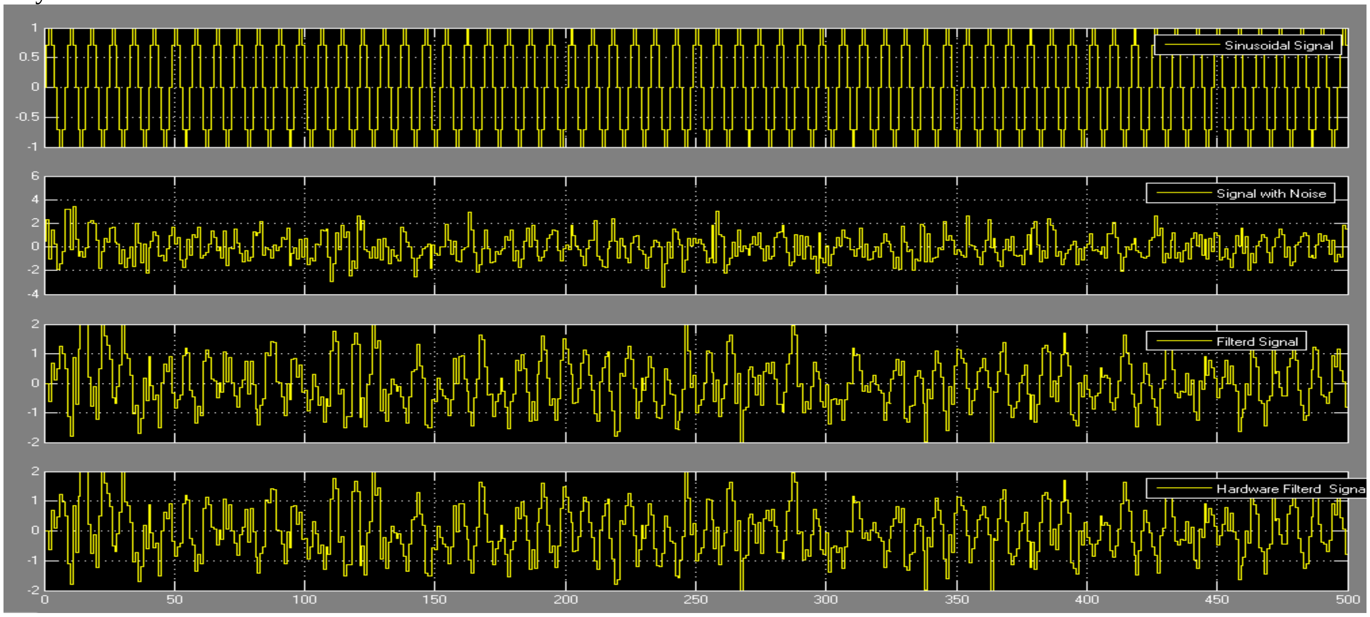


Fig. 7. Software simulation and hardware filtered output signal with original filte.

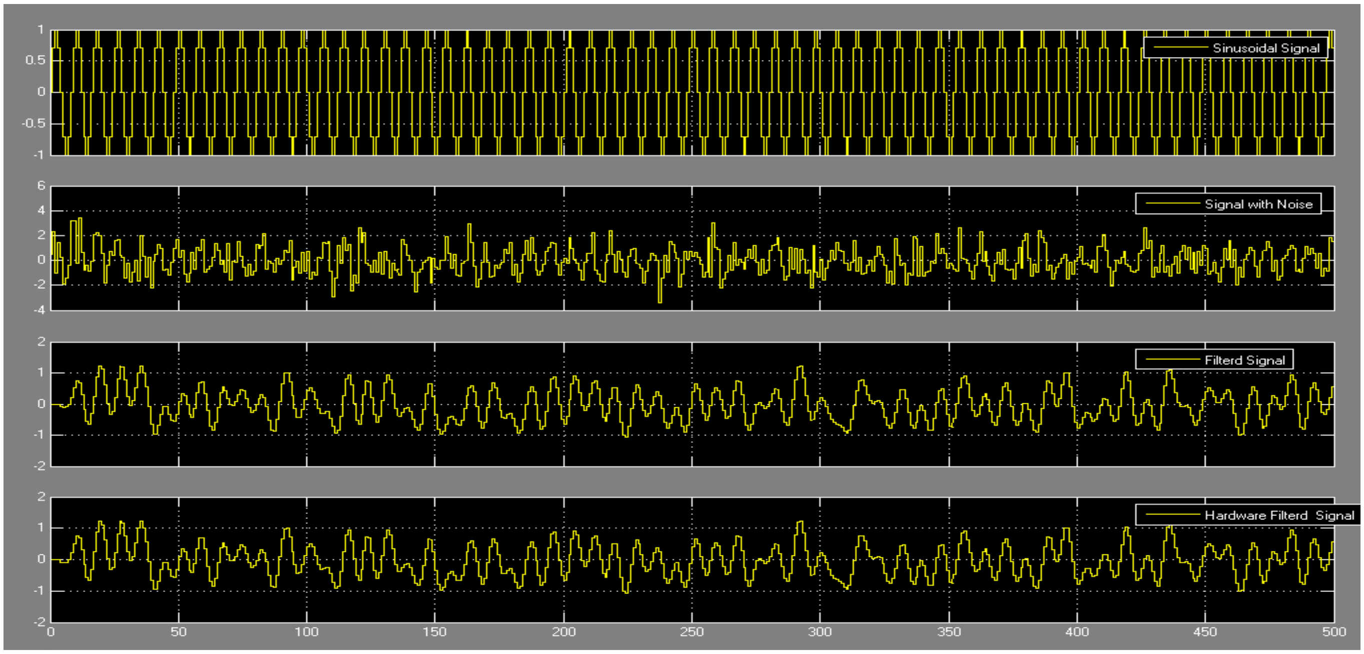


Fig. 8. Software simulation and hardware filtered output signal with optimized filter.

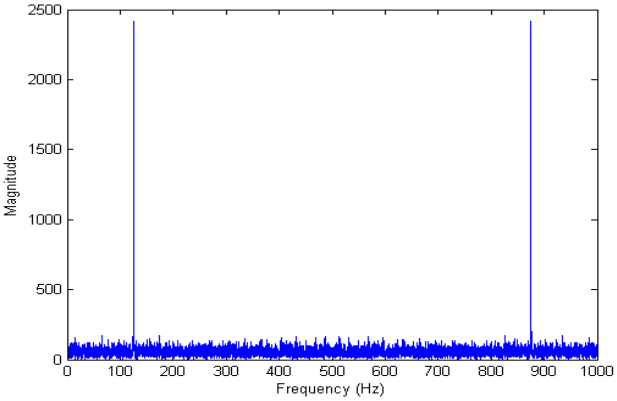


Fig. 9. Frequency spectrum of noisy signal.

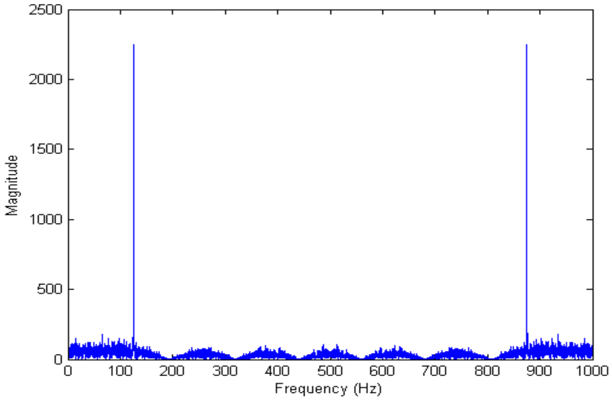


Fig. 10. Frequency spectrum of output signal of original filter.

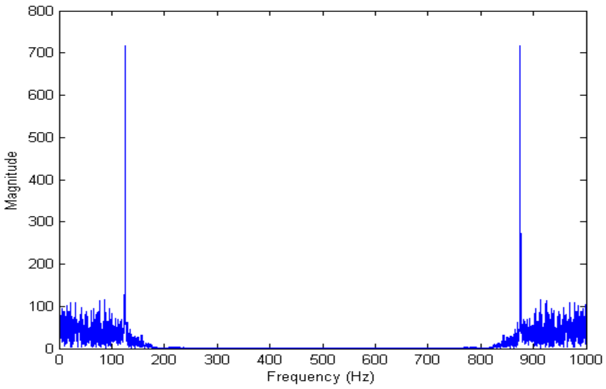


Fig. 11. Frequency spectrum of output signal of optimized filter.

In Fig. 9, frequency spectrum of noisy signal contains 125Hz original signal frequency and SNR=1 is shown. Whereas, Fig. 10 shows the output of original filter which shows the noise is present in the filtered signal and Fig. 11 shows the output of PSO based filter which contains less noise as compared to original filter. The area utilization by the PSO algorithm in FPGA given in Table II is quite small amount as compared to the available resources of the device.

TABLE II. AREA UTILIZATION OF FIR FILTER USING SPARTAN 3E KIT

Logic Unitization	**Used**	**Available**	**Utilization**
Number of Slices	82	4656	1%
Number of Slice Flip Flops	161	9312	1%
Number of IOBs	60	232	25%
Number of GCLKs	1	24	4%

V. Conclusion

In this paper, we have designed PSO based FIR filter in MATLAB that is further efficiently designed and synthesized using Xilinx System Generator in FPGA. Functional verification of the Remez Exchange Algorithm and PSO Algorithm based FIR low pass filter is performed through hardware co-simulation in Spartan 3E FPGA device. It is demonstrated that error in the PSO algorithm is successfully minimized. Area utilization of the PSO algorithm is also reported that is well below the available resources that shows much more room is available for improvement in the algorithm by increasing order of the filter.

Point to the future work is to compare this algorithm with other recursive algorithm and finally develop single-bit ternary FIR-like filter by employing these techniques.

References

[1] B. Luitel, G. K. Venayagamoorthy, “Differential Evolution Particle Swarm Optimization for Digital Filter Design,” in IEEE Cong. on Evolution Comput., pp. 3954-3961, 2008.

[2] F. Shaikh, T.D. Memon and I. H. Kalwar, “Design and Analysis of Linear Phase FIR Filter in Fpga using PSO Algorithm,” in 6th Mediterranean Conf. on Embd. Comput., Montenegro, 2017.

[3] K. Pardeep and S. Kaur, “Optimization of FIR Filters Design using Genetic Algorithm,” Int. J. of Emerg. Trends and Techno. in Comput. Sci., vol. 1, no. 3, 2012.

[4] Neha and A. P. Singh, “Design of Linear Phase low pass FIR Filter using Particle Swarm Optimization Algorithm,” Int. J. of Comput. Appl., vol. 98, no.3, pp. 0975–8887, 2014.

[5] Wei Zhong, “Linear phase FIR Digital Filter Design using Differential Evolution Algorithms,” M.S. thesis, Dept. of Elect. & Comput. Eng., University of Windsor, Ontario, Canada, 2016.

[6] A. K. Dwivedi, S. Ghosh and N. D. Londhe, “Modified Artificial Bee Colony Optimization-Based FIR Filter Design with Experimental Validation using fpga,” Inst. of Elect. and Techno. Signal Processing J., 2017: Available doi: 10.1049/iet-spr.2015.0214.

[7] A. Praneeth and P. K. Shah, “Design of FIR Filter using Particle Swarm Optimization,” Int. Adv. Research. J. in Sci, Eng. and Techno., vol. 3, no. 5, 2016.

[8] B. A. Mohamed sadek and SAKLY Anis, “FPGA Implementation of Parallel Particle Swarm Optimization Algorithm and Compared with Genetic Algorithm,” Int. J. of Adv. Comput. Sci. and Appl., vol. 7, no. 8, 2016.

[9] R. Thakur and K. Khare, “High speed FPGA Implementation of FIR filter for DSP Applications,” Int. J. of Mod. and Opt., vol. 3, no. 1, 2013.

[10] L.Tan and J. Jiang, “Finite impulse response filter design,” Digital Signal Processing Fundamentals and Applications, 2nd ed. pp 217-290, ELSVIER.

[11] P. M. Palangpour, “Fpga Implementation of PSO Algorithm and Neural Networks,” M.S. thesis, Missouri University of Science and Technology, 2010.

[12] P. Fodisch, A. Bryksa, B. Lange, W. Enghardt and P. Kaever, “Implementing high order FIR filters in FPGAs,” 2016: Available arXiv:1610.03360v2.

[13] A. Chang, T. D. Memon, Z. M. Hussain, I. H. Kalwar, and B. S. Chowdhry, "Design and Analysis of Single-Bit Ternary Matched Filter," Wireless Personal Communications, pp. 1-15, 2018.

[14] Tayab D Memon, P. Beckett, and A. Z. Sadik, "Power-Area-Performance Characteristics of FPGA based sigma-delta modulated FIR Filters," Journal of Signal Processing Systems (JSPS) vol. 70, pp. 275-288, 2013.

[15] J. Blondin, “Particle Swarm Optimization,” Tutorial, 2009.

[16] J. Kennedy and R. Eberhart, “Particle Swarm Optimization” in Proceedings of the IEEE Int. Conf. on Neural Networks, vol. 4, pp 1942–1948, 1995.

[17] S. Roy, L. Srivani and D. T. Murthy, “Digital Filter Design Using FPGA,” Int. J. of Eng. and Innovat. Techno., vol. 5, no. 4, 2015.

[18] K. Sahu and R. Sinha, “FIR filter Designing using Matlab Simulink and Xilinx System Generator,” Int. Res. J. of Eng. and Techno., vol. 2 no. 8, 2015.

[19] System Generator for DSP user guide, Xilinx UG640 v. 14.3, 2012.

A Comparative Evaluation of Dotted Raster-Stereography and Feature-Based Techniques for Automated Face Recognition

Muhammad Wasim
Department of Computer Science
Usman Institute of Technology
Karachi, Pakistan

S. Talha Ahsan
Department of Electrical Engineering
Usman Institute of Technology
Karachi, Pakistan

Lubaid Ahmed, Syed Faisal Ali, Fauzan Saeed
Department of Computer Science
Usman Institute of Technology
Karachi, Pakistan

Abstract—**Automated face recognition systems are fast becoming a need for security-related applications. Development of a fool-proof and efficient face recognition system is a challenging domain for researchers. This paper presents comparative evaluation of two candidate techniques for automated face recognition application, viz. dotted Raster-stereography and feature-based system. The relevant performance parameters — accuracy, precision, sensitivity and specificity – measured for the two techniques using IPRL Database of images are reported. The results suggest that dotted Raster-stereography based face recognition system has better accuracy, precision, sensitivity and specificity, and hence is a preferred choice as compared with feature-based system for such sensitive applications where high face recognition accuracy is required. On the other hand, feature-based technique is faster in terms of the training and testing times required. Hence such applications where volume of face recognition work is large and high speed is required with some compromise in accuracy being acceptable then feature-based technique may also be the technique of choice.**

Keywords—Raster-stereography; dotted raster-stereography; feature based; face recognition; IPRL

I. Introduction

In today's world where computer- and IT-based systems are being used everywhere for different applications, user authentication is a key requirement to ensure security of these systems and applications. Implementing system protection with user ID and password is a common practice in e-mail, online banking and ATM systems. However, if password is hacked or stolen, then these apparently secure systems become vulnerable to unauthorized access. Long and hard passwords are not easy to remember; similarly the probability to guess short and simple passwords is high. Hence for these reasons, password protected systems are not very secure.

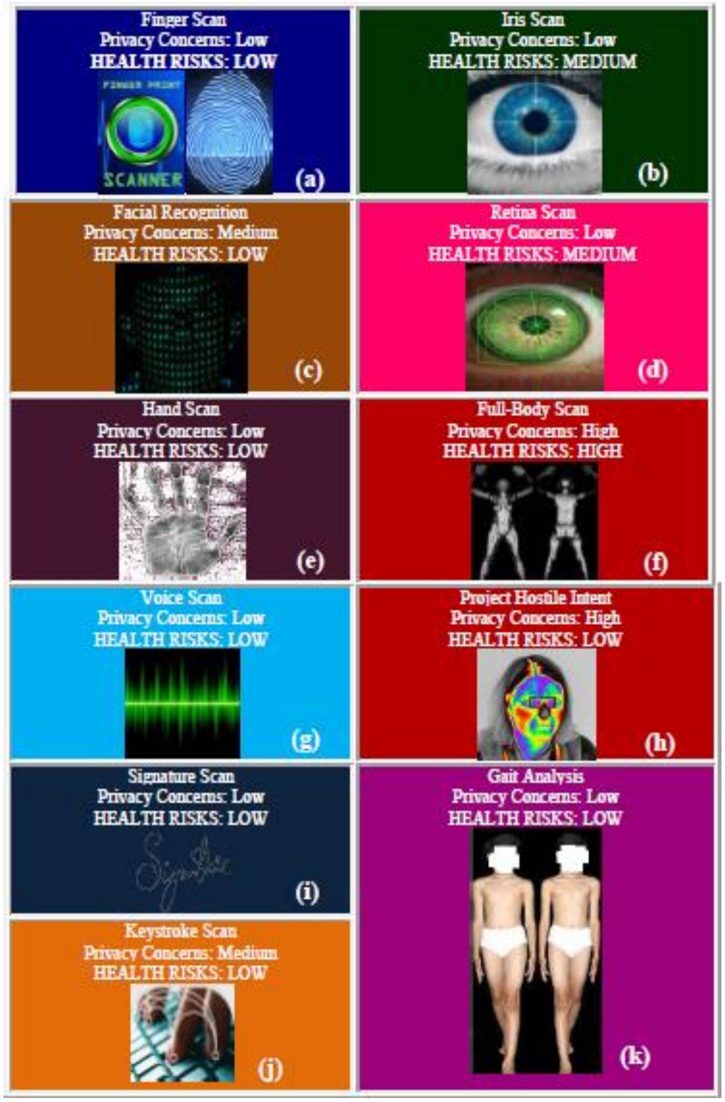


Fig. 1. Different biometric systems (a) Finger scan (b) Iris scan (c) Facial recognition (d) Retina scan (e) Hand scan (f) Full body scan (g) Voice scan (h) Project hostile intent (i) Signature scan (j) Keystroke scan and (k) Gait analysis.

In recent times, biometric-based identification or recognition is being incorporated in security technologies. A biometric feature is a distinctive and computable attribute of a human being that can be used to authenticate an individual. A biometric system is capable to compute both physiological and behavioral patterns of a person for the purpose of recognition [1]. The physiological domain is based on measurement or scan of a part of the human body, e.g. face [2], finger [3], hand [4], iris and retina [5-6]. On the other hand, the behavioral domain is based on measurement from some specific action of a person, e.g., gait characteristics, keystroke scan, signature scan, voice scan [1] and hostile intent [7]. Each of these biometric techniques has its own privacy concern and health risk [8], as shown in Fig. 1. Face recognition system [9] is a recommended technique for person identification as compared with the other biometric systems.

II. Face Recognition Systems

A. Background

Face recognition or identification is a process of matching a face from all available face-images in the system database [9]. Facial identification is a stereo-photogrammetric system, which has less health related issues. It is non-contact, non-destructive and radiation free technique [10]. A face can be recognized even with partial facial information, which can be captured with inexpensive cameras. From the last couple of years, automated face recognition, being one of the prominent applications of facial image analysis has become an attractive domain for researchers.

B. Literature Review

The preliminary work related to human facial identification system described the domain of psychology [11] and engineering [12]. Kelly [13] commenced research for automated facial identification; later Kanade [14] continued the significant work in the same domain. Darwin [15] and Galton [16] described facial recognition technique as a biometric system. Eigenfaces [17], [18] as well as Fisher faces [19]-[21] techniques recognized as a reliable mechanism especially for huge datasets. On the other hand a graph-matching method was one of a useful technique related to feature based method [22], [23]. Face recognition and tracking system in videos are now very interesting area for the last few years [24].

Sun *et al.* [25] used a specialized form of machine learning based on high level face features to identify human faces. Researchers have discussed ways to improve the performance of face recognition techniques [26]-[32]. In [33], [34], classification-based methods were used for the purpose of face identification. The performance of this technique was good. A component-based technique for face recognition is discussed in [35]. Xu *et al.* [36] specified about the change in performance of face recognition techniques with change of environment, lighting and expression of faces.

C. Face Recognition Techniques

Facial recognition systems are now one of more attractive and popular area for the researchers. The techniques required to recognize human faces can be categorized in three diverse approaches, mainly known as: i) Holistic; ii) Feature-based; and iii) Hybrid (Fig. 2).

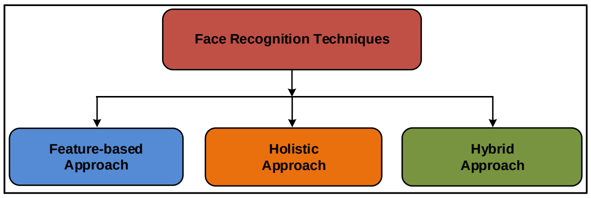


Fig. 2. Categorization of face-recognition techniques.

In this paper, two approaches of machine-based face recognition have been discussed: i) dotted Raster-stereography [37]; ii) feature-based [38]. Dotted Raster-stereography is a holistic approach, in which identification takes place on the basis of global attributes i.e. curvature patterns of human face in term of pixels and their corresponding coordinate values. In featured-based technique, identification is performed on the basis of local features obtained from the human face (nose-eyes-mouth). The research work presented here compares, for these two techniques, the performance measures of accuracy, precision, sensitivity and specificity, using the Image Processing Research Lab (IPRL) database of images.

III. Image Processing Research Lab (IPRL) Database of Images

The images of IPRL database were recorded in Image Processing Research Laboratory (IPRL) at Usman Institute of Technology (UIT), Karachi, Pakistan. This database contains a variety of facial images recorded with different illumination, facial orientation and expression of human faces. Each facial image has its corresponding facial curvatures as well in term of dotted patterns (Fig. 3). This database currently holds 800 human face images with five orientations (frontal, left, right, up and down). It also holds the facial key features of these faces in the form of dots. To describe face recognition based on dotted raster-stereography mean (M) and Gaussian (G) are obtained as decision parameters. Table I presents a snapshot of IPRL database. Fig. 3 shows a snap shot of IPRL database images.

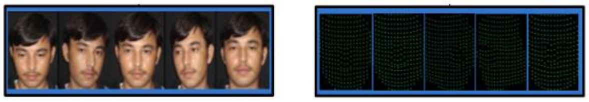


Fig. 3. IPRL Database sample (left: original facial images, right: their dotted-curvature-patterns).

TABLE I. Snapshot of IPRL Database

Attributes	Description
Total no. of human faces	*800 human faces (both male & female)*
Classifications	*4000 human faces, 4000 facial curvatures*
Nature of images	*Static*
Single or Multiple faces	*Single*
Attributes of images	*Colored*

Resolution	*Various*
Facial Pose	*Frontal, left (-15°), right (+15°), down (-15°), up (+15°)*
Facial Expression	*N/A*
Illumination	*N/A*
Light Condition	*Dark Illumination*
Accessories	*With and without Beard, with and without glasses*
3D Data	*N/A*

IV. Dotted Raster-Stereography Technique

The proposed face recognition technique using dotted Raster-stereography consists of three basic steps: registration of face, identification of face; and recognition of face as given in the flow chart shown in Fig. 4.

Dotted raster grid, consisting of green color-dots of equal spacing (0.6 mm), was projected on human face. Because of the curved surface of human face, distortions were created and recorded in term of coordinate values. Fundamental curvatures κ_1 and κ_2 were calculated respectively, which are the main source of curvatures mean (M) and Gaussian (G). These dotted curvature patterns in Table II are easily picks using image processing algorithms. Pixels and their corresponding coordinate values of each dot in curved patterns are representing the face curvature patterns of each person.

A. Mathematical Model of Dotted Raster-stereography Technique

Human faces can be mapped as curved surfaces, curvature patterns were generated when dotted grid (raster) projected on facial surface of human using multimedia projector. Using a digital camera these curvature patterns were recorded and use as an input for the computer system. Designed algorithm was used to find the pixel and corresponding coordinate values in term of (*x*, *y*). Using the developed mathematical model the horizontal and vertical curvatures 'κ_1' and 'κ_2' were calculated. These two facial curvatures were used to calculate decision parameters mean (M) and gaussian (G) curvatures. In our work M and G were the main source of face recognition.

B. Fundamental Curvatures (κ1, κ2)

The fundamental curvatures 'κ_1' and 'κ_2' represents the horizontal and vertical curvatures simultaneously of human face in the face-recognition system described in this work (Fig. 5).

C. Mean and Gaussian Curvatures (M, G)

Mean and Gaussian facial curvatures can be obtained using equations (1) and (2), while $\kappa\rho = 1$, where 'ρ' is the radius of curve generated by facial curvatures of human face.

$$Mean = \frac{\kappa_1+\kappa_2}{2} \ldots\ldots..(1)$$

$$Gaussian = \kappa_1 \times \kappa_2 \ldots\ldots(2)$$

There are four basic types of curvatures of a small surface element [39] as shown in Fig. 4 and the fundamental curvatures (κ_1 and κ_2) decision is described.

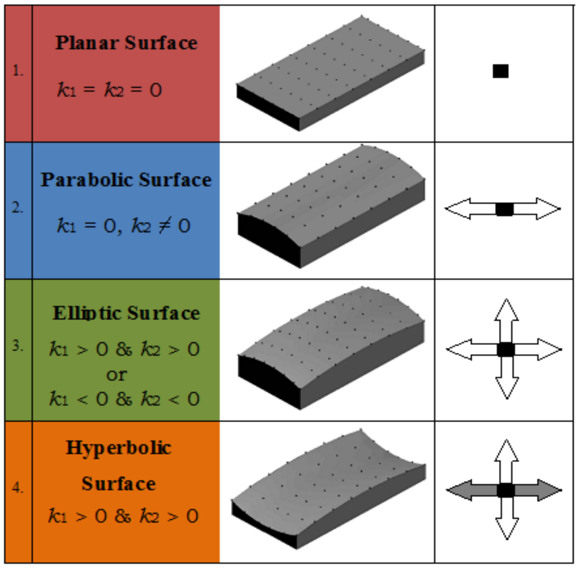


Fig. 4. Four basic types of curvatures

The following Fig. 5 shows the flow of proposed system technique.

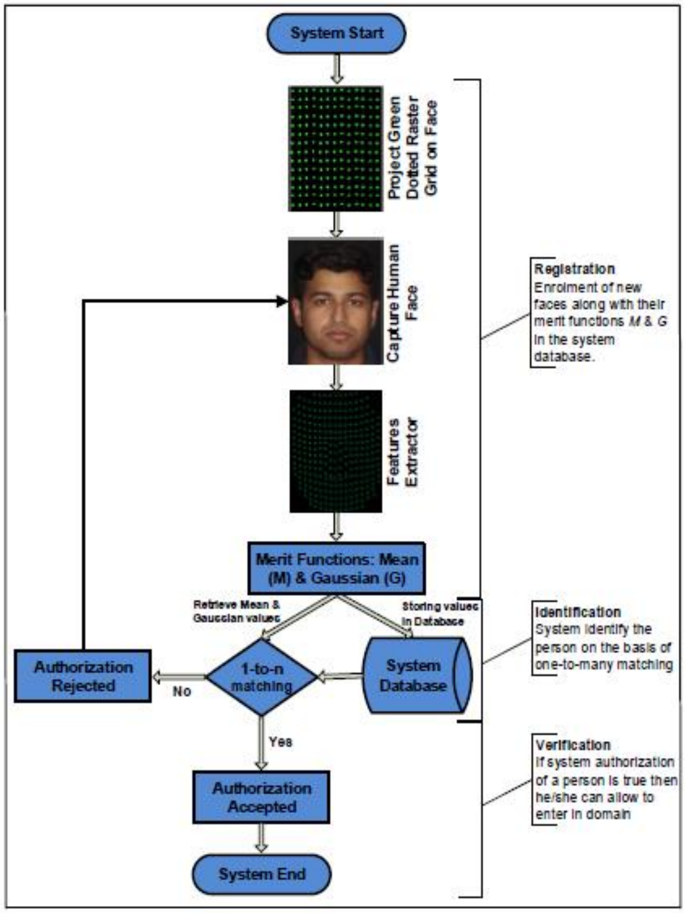
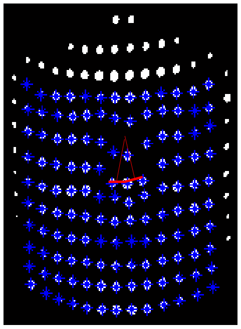

System Start
Project Green Dotted Raster Grid on Face
Capture Human Face
Features Extractor
Merit Functions: Mean (M) & Gaussian (G)
Retrieve Mean & Gaucсian values
Storing values in Database
1-to-n matching
System Database
No
Authorization Rejected
Yes
Authorization Accepted
System End
Registration
Enrolment of new faces along with their merit functions M & G in the system database.
Identification
System identify the person on the basis of one-to-many matching
Verification
If system authorization of a person is true then he/she can allow to enter in domain

Fig. 5. Flow diagram of dotted Raster-stereography system.

TABLE II. OBSERVATION TABLE: CURVATURE PATTERNS, COORDINATE VALUES AND DECISION PARAMETERS

Curvature Patterns	***Pixels for 10×13 order Face Points***	***Coordinate Values (x, y)***


Horizontal and Vertical curvatures (κ_1 & κ_2)
	11 17 29 39 50 59 72 89 ....
	9 19 34 39 52 59 75 98 ....
	11 18 34 47 58 71 82 91 ....
	13 28 31 40 58 67 79 89 ....
	13 19 32 44 57 68 75 88 ....
	14 20 29 39 52 68 82 90 ....
	14 21 30 45 57 69 80 95 ....
	15 24 35 44 62 75 84 99 ....
	16 24 36 51 59 68 80 98 ....
	16 26 36 49 59 70 87 100 ....
		(191.1517, 274.2141) …
		(186.0023, 287.0543) …
		(196.457, 287.1254) …
		(215.3498 292.1123) …

		Mean & Gaussian
		M = 8.0250 cm^{-1} *G =20.1000 cm*$^{-1}$

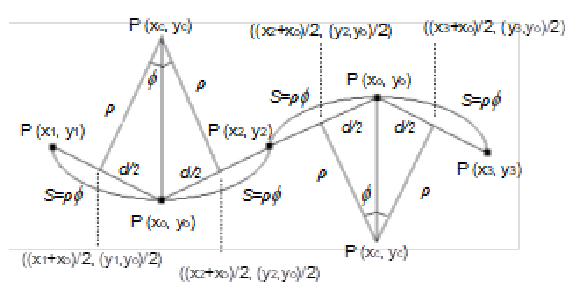


Fig. 6. Mathematics of Raster-stereography.

D. *Mathematics of Curvature Extraction*

In our mathematical model, '*d*' was the small gap (distance) between two points in the original grid. When this raster grid was projected on the face of human, grid were distorted and provided the facial curvature information of human face. When distorted grid generated linear spacing '*d*' is converted into arc length '*s*' (Fig. 6). The study of this arc length provided the detail of facial curvatures of human face. From the geometry of Fig. 6, it is obtained that angular distance = κs and $\kappa\rho$ =1), the result for curvatures κ, we obtained:

$$\kappa \approx \pm\frac{1}{s}\sqrt{24\left(1-\frac{d}{s}\right)} \text{ .... } (3)$$

κ_1 and κ_2 are the horizontal and the vertical curvatures.

V. FEATURE-BASED TECHNIQUE

Feature-Based technique described in this work is based on core features selection of the human face i.e. mouth, nose and eyes. In this technique, edge density (ED) and sum of square for error (SSE) are the two parameters required to recognize a face. In the first step, ED is calculated. An ED can be described as it is an edge, which belongs to the boundaries among two dissimilar classes of intensity. Mathematically ED is a ratio between perimeter (E) and area (A) as given in Fig. 7. This relation is described in the following equation 4.

$$\mathrm{ED} = \frac{\mathrm{E}}{\mathrm{A}} \text{ .... } (4)$$

In second step, the SSE is calculated for the purpose of face recognition [39].

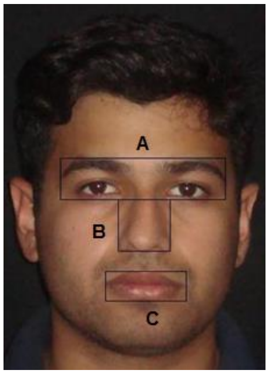


Fig. 7. The feature-based approach is based on edge density.

A. *Mathematical Model of Feature Based Technique*

Edge density (*ED*) is a ratio between perimeter (*E*) and the area (A_e) where perimeter E is the average of perimeters A, B and C; A_e is the area of eye rectangle, whose perimeter is A.

$$\mathrm{ED} = \frac{\mathrm{E}}{\mathrm{A_e}} \qquad (5)$$

where $\mathrm{E} = \frac{(\mathrm{A+B+C})}{3}$

A = perimeter of eye $= 2(\mathrm{L_e} + \mathrm{W_e})$
B = perimeter of nose $= 2(\mathrm{L_n} + \mathrm{W_n})$
C = perimeter of mouth $= 2(\mathrm{L_m} + \mathrm{W_m})$

$\mathrm{L_e}$, $\mathrm{L_n}$ and $\mathrm{L_l}$ are the length of selected rectangular of eyes, nose and mouth.

$$\text{Area} = \mathrm{A_e} = \mathrm{L_e} \times \mathrm{W_e} \qquad (6)$$

B. *Calculation of Sum of Square for Error*

Table III shows the calculation of sum of square for error (SSE) for sample face of *IPRL-UIT-2015061-01* from IPRL Database.

TABLE III. CALCULATION OF SUM OF SQUARE FOR ERROR

$\mathbf{x}$	$\mathbf{y}$	$\mathbf{x - M_x}$	$\mathbf{y - M_y}$	$\mathbf{(x - M_x)^2}$	$\mathbf{(y - M_y)^2}$	$\mathbf{(x - M_x)(y - M_y)}$
67	181	-106.41	-93.08	11323.0881	8663.8864	9904.6428
262	182	88.59	-92.08	7848.1881	8478.7264	-8157.3672
262	231	88.59	-43.08	7848.1881	1855.8864	-3816.4572
68	231	-105.41	-43.08	11111.2681	1855.8864	4541.0628
120	316	-53.41	41.92	2852.6281	1757.2864	-2238.9472
197	231	23.59	-43.08	556.4881	1855.8864	-1016.2572
197	292	23.59	17.92	556.4881	321.1264	422.7328
136	292	-37.41	17.92	1399.5081	321.1264	-670.3872
120	316	-53.41	41.92	2852.6281	1757.2864	-2238.9472
216	315	42.59	40.92	1813.9081	1674.4464	1742.7828
316	351	142.59	76.92	20331.9081	5916.6864	10968.0228
120	351	-53.41	76.92	2852.6281	5916.6864	-4108.2972
2081	**3289**	**0.08**	**0.04**	**71346.9172**	**40374.9168**	**5332.5836**

$$\text{Mean of x} = \mathrm{M_x} = 173.41$$

$$\text{Mean of y} = \mathrm{M_y} = 274.08$$

$$\text{Variance of x} = \mathrm{var(x)} = 5945.57$$

$$\text{Variance of y} = \mathrm{var(y)} = 3364.57$$

$$\text{Co} - \text{variance } = \mathrm{cov(x,y)} = 484.78$$

$$\frac{\mathrm{cov}(x,y)}{\mathrm{var}(x)} = 0.0815$$

$$\mathrm{SSE} \ = (\mathrm{n}-1)[\mathrm{var(y)} - \frac{\mathrm{cov(x,y)}}{\mathrm{var(x)}}] = 370.0937$$

VI. RESULTS

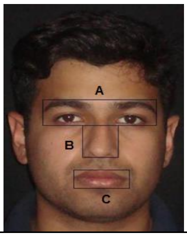
A. *Results of Dotted Raster-stereography and Feature-Based Technique*

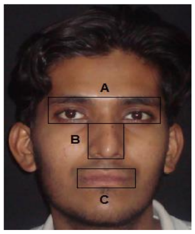
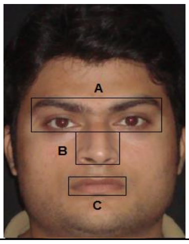
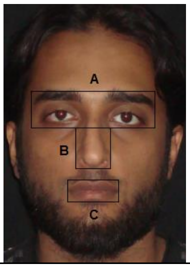
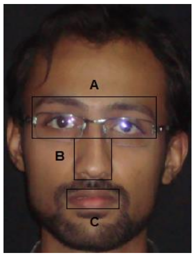
Table IV shows the results of dotted Raster-stereography and feature-based techniques.

B. *Accuracy, Precision, Sensitivity and Specificity of Dotted Raster-stereography and Feature-Based Systems*

This dotted raster-stereography technique is based on Mean and Gaussian curvatures. Both mean and Gaussian curvatures are based on two fundamental curvatures κ_1 and $\kappa_{2,}$ as shown in Table V.

TABLE IV. RESULTS OF DOTTED RASTER-STEREOGRAPHY AND FEATURE-BASED TECHNIQUES

S. No	**Face ID**	**Sample Faces**	**Raster-stereography Technique**		**Sample Faces**	**Feature-Based Technique**	
			Mean (M) ***cm^{-1}***	**Gaussian (G)** ***cm^{-1}***		**Edge Density (ED)** ***cm^{-1}***	**Sum of Square Error (SSE)**
1	IPRL-UIT-2015061-01		03.0500	06.0600		0.0355	370.0937

2	IPRL-UIT-2015061-02		8.0250	20.1000		0.0427	869.156
3	IPRL-UIT-2015061-03		13.0500	48.1100		0.0353	866.246
4	IPRL-UIT-2015061-04		12.0000	33.7500		0.0364	1238.918
5	IPRL-UIT-2015061-05		15.6000	55.6700		0.0322	898.414

$$\text{Mean}= \text{M}= (\kappa_1 + \kappa_2)/2$$
$$\text{Gaussian}=\text{G}= \kappa_1 \times \kappa_2$$

In our test run, 100 faces have been tested in IPRL using dotted Raster-stereography and feature-based techniques. For dotted raster-stereography, 95 faces were correctly recognized. Using feature-based technique, 82 faces were correctly recognized. Table IV shows the facts recorded during test runs for dotted Raster-stereography and feature-based systems.

Table VI shows the results of recognition rate, training and testing time of both the techniques using IPRL database. Runtime of image normalization and alignment excludes the training and testing times.

TABLE V. FACTS RECORDED DURING TEST RUNS IN IPRL DATABASE FOR DOTTED RASTER-STEREOGRAPHY AND FEATURE-BASED TECHNIQUES

Parameters	**Dotted Raster-stereography**	**Feature-Based**	**Parameters**	**Dotted Raster-stereography (%)**	**Feature-Based (%)**
Number of true positive (TP)	95	82	$Accuracy = \frac{TP+TN}{TP+FP+FN+TN}$	96.00	88.00
Number of true negative (TN)	01	6	$Precision = \frac{TP}{TP+FP}$	96.93	91.11
Number of false positive (FP)	03	8	$Sensitivity\ = \frac{TP}{TP+FN}$	98.95	95.34

Number of false negative (FN)	01	4	$Specificity = \frac{TN}{FP + TN}$	25.00	42.85
-	-	-	$Positive\ predictive\ value = \frac{TP}{TP + FP}$	96.93	91.11
-	-	-	$Negative\ predictive\ value = \frac{TN}{FN + TN}$	50.00	60.00

TABLE VI. RECOGNITION RATE AND TRAINING & TESTING TIMES FOR BOTH TECHNIQUES

Technique	Recognition Rate	Training Time (second)	Testing Time (second)
Dotted Raster-stereography	96.00 %	260.5	2.1
Feature-Based	88.00 %	70	1.2

VII. CONCLUSIONS

Each of Dotted Raster-stereography and feature based techniques was tested for 100 faces in Image Processing Research Laboratory, using IPRL database. In case of dotted Raster-stereography technique, following values of performance measures were found: accuracy 96.00 %, precision 96.93%, sensitivity 98.95% and specificity 25.00%. For feature-based technique, the same measured values were: accuracy 88.00%, precision 91.11%, sensitivity 95.34% and specificity 42.85%. Consequently, Dotted Raster-stereography technique is a better approach for face recognition as far as these performance measures are concerned. Feature-based technique is faster in terms of the training and testing times required. Thus overall, such sensitive applications where high face recognition accuracy is required, dotted raster-stereography should be preferred. On the other hand, such applications where volume of face recognition work is large and high speed is required with some compromise in accuracy being acceptable, then feature-based technique may also be the technique of choice.

REFERENCES

[1] Delac, Kresimir, and Mislav Grgic. "A survey of biometric recognition methods." Electronics in Marine, 2004. Proceedings Elmar 2004. 46th International Symposium. IEEE, 2004.

[2] Tistarelli, M. and E. Grosso (1997). Active face recognition with a hybrid approach. Pattern Recognition Letters, 18: 933-946.

[3] Cappelli, Raffaele, Matteo Ferrara, and Davide Maltoni. "Minutia cylinder-code: A new representation and matching technique for fingerprint recognition." IEEE Transactions on Pattern Analysis and Machine Intelligence 32.12 (2010): 2128-2141.

[4] Bulatov, Yaroslav, et al. "Hand recognition using geometric classifiers." Biometric Authentication. Springer, Berlin, Heidelberg, 2004. 753-759.

[5] Lim, Shinyoung, et al. "Efficient iris recognition through improvement of feature vector and classifier." ETRI journal23.2 (2001): 61-70.

[6] Ali, Jafar MH, and Aboul Ella Hassanien. "An iris recognition system to enhance e-security environment based on wavelet theory." AMO-Advanced Modeling and Optimization 5.2 (2003): 93-104.

[7] Ko, Teddy. "A survey on behavior analysis in video surveillance for homeland security applications." Applied Imagery Pattern Recognition Workshop, 2008. AIPR'08. 37th IEEE. IEEE, 2008.

[8] Wasim, M. and Shaikh A. Book on "Raster-stereography based partial face recognition". Lambert Academic Publishing (2014).

[9] Vilas H. Gaidhane, Yogesh V. Hote, Vijander Singh. "An efficient approach for face recognition based on common eigenvalues". Pattern Recognition 47 (2014) 1869-1879. Elsevier.

[10] Jafri, R. and Arabnia, H. R. 2009. A survey of face recognition techniques. Journal of information processing system, 5 (2), 41-68.

[11] Bruner, I. S. and Tagiuri., 1954. The perception of people. In Handbook of Social Psychology, Vol. 2, G. Lindzey, Ed., Addison-Wesley, Reading, MA, 634–654.

[12] Bledsoe, W. W., 1964. The model method in facial recognition. Tech. rep. PRI:15, Panoramic research Inc., Palo Alto, CA.

[13] Kelly, M. D. 1970. Visual identification of people by computer. Tech. rep. AI-130, Stanford AI Project, Stanford, CA.

[14] Kanade, T., 1973. Computer recognition of human faces. Birkhauser, Basel, Switzerland, and Stuttgart, Germany.

[15] Darwin, C., 1972. The Expression of the Emotions in Man and Animals. John Murray, London,U.K.

[16] Galton, F., 1988. Personal identification and description. Nature, (June 21), 173–188.

[17] Kirby, M., and Sirovich, L. 1990. Application of the Karhunen-Loeve procedure for the characterization of human faces. IEEE Trans. Patt. Anal. Mach. Intell. 12.

[18] Turk, M., and Pentland, A. 1991. Eigenfaces for recognition. J. Cogn. Neurosci. 3, 72–86.

[19] Belhumeur, P. N., Hespanha, J. P., and Kriegman, D. J. 1997. Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection. IEEE Trans. Patt. Anal. Mach. Intell. 19, 711–720.

[20] Etemad, K. and Chellappa, R. 1997. Discriminant analysis for recognition of human face images. J. Opt. Soc. Am. A 14, 1724–1733.

[21] Zhao,W., Chellappa, R., and Krishnaswamy, A. 1998. Discriminant analysis of principal components for face recognition. In Proceedings, International Conference on Automatic Face and Gesture Recognition. 336–341.

[22] Wiskott, L., Fellous, J.-M., and Von Der Malsburg, C. 1997. Face recognition by elastic bunch graph matching. IEEE Trans. Patt. Anal. Mach. Intell. 19, 775–779.

[23] Cox, I. J., Ghosn, J., and Yianilos, P. N. 1996. Feature-based face recognition using mixturedistance. In Proceedings, IEEE Conference on Computer Vision and Pattern Recognition. 209–216.

[24] Lu, Ke, et al. "Video-based face recognition." Image and Signal Processing (CISP), 2010 3rd International Congress on. Vol. 1. IEEE, 2010.

[25] Yi Sun, Xiaogang Wang & Xiaoou Tang. Deep Learning Face Representation from Predicting 10,000 Classes. IEEE Conference on Computer Vision and Pattern Recognition, 2014.

[26] Junlin Hu1, Jiwen Lu, Yap-Peng Tan. Discriminative Deep Metric Learning for Face Verification in theWild. IEEE Conference on Computer Vision and Pattern Recognition, 2014.

[27] Z. Cui, W. Li, D. Xu, S. Shan, and X. Chen. Fusing robust face region descriptors via multiple metric learning for face recognition in the wild. In CVPR, pages 3554–3561, 2013.

[28] M. Guillaumin, J. Verbeek, and C. Schmid. Metric learning approaches for face identification. In ICCV, pages 498–505, 2009.

[29] G. B. Huang, H. Lee, and E. G. Learned-Miller. Learning hierarchical representations for face verification with convolutional deep belief networks. In CVPR, pages 2518–2525, 2012.

[30] K. Simonyan, O. M. Parkhi, A. Vedaldi, and A. Zisserman. Fisher vector faces in the wild. In BMVC, 2013.

[31] L. Wolf, T. Hassner, and I. Maoz. Face recognition in unconstrained videos with matched background similarity. In CVPR, pages 529–534, 2011.

[32] L. Wolf and N. Levy. The svm-minus similarity score for video face recognition. In CVPR, pages 3523–3530, 2013.

[33] Can-Yi Lu, Hai Min, Jie Gui, Lin Zhu & Ying-Ke Lei. Face recognition via Weighted Sparse Representation. J. Vis. Commun. Image R. 24, 111–116, Elsevier, 2013.

[34] J. Wright, A.Y. Yang, A. Ganesh, S.S. Sastry, Y. Ma. Robust face recognition via sparse representation. IEEE Transactions on Pattern Analysis and Machine Intelligence 31, 210–227, 2009.

[35] Kathryn Bonnen, Brendan F. Klare, Anil K. Jain. Component-Based Representation in Automated Face Recognition. IEEE transactions on information forensics and security, vol. 8, no. 1, 2013.

[36] Yong Xu, Xiaozhao Fang, Xuelong Li, Jiang Yang, Jane You, Hong Liu, Shaohua Teng. Data Uncertainty in Face Recognition. IEEE transactions on cybernetics, vol. 44, no. 10, 2014.

[37] Wasim M., Kamal S. A., Shaikh A., A Security System Employing Edge-Based Raster-stereography. Int. j. biol. biotech. 10(4): 483-501, 2013.

[38] Reza Moradi Rad, Abdolrahaman Attar, Reza Ebrahimi Atani. A Robust Face Recognition Method Using Edge-Based Features. IEEE symposium on computers & informatics, 2012.

[39] W. Frobin, E.Hierholzer, B. Drerup. Shape Analysis of Surfaces: Extraction of Shape from Coordinate Data. Moiré Fringe Topography, Gustav, Stuttgart, New York, 1983.

Empirical Evaluation of Modified Agile Models

Shabib Aftab, Zahid Nawaz, Faiza Anwer, Muhammad Salman Bashir, Munir Ahmad, Madiha Anwar
Department of Computer Science
Virtual University of Pakistan
Lahore, Pakistan

***Abstract*—Empirical evaluation is one of the widely accepted validation method in the domain of software engineering which investigates the proposed technique via practical experience and reflects its benefits and limitations. Due to various advantages, agile models have been taking over the conventional software development methodologies since last two decades. However besides the benefits, various limitations have been noticed as well by the researchers and software industry in agile family. To achieve the maximum benefits it is vital to fix the limitations by customizing the development structure of agile models. This paper deals with the empirical analysis of modified agile models called Simplified Extreme Programing (SXP) and Simplified Feature Driven Development (SFDD), which are the modified forms of Extreme Programing (XP) and Feature Driven Development (FDD). SXP was presented to eliminate the issues of conventional XP such as, lack of documentation, poor architectural structure and less focus on design. SFDD was proposed to take care of reported issues in FDD such as explicit dependency on experienced staff, little or no guidance for requirement gathering, rigid nature to accommodate requirement changes and heavy development structure. This study evaluates SXP and SFDD through implementing client oriented projects and discusses the results with empirical analysis.**

***Keywords*—Agile models; SXP; SFDD; Modified XP; modified FDD; empirical evaluation; comparative analysis**

I. Introduction

Conventional software process models are replaced by lightweight agile development methodologies. The reason behind the widely acceptance of agile family by the software industry is the features these models provide such as: light weight approach for development, early delivery of partially working software (module), welcome changes at any stage of development and quick response. Agile models shifted the focus from process to people and valued those factors which were neglected by traditional models [7], [25], [26]. Some of the famous agile models are: Extreme Programming (XP), Scrum, Test Driven Development (TDD), Dynamic System Development Model (DSDM), Crystal methods and Feature Driven Development (FDD), etc. [7], [8]. These models follow the values, principles and practices given by agile manifesto which is considered a parent document of all agile models and contains twelve foundation principles of software development. XP and FDD, both are the widely used agile models in software industry [12], [40]. XP was developed by Kent Beck and mainly focuses to overcome the limitations of traditional software process models. The working of XP consists of certain principles, values and practices, which work together rigorously to develop high quality software [9], [29], [34], [35], [39]. XP provides a flexible and adaptive development approach which can handle the changing business needs in an effective way due to its well-known requirements gathering technique, "story cards". Its 12 practices provide the guidelines to govern the whole development process in an effective and efficient way. Besides the advantages, XP reflects some limitations as well. Drawbacks of XP include poor architecture, weak system design and lack of documentation [29], [32], [36], [37]. Moreover its practices: ‘pair programming’ and ‘on-site customer’ are controversial and cannot be applicable in every situation [38], [39]. Due to these drawbacks, XP is suitable only for small scale and low risk projects. On the other hand FDD follows the process oriented approach [9]-[11]. It is highly adaptive and mainly focuses on design and building aspects of development. As its name reflects, features are the basic building blocks of this model. Feature is considered as a functionality which user wants in the software. Benefits provided by FDD model includes the iterative and incremental approach along with ETVX pattern which ensures the development of high quality software according to client valued features. However along with advantages, some limitations of FDD were also reported such as: little or no guidance for requirement gathering, explicit dependency on experience staff, rigid nature to handle changing requirements and heavy development structure including various activities and team roles. All these issues make it only suitable for medium or large scale projects. SXP [40] and SFDD [12] were proposed to overcome the limitations of XP and FDD respectively. This study empirically valuates the proposed models through empirical case studies conducted in software industry.

II. Related Work

Drawbacks of agile models have to be eliminated in order to achieve the maximum benefits, for this purpose many researchers have proposed the modifications in agile models. XP and FDD were discussed and optimized in many studies from which some of are discussed here. In [13], researchers presented the Tailored Extreme Programming (TXP) model which was specifically designed for small scale projects where requirements have fewer or no tendencies to change. In [14], researchers proposed the feature of reusability in XP model. They introduced a framework to add the ability of component based architecture refinement reusability in traditional XP. The used framework provided a way to develop simple and loosely coupled design which can be modified easily in future. Researchers in [15] customized the XP by introducing parallel refinement iteration to the development activities in order to enhance the quality; however the proposed model is not suitable for software projects having a lot of inter dependencies among modules. In [16], authors customized the software maintenance model by using many XP practices such as: on-site customer, planning game, small releases, pair programming, metaphor, test driven development and refactoring. In [18], researchers integrated Personal Software Process (PSP) with XP. The proposed model introduced "Personal Planning Phase" in which developer can plan the activities by using PSP practices. Six important practices from each model (XP and PSP) are integrated in proposed model. In [19], XP was customized to develop medium scale projects with large team by eliminating its drawbacks such as weak design and lack of documentation. Moreover a phase named "Analysis and Risk Management" was introduced to handle the failure risks. In [30], Analytical Hierarchy Process (AHP) was used with CRC cards during designing phase of XP. AHP was used to design a systematic approach of CRC cards prioritization. AHP is a hierarchal model consists of five steps which reflect the human thinking process. By using AHP the developers can select, design and implement the most important classes first. In [31], XP was customized for medium to large scale projects. The research highlighted the drawbacks of classical XP such as weak design, poor architecture, lack of risk management and lack of documentation. These issues of XP make it suitable only for small scale projects. To eliminate these issues, new phases were introduced in modified proposed model. Author in [1] proposed Feature Driven Reuse Development (FDRD), an enhanced version of FDD which considered re-useable feature-sets for development along with the new requirements. Author in [2] presented Competitor Driven Development (CDD), a hybrid process model which integrated the practices of Extreme Programming (XP) and Feature Driven Requirement Reuse Development (FDRD). The proposed model is a self-realizing requirement generation model which keeps track of market trends as well as competitor's next product launch to extract requirements. Moreover CDD considers the market orientation of product to guess the product's success rate. In [3], authors proposed a hybrid model SCR-FDD, an integration of Scrum and FDD. The proposed model covered the imitations of both models by taking the schedule related aspects from Scrum and quality related aspects from FDD. In [4], researchers presented Feature-Driven Methodology Development (FDMD), a modified version of FDD which integrated the features of object oriented approach with Situational Method Engineering (SME). In the proposed model requirements are represented as features, which are based on object oriented principles. The feature is defined by using action, result and object. Authors in [5] proposed Secure Feature Driven Development (SFDD), an enhanced version of FDD which introduced some changes in classical FDD to cover security related issues. The proposed model introduced two phases in classical FDD named "Build security by feature" and "Test security by feature" along with the "In-phase Security" element in each phase. Moreover, a new role is also added called security master to ensure the secure software development. Authors in [6] proposed an ontology based approach in FDD for semantic web application. The proposed model used the concepts of domain ontology from domain knowledge modeling. Ambiguity and inconsistency regarding Language is handled by RDF and OWL however the agility of FDD can be compromised by adding the concepts of domain ontology in each phase.

III. Modified Agile Models

The proposed Simplified Extreme Programming (SXP) is focused to overcome the limitations of classical XP. It provides more flexible and simple approach for small to medium scale projects. The issues of pair programming and on-site customer are handled in an effective way. On the other hand, SFDD [12] was proposed to overcome the limitations of FDD such as explicit dependency on experienced staff, little or no guidance for requirement gathering, rigid nature to accommodate changes in requirements, heavy development structure. SFDD focused on small to medium scale projects along with an effective requirement elicitation technique of story cards which simplified the requirement change process. Both the proposed models are briefly explained below.

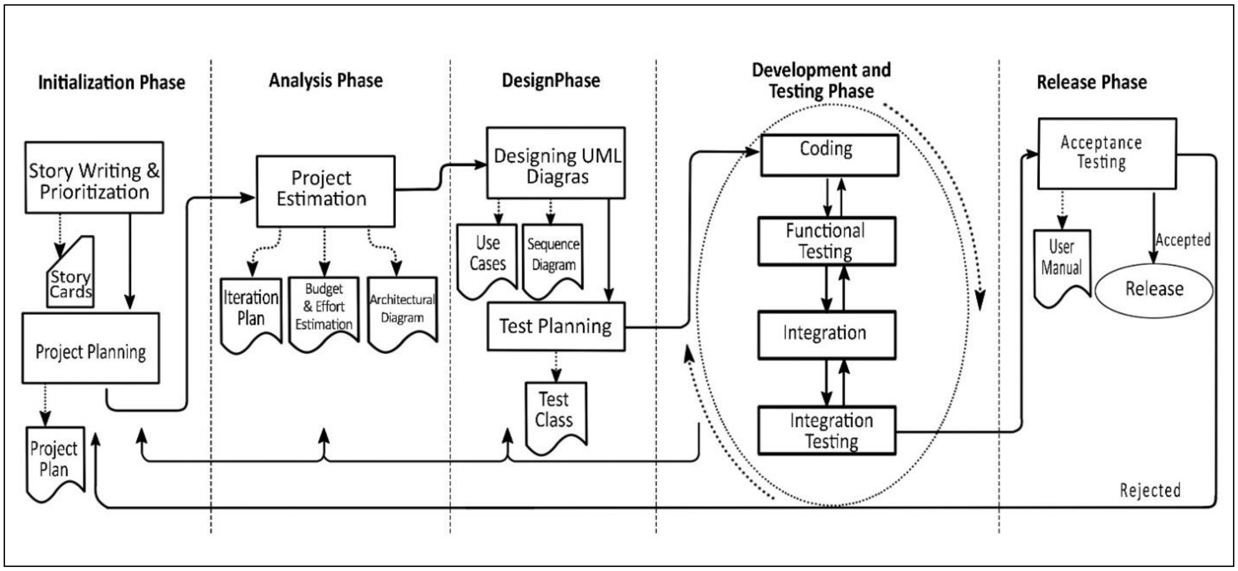


Fig. 1. SXP.

A. SXP

Simplified Extreme Programing (SXP) consists of five phases; Initialization, Analysis, Design, Development & Testing and Release as shown in Fig. 1. In the proposed model, customer involvement is restricted to initialization and release phase only and all other phases are executed by development team with the complete coordination. Necessary documentation is produced during each phase that helps to resolve change management issues. "Initialization" is the first phase of SXP and is responsible to extract and manage the requirement as well as to create an overall plan for project. Requirements are extracted and managed through story cards, a story card consists of following features: functionality name, type, priority and the short description without any technical detail. Type defines whether the functionality is functional and nonfunctional and priority is assigned with number so that higher priority features can be developed in early iterations. Project planning includes the decisions regarding project scope, cost and tools to be used for the development. "Analysis" is the second phase and deals with budget and schedule related activities which are performed by development team only. In this phase required budget is estimated and documented. An iteration plan is also formed which includes the detail about number of iterations, number of stories implemented in each iteration and the time of each iteration. A training session is also conducted to make the development team familiar with the tools and technology (if the team members are not already familiar). "Design Phase" is third phase of SXP which deals with two activities: “Designing UML Diagrams” and “Test Planning”. Conventional XP does not include any documentation which makes requirement change management very difficult. This issue is effectively solved by SXP by focusing on system design with use case diagrams and sequence diagrams. Test cases are also developed in this phase. Writing tests prior to code help the development team to understand different design opportunities. "Development and Testing" is the fourth phase and works in an iteratively. Activities of this phase include coding, functional testing, integration and integration testing. Developer writes the code for selected stories by keeping in view the design document which was developed during design phase. Functional testing is performed by using test cases, developed during test planning activity. Coding activity is repeated if any issue is reported in functional testing. These tests are performed by programmers and results are noted to keep the track of defects. Code is integrated with previous developed module in case of successful functional testing followed by another testing known as integration testing. "Release" is the last phase in which customer performed acceptance testing. The developed workable product is released after the customer’s approval along with the User manual. If the customer is not satisfied with the developed product then whole development process can be repeated again with changed or modified set of requirements.

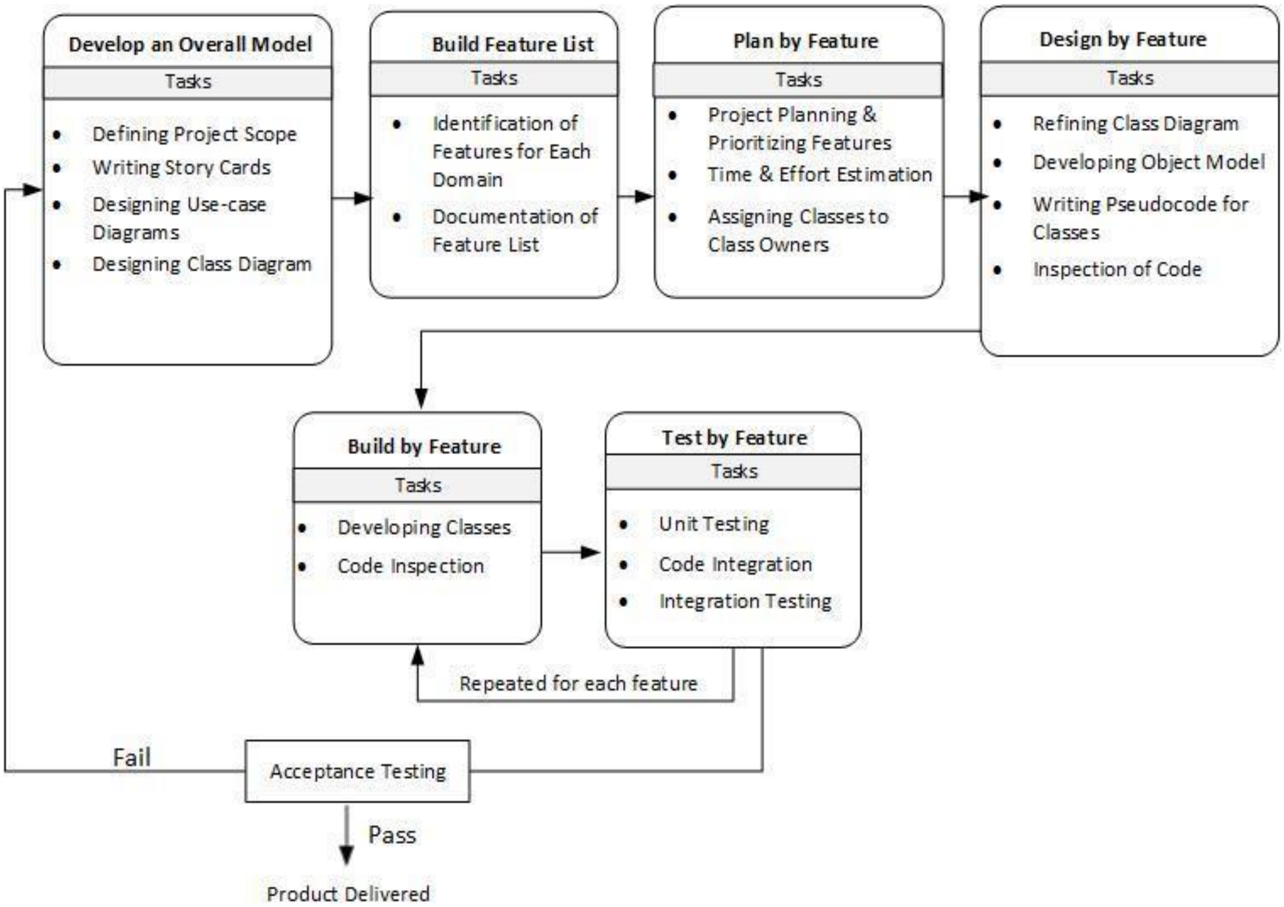


Fig. 2. SFDD.

B. SFDD

Simplified Feature Driven Development (SFDD) consists of six phases and various activities as shown in Fig. 2. "Develop an Overall Model" is the first phase which deals with the identification of requirements and scope of project. Domain expert and chief programmer are the main roles of this phase. Domain expert provides the project requirements through story cards and chief programmer finalizes the project scope by keeping in view the provided requirements moreover use case diagrams and class diagrams are also developed in this phase. "Build Feature List" is the second phase of SFDD and deals with the extractions of features from the documents developed in first phase. Features are basically the functions which a customer wants in the software. Related features are collected in a list called feature list. Chief programmer converts the requirements in to feature lists in this phase. "Plan by Feature" is the third phase which deals with the project planning activities and starts with a meeting where domain expert and chief programmer finalize the budget and time frame of the project. Chief programmer further finalizes the number of iterations and assigns features to iterations by keeping in view the priorities. This phase also includes the estimation of effort (resource persons) and hardware/software resources which are needed for the project. At the end of the phase classes are assigned to class owners (developers). "Design by Feature" is the fourth phase and deals with the process of refining the class diagrams developed in the first phase. Object model is finalized in this phase and class owner completes the pseudo code for the assigned classes. To ensure the quality, a role of QA manager is introduced in this phase. "Build by Feature" is the fifth phase of model and first phase of iteration. Development actually starts in this phase according to the pseudo code, written in previous phase. QA manager makes sure that the developing module is according to the features. Test by feature is the last phase of model and second phase of iteration which deals with the testing activities and starts with unit testing to make sure that the developed module is bug free and working properly, if passed then integrated with already developed module.

Integration testing is then performed to check the integrated working of modules. Finally domain expert performs the acceptance testing. Proposed model simplified the structure of FDD through effective customization.

IV. Empirical Evaluation

This research aims to perform the empirical evaluation of proposed modified agile models. For this purpose two case studies are conducted in which both models, SXP and SFDD were used to develop small scale web based projects. The selected case studies were part of an empirical research project in which multiple agile models were used to develop various client oriented applications in a software house, situated in Islamabad, capital of Pakistan. The software house consists of experienced staff with dominating knowledge of software development along with higher degrees in computer science disciplines. The developers were using agile methods for most of the projects. Both case studies were implemented in same working environment but with different teams. Most of the characteristics of applications are same such as size of project, no of iterations, no of team members, and the tools used in development. The detail regarding the characteristics of developed projects is given in Table I. The case study of SXP is implemented by the team which had significance experience of agile development. On the other hand, to implement the SFDD, the chosen team had less or no experience of agile development however training session of 10 days was organized.

For SFDD, less experienced team was selected as the authors of proposed model (SFDD) claimed that the issue in classical FDD regarding the dependency on experienced staff has been eliminated. The detailed empirical results collected during the development are shown in Table II. Partial and aggregated results of selected case studies are discussed in [39], [33]. However this paper demonstrates the complete results of empirical experiment including all the iterations by keeping in view the guidelines extracted from [17], [27], [28], [19]. Both case studies are implemented with four iterations. After each iteration, partial working software (module) was released for the client.

TABLE I. Case Studies Detail

Characteristics	**SXP**	**SFDD**
Product Type	Human Resource Management	Human Resource Management
Size	Small	Small
Iterations	4	4
Programming Approach	Object Oriented	Object Oriented
Language	C#, ASP.NET	C#, ASP.NET
Documentation	MS Office	MS Office
Testing	Browser Stack	Browser Stack
Web Server	IIS	IIS
Project Type	Average	Average
Team Size	5 Member	5 Member
Feedback	Weekly	Weekly
Development Environment	Visual Studio 2012	Visual Studio 2012
Other Tools	MS Visio	MS Visio
Reports	Crystal Report	Crystal Report

TABLE II. EMPIRICAL RESULTS

Sr. No	Software Metric	Release 1		Release 2		Release 3		Release 4		Total	
		SXP	SFDD	SXP	SFDD	SXP	SFDD	SXP	SFDD	SXP	SFDD
1	Completion Time (weeks)	1	0.9	0.9	0.8	0.9	0.8	1	0.7	3.8	3.2
2	Number of Modules	2	1	1	1	1	1	2	1	6	4
3	No of User Stories	8	21	4	20	3	15	6	9	21	65
4	Budgeted Work Effort (h)	200	180	180	160	180	160	200	140	760	640
5	Actual Work Effort (h)	180	180	165	147	175	140	175	125	695	592
6	Number of User Interfaces	6	3	3	3	3	2	2	2	14	10
7	No of Classes	4	7	3	5	2	4	2	4	11	20
8	Lines of Code	820	4300	734	3450	860	2760	646	2600	3060	13110
9	KLOC	0.820	4.3	0.734	3.4	0.860	2.7	0.646	2.6	3.060	13.1
11	No of Code Integrations	10	7	8	5	12	3	7	3	37	18
12	Post Release Defects	2	2	4	1	6	1	3	1	15	5
13	Post Release defects / KLOC	2.4	0.465	5.45	0.294	6.97	0.37	4.64	0.38	4.902	0.381
14	Productivity (= line of code/ actual time spent in hours)	4.56	23.88	4.44	23.46	4.91	19.71	3.69	20.80	4.4	22.14
16	No of Pre-release Change Requests	2	6	3	3	4	1	1	2	10	12
17	Total Change requests/KLOC	2.44	1.395	4.09	0.882	4.65	0.370	1.55	0.769	3.27	0.916
18	Time to Implement Changes (h)	3	4	2	3	4	3	2	1	11	11

The second column of Table II represents the attributes/metrics which are measured in each release for both the models and the last column contains the cumulative/average values of metrics from all four releases. The remaining columns (release 1 to release 4) present the values of metrics (column 2) in each release for SXP and SFDD. Metrics are used to measure the software in terms of development, cost, working, productivity, quality, effectiveness and efficiency from various aspects [20]-[24].

V. CRITICAL ANALYSIS

From the detailed empirical results (Table II), significant differences can be seen among the performances of both the models. Even though the working environment as well as the size and nature of both the applications were same, but SFDD performed much better than SXP. KLOC of the application developed using SXP are 3.069 with the implementation of 21 user stories however on the other hand SFDD implemented 65 user stories with 13.1 KLOC (Fig. 3 and 4).

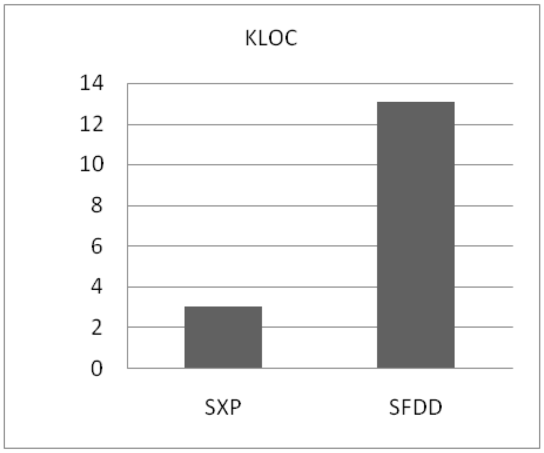


Fig. 3. KLOC.

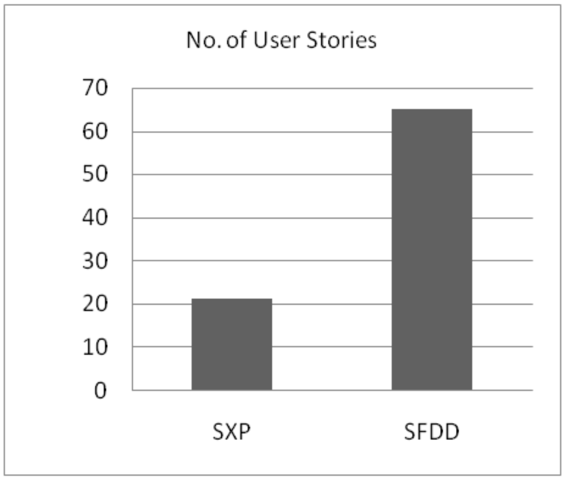


Fig. 4. Implemented user stories.

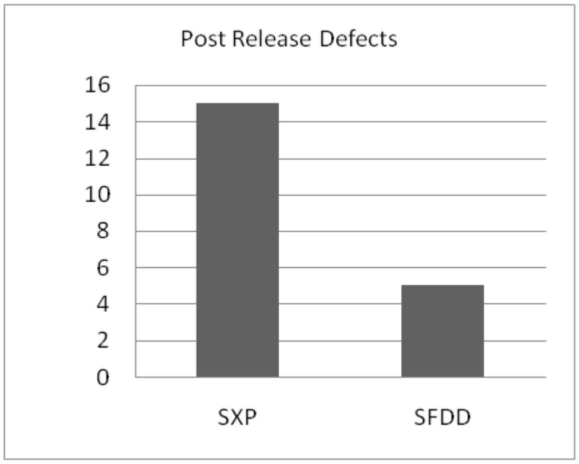


Fig. 5. Post release defects.

No. of post release defects is an important software metric which reflects the quality of developed application as well as the satisfaction of customer. After the release 15 defects were reported in the application developed using SXP however only 5 defects were reported in the application developed using SFDD (Fig. 5).

Time to implement pre-release change requests is also considered as one of the important quality metric which reflects the change management feature of software process model. 10 pre-release changes were proposed during SXP case study which took 11 hours to implement however no. of pre-release change requests in SFDD case study were 12 which took the same time for the implementation (Fig. 6) as in SXP (11 hours).

Software productivity reflects the team effort during the application development. Productivity of the application developed by SXP was far lower than the application of SFDD (Fig. 7). During SXP case study, 3060 lines of code were written in 695 hours (Actual Work Effort) with the productivity of 4.4 however during the implementation of SFDD, 13110 lines were written in 592 hours and reflected productivity of 22.14. As compared to SFDD, SXP showed very poor performance by keeping in view the empirical results. SFDD performed very well according to all software parameters (Table II) even with the team having less experience with agile methodologies. There might be various reasons of poor performance in SXP case study. Complexity level of the developed application in SXP case study may be higher than the application of SFDD however according to best of our knowledge the nature and complexity level of both the application were same. As the performance of SXP is lower in every release so, there might be issues in code integrations as there were total of 37 integrations in SXP and only 18 in SFDD. Moreover the issue of communication among the team members can also be a reason of lower performance. The issue of awareness with agile development cannot be considered as the team of SXP was experienced with agile and team of SFDD had less or no experience with agile development.

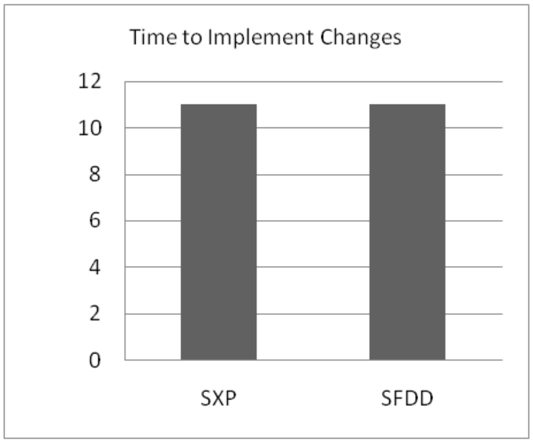


Fig. 6. Time to implement pre-release change requests.

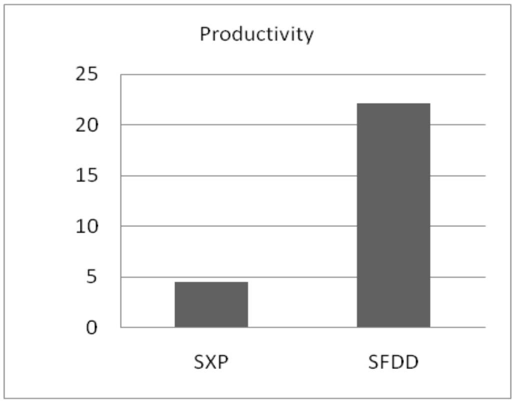


Fig. 7. Productivity.

VI. Conclusion and Future Work

This paper evaluated the proposed modified agile models, SXP and SFDD through empirical case studies. SXP focused to reduce the reported issues of conventional XP such as: Lack of documentation, poor architectural structure and less focus on design. Due to these issues, XP is only suitable for small scale and low risk projects. SFDD has taken care of the issues reported in FDD, such as explicit dependency on experienced staff, no guidance for requirement gathering, rigid nature to accommodate requirement changes and heavy development structure. Empirical analysis was performed via development of client oriented projects by using SXP and SFDD. Both projects were related to Human Resource Management (HRM) and also were same in nature as well as in size and complexity level. The development team for SXP case study was experienced in agile development however for SFDD case study the chosen team had less experience of agile as the proposed SFDD eliminated the dependency on experienced staff. According to empirical results, SFDD performed much better than SXP even with the less experienced team. In comparison of SFDD, SXP performance was very poor in each metric such as lines of code, implemented user stories, post release defects, productivity and time required to implement pre-release change requests. There might be various reasons of poor performance of SXP model such as complexity level, integration issues and communication problems within the development team. It is suggested that both the models should be further tested with large and complex projects.

References

[1] S. Thakur and H. Singh, “FDRD: Feature driven reuse development process model,” in Proceedings of 2014 IEEE International Conference on Advanced Communication, Control and Computing Technologies, ICACCCT 2014, 2015, pp. 1593–1598.

[2] V. P. Doshi and V. Patil, “Competitor driven development: Hybrid of extreme programming and feature driven reuse development,” 1st Int. Conf. Emerg. Trends Eng. Technol. Sci. ICETETS 2016 - Proc., no. Cdd, p. 7602985, 2016.

[3] S. Ali, S. S. Tirumala, and A. Babu G, “A Hybrid Agile model using SCRUM and Feature Driven Development,” Int. J. Comput. Appl., vol. 156, no. 5, pp. 1–5, 2016.

[4] R. Mahdavi-Hezave and R. Ramsin, “FDMD: Feature-Driven Methodology Development,” Proc. 10th Int. Conf. Eval. Nov. Approaches to Softw. Eng., pp. 229–237, 2015.

[5] A. Firdaus, I. Ghani, and S. R. Jeong, “Secure Feature Driven Development (SFDD) Model for Secure Software Development,” Procedia - Soc. Behav. Sci., vol. 129, pp. 546–553, 2014.

[6] F. Siddiqui and Alam, M. Afshar, “Ontology Based Feature Driven Development Life Cycle,” Int. J. Comput. Sci. Issues, vol. 9, no. 1, 2010.

[7] F. Anwer, S. Aftab, S. S. M. Shah, and U. Waheed, “Comparative Analysis of Two Popular Agile Process Models: Extreme Programming and Scrum,” Int. J. Comput. Sci. Telecommun., vol. 8, no. 2, 2017.

[8] G. Rasool, S. Aftab, S. Hussain, and D. Streitferdt, “eXRUP: A Hybrid Software Development Model for Small to Medium Scale Projects,” J. Softw. Eng. Appl., vol. 6, no. 9, pp. 446–457, 2013.

[9] P. Abrahamsson, O. Salo, J. Ronkainen, and J. Warsta, “Agile Software Development Methods: Review and Analysis,” 2017.

[10] S. R. Palmer and M. Felsing, A Practical Guide to Feature Driven Development. 2002.

[11] D. Ph, “Major Seminar On Feature Driven Development Agile Techniques for Project Management Software Engineering By Sadhna Goyal Guide : Jennifer Schiller Chair of Applied Software Engineering,” p. 4, 2007.

[12] Z. Nawaz, S. Aftab, and F. Anwer, “Simplified FDD Process Model,” Int. J. Mod. Educ. Comput. Sci., vol. 9, no. 9, pp. 53–59, 2017.

[13] F. Anwer, S. Aftab, and I. Ali, “Proposal of Tailored Extreme Programming Model for Small Projects,” Int. J. Comput. Appl., vol. 171, no. 7, pp. 23–27, 2017.

[14] N. Swamy, L. M. Rao, and K. S. Praveen, “Component Based Software Architecture Refinement and Refactoring Method into Extreme Programming,” vol. 5, no. 12, pp. 398–401, 2016.

[15] M. R. Jameel Qureshi and J. S. Ikram, “Proposal of Enhanced Extreme Programming Model,” Int. J. Inf. Eng. Electron. Bus., vol. 7, no. 1, pp. 37–42, 2015.

[16] J. Choudhari and U. Suman, “Extended iterative maintenance life cycle using eXtreme programming,” ACM SIGSOFT Softw. Eng. Notes, vol. 39, no. 1, pp. 1–12, 2014.

[17] S. Ashraf and S. Aftab, “Pragmatic Evaluation of IScrum & Scrum,” Int. J. Mod. Educ. Comput. Sci., vol. 10, no. 1, pp. 24–35, 2018.

[18] N. Iqbal, M. ul Hassan, A. Rehman Osman, and M. Ahmad, “A framework for partial implementation of PSP in Extreme programming,” Int. J. Eng. Res. Appl. www.ijera.com, vol. 3, no. 2, pp. 604–607, 2013.

[19] M. R. J. Qureshi, “Estimation of the New Agile XP Process Model for Medium-Scale Projects Using Industrial Case Studies,” Int. J. Mach. Learn. Comput., vol. 3, no. 5, pp. 393–395, 2013.

[20] N. E. Fenton, and S. L. Pfleeger, "Software Metrics: A Rigorous and Practical Approach: Brooks," 1998.

[21] S. H. Kan, Metrics and models in software quality engineering. Addison-Wesley Longman Publishing Co., Inc. 2002.

[22] C. Jones, "Applied Software Measurement", McGraw Hill, 1991.

[23] N. Fenton and J. Bieman, “Software Metrics: Roadmap,” It Prof., vol. 2, pp. 38–42, 2014.

[24] S. Ashraf and S. Aftab, “Scrum with the Spices of Agile Family: A Systematic Mapping,” Int. J. Mod. Educ. Comput. Sci., vol. 9, no. 11, pp. 58–72, 2017.

[25] S. Ashraf and S. Aftab, “Latest Transformations in Scrum: A State of the Art Review,” Int. J. Mod. Educ. Comput. Sci., vol. 9, no. 7, pp. 12–22, 2017.

[26] S. Ashraf and S. Aftab, “IScrum: An Improved Scrum Process Model,” Int. J. Mod. Educ. Comput. Sci., vol. 9, no. 8, pp. 16–24, 2017.

[27] S. U. Nisa and M. R. J. Qureshi, “Empirical Estimation of Hybrid Model: A Controlled Case Study,” Int. J. Inf. Technol. Comput. Sci., vol. 1, no. July, p. 8, 2012.

[28] M. Qureshi, “Empirical Evaluation of the Proposed eXSCRUM Model: Results of a Case Study,” Int. J. Comput. Sci. Issues., vol. 8, no. 3, pp. 150–157, 2012.

[29] F. Anwer, S. Aftab, U. Waheed, and S. S. Muhammad, “ Agile Software Development Models TDD, FDD, DSDM, and Crystal Methods : A Survey,” Int. J. Multidiscip. Sci. Eng., vol. 8, no. April, pp. 1–10, 2017.

[30] S. Alshehri and L. Benedicenti, “Prioritizing CRC cards as a simple design tool in extreme programming,” Can. Conf. Electr. Comput. Eng., pp. 13–16, 2013.

[31] M. R. J. Qureshi, “Agile software development methodology for medium and large projects,” IET Softw., vol. 6, no. 4, p. 358, 2012.

[32] F. Anwer and S. Aftab, “Latest Customizations of XP: A Systematic Literature Review,” Int. J. Mod. Educ. Comput. Sci., vol. 9, no. 12, pp. 26–37, 2017.

[33] S. Aftab, Z. Nawaz, M. Anwar, F. Anwer, M. S. Bashir, and M. Ahmad, “Comparative Analysis of FDD and SFDD,” Int. J. Comput. Sci. Netw. Secur (IJCSNS)., vol. 18, no. 1, pp. 63–70, 2018.

[34] E. Mnkandla and B. Dwolatzky, “A Survey of Agile Development Methodologies,” no. December 2004, pp. 209–227, 2007.

[35] I. Journal et al., “Extreme Programming : Newly Acclaimed Agile System,” vol. 3, no. 2, pp. 699–705, 2010.

[36] R. Crocker, “The 5 reasons XP can’t scale and what to do about them,” Proc. 2nd Int’l. Conf. Extrem. Program. Agil. Process. Softw. Eng., pp. 62–65, 2001.

[37] A. Dalalah, “Extreme Programming: Strengths and Weaknesses,” î Comput. Technol. Appl., vol. 5, no. 1, 2014.

[38] S. Beecham, H. Sharp, N. Baddoo, T. Hall, and H. Robinson, “Does the XP environment meet the motivational needs of the software developer? An empirical study,” Proc. - Agil. 2007, pp. 37–48, 2007.

[39] F. Anwer, S. Aftab, M. S. Bashir, Z. Nawaz, M. Anwar, and M. Ahmad, “Empirical Comparison of XP & SXP,” Int. J. Comput. Sci. Netw. Secur (IJCSNS)., vol. 18, no. 3, pp. 161–167, 2018.

[40] F. Anwer and S. Aftab, “SXP: Simplified Extreme Programing Process Model,” Int. J. Mod. Educ. Comput. Sci., vol. 9, no. 6, pp. 25–31, 2017.

Information System Evaluation based on Multi-Criteria Decision Making: A Comparison of Two Sectors

Ansar DAGHOURI1, Khalifa MANSOURI1, Mohammed QBADOU1
1Laboratory: Signals, Distributed Systems and Artificial Intelligence (SSDIA)
ENSET of Mohammedia, University Hassan II of Casablanca

***Abstract*—In this article, our purpose is to introduce the results of a new approach to assess the information system success. It is based on the DeLone and McLean model and was applied on two domains. The chosen domains are banking sector being the most customer of information technology and construction industry as the least computer-intensive sector. The work methodology used to evaluate the information system performance is a combined approach of the two most popular multi-criteria decision making techniques: AHP and TOPSIS. Based on the results of this technique applied on studied sectors, we can obtain a horizontal comparison at the sector level and optimize the choice of the best system.**

Keywords*—Information system success; multi criteria decision; AHP and TOPSIS methods; criteria

I. Introduction

The emergence of computing and communication technology commonly known as information technology and the relatively cheap of hardware, the managers of organizations invest massively on information system (IS).

IS [1] is a set of data, hardware, software to treat data and procedure to help personnel establish the several objectives of organization.

The literature is rich with different models to assess information system [2] that offer the possibility to highlight the importance of evaluation in this field.

Concurrence, economic evolution and technological progress impose organizations to invest heavily on information system regardless their domain.

In this work, we will take two paradoxes sectors on terms of information system use. Banking sector being the most customer of IS, which has a primordial role at the internal and external environment. The role of banking information system is more complex considering the nature of products and services that offer also the requirements of these customers.

In contrary, construction industry sector who do not generally appear to appreciate the positive influence and changes that an IS provide. In general, construction sector cannot headily be compared to the banking sector in regards to adoption of information systems. Some major reasons for such situation are [3]: the nature of construction industry, the traditionalism of construction and low level of investments into research and development within this sector.

This paper present the results of a new technique developed to assess information system success being tested on two sectors. The proposed technique has adopted the D&M success model (2003) [4] and used two famous multi criteria decision making approaches; AHP and TOPSIS [5], [6] to evaluate the IS success. As a contribution of this work, this article present a benchmark study of the two studies sectors on information system success based on the same work methodology and comparing the results.

II. Information System Success Models

In the last two decades, many models have been cited to evaluate the IS and to explain the dimensions that makes IS successful [4], [7]-[12]. The majority of these models have been validated empirically by researchers in several domains.

A. The DeLone and McLean Model

The evaluation of IS success is among the most delicate areas, because of the complexity and the multidimensional aspect of IS. Among the first attempt at solving this problem that of D&M in 1992 [8]. However, this first model has undergone several criticisms which pushed the authors to update their model [4].

The original (Fig. 1) and updated (Fig. 2) DeLone and McLean IS success models are the most cited and used models in the field of IS evaluation [13].

Based on empirical studies, suggestions and tests, DeLone and McLean have updated the first model. That is way this updated model is the most adopted and cited version in the literature review in the field on IS assessment.

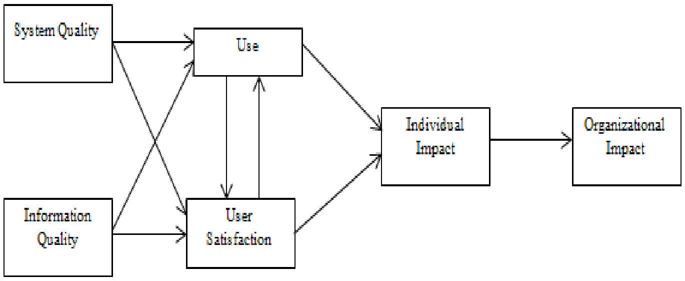


Fig. 1. Original DeLone and McLean IS success model.

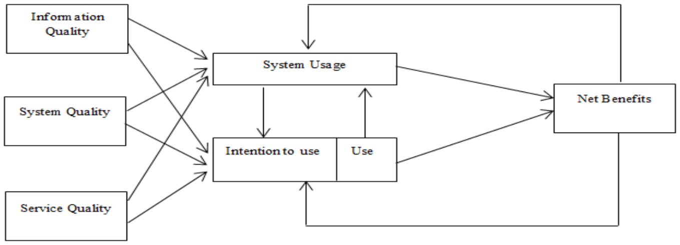


Fig. 2. Updated DeLone and McLean IS success model.

For the reasons mentioned before, in this work, our research methodology is based on the updated IS success model which is composed by six interrelated dimensions as shown in the figure at the top.

III. MCDM

Multiple criteria decision-making (MCDM), is a part of operations research [15], its objective is to facilitate decision facing problems involving multiple criteria. The complexity of such as problem is the lack of a single solution, that way it is necessary to use maker's preferences to obtain the best solution from a collection of alternatives under a number of criteria and even sub-criteria [16].

Several MCDM methods such as: [14], [17]: AHP, ANP, Electre, GP, MAUT, MAVT, TOPSIS, WSM… are employed for different applications and domains due to the particularity of each one.

A. Analytical Hierarchy Process (AHP)

AHP is an analytical technique based on a hierarchy process [18], [19]. The principal is to decompose a problem into hierarchies of goals, criteria (sub-criteria) and alternatives (Fig. 3).

AHP is classified as one of the the most cited approach. It has the possibility to treat both tangibles and non-tangibles criteria.

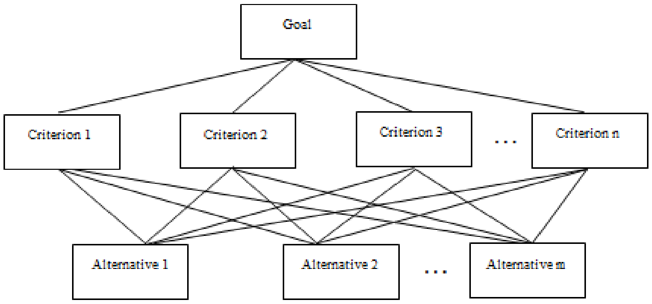


Fig. 3. Analytical hierarchy process tree.

In mathematical way, the basic AHP equations are as follow [19]:

Stage 1: Decomposes the initial problem into hierarchical presentation (goal, criteria and alternatives)

Stage 2: Develop the pairwise comparison: a_{ij}=1 when i=j and a_{ji}=1/a_{ij}

$$A_{nn}= \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{bmatrix}$$

Stage 3: Construct normalized decision matrix

$$c_{ij}=a_{ij}/\sum_{j=1}^{n} a_{ij} \qquad (1)$$

i=1, 2, 3…n and j=1, 2, 3… n

Stage 4: Construct the weighted normalized decision matrix

$$w_i=\sum_{j=1}^{n} c_{ij} \ / \ n, i=1, 2, 3...n \qquad (2)$$

$$W= \begin{bmatrix} w_1 \\ w_2 \\ \cdots \\ w_n \end{bmatrix} \qquad (3)$$

Stage 5: Calculate Eigenvector and Row matrix

$$E=N^{th}\text{rootvalue}/\sum N^{th}\text{rootvalue} \qquad (4)$$

$$\text{Rowmatrix}=\sum_{j=1}^{n} a_{ij} * e_{j1} \qquad (5)$$

Stage 6: Calculate the maximum Eigenvalue, λ_{max}

$$\lambda_{max}=\text{Rowmatrix/E} \qquad (6)$$

Stage 7: Calculate the consistency index and consistency ratio

$$\text{CI}= (\lambda_{max}\text{-n}) / (\text{n-1}) \qquad (7)$$

$$\text{CR}= \text{CI/RI} \qquad (8)$$

B. Technique for Order Preference by Similarity to Ideal Solution (TOPSIS)

The TOPSIS method was developed in 1981 [20], the basic concept is to choose the best alternative (Fig. 4) depending on closest and most distance respectively to positive ideal solution and negative ideal solution.

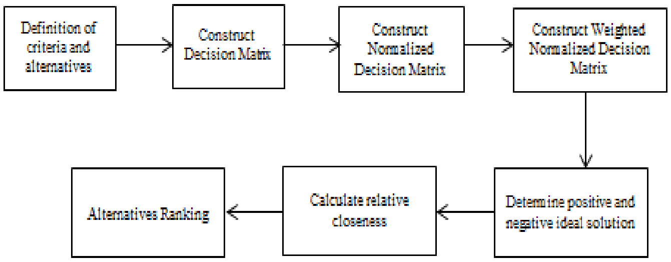

Definition of criteria and alternatives
Construct Decision Matrix
Construct Normalized Decision Matrix
Construct Weighted Normalized Decision Matrix
Determine positive and negative ideal solution
Calculate relative closeness
Alternatives Ranking

Fig. 4. TOPSIS methodology.

The TOPSIS has the following stages [20]:

Stage 1: Construct the decision matrix:

$$A_{mn=\{a_{ij}/i\epsilon(1,2,\dots,m) and\ j\epsilon(1,2,\dots,n)\ \ \}} \qquad (9)$$

Stage 2: Construct normalized decision matrix:

$$r_{ij}=x_{ij}/\sqrt{\sum_{j=1}^{J} x_{ij}^2}\ , j=1,\ 2,\ 3...\ J \text{ and } i=1,\ 2,\ 3...n \qquad (10)$$

Stage 3: Construct the weighted normalized decision matrix:

$$v_{ij}=w_i * r_{ij}, j=1, 2, 3... J \text{ and } i=1, 2, 3 \ldots n \tag{11}$$

Stage 4: Determine the positive ideal (PIS) and negative ideal solution (NIS)

$$A^{*}= \{v_1^{*}, v_2^{*}, \ldots, v_n^{*}\} \text{ maximum values} \tag{12}$$

Where $v_i^{*}= \{\max (v_{ij})\ if\ j\epsilon J; \min\left(v_{ij}\right) if j\ \epsilon J^{-}\}$

$$A^{-}= \{v_1^{-}, v_2^{-}, \ldots, v_n^{-}\} \text{ minimum values} \tag{13}$$

Where $v^{-}= \{\min (v_{ij})\ if\ j\epsilon J; \max\left(v_{ij}\right) if j\ \epsilon J^{-}\}$

Stage 5: Calculate the separation measures of each alternative from PIS and NIS

$$d_i^{*}=\sqrt{\sum_{j=1}^{n}(v_{ij}-v_j^{*})^2}, j=1, 2, 3... J \tag{14}$$

$$d_i^{-}=\sqrt{\sum_{j=1}^{n}(v_{ij}-v_j^{-})^2}, j=1, 2, 3... J \tag{15}$$

Stage 6: Calculate the relative closeness coefficient to the ideal solution:

$$CC_i = \frac{d_i^{-}}{d_i^{*}+d_i^{-}} \quad i=1, 2, 3... J \tag{16}$$

Stage 7: Rank the preference order.

IV. Work Methodology

A. *Purpose of Study*

This study provide a comparison between IS success in two sectors and presents the results of a framework based on the MCDM approaches for ranking these information systems. Our work methodology was applied on two sectors: banking sector [21] and construction industry sector [22]. Firstly, D&M model (2003) was adopted to construct the analytical hierarchy process; the six dimensions are considered as main criteria and sub-criteria were taken from literature. Using AHP method the weights of criteria (sub-criteria) are obtained. Then, TOPSIS method was applied to rank information systems. For the purpose of testing and verifying the proposed framework on each sector, we were done an online questionnaire to collect data.

B. *Studied Sectors*

As mentioned, this study compare the results of a framework applied on two sectors, in the following sections we will explain the particularities of each sector.

1) Banking Sector: Banks are the first companies that have invested heavily in the technologies of information and communication. The banking sector is one of the sectors most risk-sensitive [23] with this multitude of risk types, the IS is considered both as a risk factor and a tool for identifying, evaluating and monitoring risks.

Banking IS is generally subject to stricter rules [24]: The availability of the IS even after malfunctions, the security of data especially the confidentiality of customer information, archiving and data traceability to facilitate internal control and audit and finally the integrity of data.

A last particularity of the banking IS is the opening to the outside and that it is to be functional 24 hours a day, banks offer today sales and purchase transactions, consultation or even transaction from Internet.

2) Construction Industry Sector: The construction industry sector use information systems [25] to enhance organizational flexibility, improve decision making capability, reduce project completion time and cost and to present an ideal schedule for the factory construction work. In general way, it used to achieve project mission objectives within specified constraints that construction industry knows. Despite, the construction industry is one of the sector that doesn't use heavily IS [26].

The following figure presents the basic idea of the study (Fig. 5):

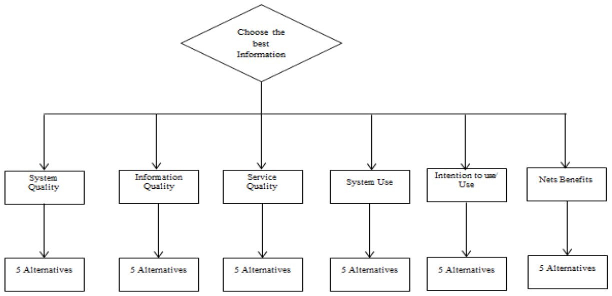


Fig. 5. General work research.

C. *Research Methodology*

The evaluation procedure consists of six main steps as summarized in the following figure (Fig. 6):

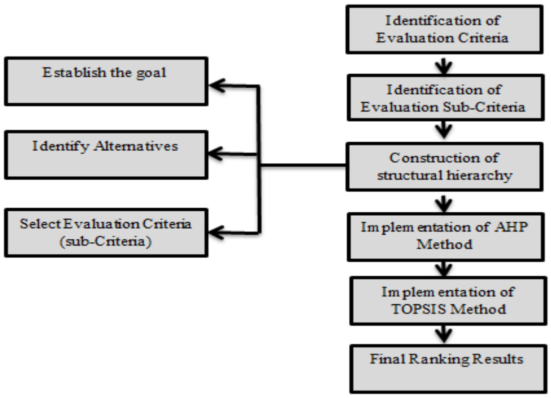

Identification of Evaluation Criteria
Identification of Evaluation Sub-Criteria
Construction of structural hierarchy
Implementation of AHP Method
Implementation of TOPSIS Method
Final Ranking Results
Establish the goal
Identify Alternatives
Select Evaluation Criteria (sub-Criteria)

Fig. 6. Steps of evaluation procedure.

Step 1: Identify the evaluation criteria using the information system success model (DeLone & McLean 2003);

Step 2: Identify sub-criteria;

Step 3: Construct the structural hierarchy; establish goal which is in our case the evaluation of IS success, identify the alternatives which are the methods that change the preliminary condition into preferred condition and select the main criteria (sub-criteria);

Step 4: Calculate the weights of each criterion using AHP;

Step 5: Apply the TOPSIS method;

Step 6: Achieve the final ranking results.

The description of each step will be given in the following sections.

V. Results and Discussions

A. Implementation of AHP Method

The AHP hierarchy for decision making in this paper as shown in Table I is constitute of six main criteria which are the several dimensions of the updated model (D&M 2003) and sub-criteria were inspired from literature.

This study utilized a questionnaire survey to collect data from the different decision makers. According to the steps defined in Section III-A, (Fig. 7) display the pairwise comparison matrix using (1) and (Fig. 8) shows the normalized decision matrix which is calculated using (2). In this part, we choose to only present results of banking sector.

$$C = \begin{bmatrix} 1 & 5 & 7 & 5 & 7 & 3 \\ 0.2 & 1 & 5 & 3 & 3 & 3 \\ 0.14 & 0.2 & 1 & 7 & 3 & 3 \\ 0.2 & 0.33 & 0.14 & 1 & 3 & 5 \\ 0.14 & 0.33 & 0.33 & 0.33 & 1 & 3 \\ 0.33 & 0.33 & 0.33 & 0.2 & 0.33 & 1 \end{bmatrix}$$

Fig. 7. Aggregated pairwise comparison matrix.

$$C = \begin{bmatrix} 0.49 & 0.69 & 0.50 & 0.30 & 0.40 & 0.16 \\ 0.09 & 0.13 & 0.36 & 0.18 & 0.17 & 0.16 \\ 0.69 & 0.02 & 0.07 & 0.42 & 0.17 & 0.16 \\ 0.09 & 0.04 & 0.01 & 0.06 & 0.17 & 0.27 \\ 0.06 & 0.04 & 0.02 & 0.01 & 0.05 & 0.16 \\ 0.16 & 0.04 & 0.02 & 0.01 & 0.01 & 0.05 \end{bmatrix}$$

Fig. 8. Normalized decision matrix.

Then the priority weights are calculated using (2):

w_1=2.32/6= 0.42 w_2=1.04/6=0.11

w_3=1.36/6= 0.18 w_4=1.02/6=0.06

w_5=0.65/6= 0.15 w_6=0.38/6=0.05

The same calculation steps were followed to determine the weights of the sub-criteria. We chose not to present them so as not to load our article by several tables. These results were used for the TOPSIS method.

TABLE I. Hierarchical Presentation of Criteria

Criteria	**Sub-Criteria**
System Quality (C_1)	Availability(C_{11}),EmployeesOccupancy(C_{12}), Longest Delay(C_{13}),Answer Speed (C_{14}), Abandons(C_{15}), Blockage(C_{16}), Average hour of operation(C_{17}), Self-service and availability (C_{18})
Information Quality (C_2)	Grammar and spelling(email) (C_{21}), Data accuracy(C_{22}),Secure(C_{23}),Complete(C_{24}), Relevant and correct(C_{25}) and Data Understandability(C_{26})
Service Quality (C_3)	On Time delivery(C_{31}),Knowledge and competency(C_{32}), Error Network(C_{33}), Availability(C_{34}), Access(C_{35}), Rate Delay(C_{36}) and Reliability (C_{37})
Use (C_4)	Frequency of use(C_{41}), Amount of use(C_{42}), Number of reports generated(C_{43}), Technical support(C_{44}), Managerial support(C_{45}) and Financial transactions use (C_{46})
User Satisfaction (C_5)	Handle Time(C_{51}), Average Number of employees connected(C_{52}), Training Investment(C_{53}), Employee Turnover(C_{54}) and Average Satisfaction(C_{55})
Net Benefits (C_6)	Return on Investment(C_{61}), Productivity(C_{63}), Profit(C_{63}), Market Share(C_{64}), Growth in customer base(C_{65}) and Increased Sale (C_{66})

To calculateλ_{max}, we used (4), (5), and (6): λ_{max} =6.09

(CI) and (CR) are calculated through (7) and (8), respectively (for RI=1.24, according to table below (Table II)):

CI=0.018 and CR= 0.014

CR=0.014<0.10, it is accepted.

TABLE II. random Index (RI) [18]

N	1	2	3	4	5	6	7	8	9
RI	0	0	0.58	0.90	1.12	1.24	1.32	1.41	1.45

B. Implementation of TOPSIS Method

The second MCDM methods we used in this work is TOPSIS. AHP weighted scores are used by TOPSIS for each sector's alternative to obtained performance ranks of systems.

The five banking companies are referenced: B_1 B_2 ,B_3,B_4, B_5 and the companies operating in the construction industry sector are referenced as:C_1, C_2 ,C_3, C_4,C_5.

The different steps of the TOPSIS method were followed applying equations. In the final step, using (16), the relative closeness to ideal solution is calculated and the companies were ranked as shown in Tables III and IV.

TABLE III. ranking of banking companies

Alternatives	d_i^*	d_i^-	CC_i	Result-Ranks
B_1	0,196	0,116	0,372	4
B_2	0,203	0,164	0,446	2
B_3	0,209	0,094	0,310	5
B_4	0,166	0,204	0,551	1
B_5	0,187	0,130	0,409	3

TABLE IV. RANKING OF COMPANIES ON CONSTRUCTION INDUSTRY

Alternatives	d_i^*	d_i^-	CC_i	Result-Ranks
C_1	0,461	0,221	0,323	4
C_2	0,593	0,145	0,196	5
C_3	0,172	0,580	0,771	1
C_4	0,381	0,283	0,425	3
C_5	0,356	0,316	0,470	2

According to the value of$\mathrm{CC_i}$, the alternative having highest closeness coefficient in bank is alternative 4 with $\mathrm{CC_i}$ =0,551 and in construction industry is alternative 3 with$\mathrm{CC_i}$ =0,771. These alternatives are selected as the two best companies among studied alternatives in terms of information system performance.

C. Elaboration of a Prototype

This work is based on a proposed framework to evaluate the IS success, to implement it we propose a prototype that consists of three interconnected parts (Fig. 9)

- The different actors which in our case are the decision makers to specify the weights of each criterion (sub-criterion) to evaluate the IS.
- The user interface that allows adding details of criterion to assess the studied system.
- Functional part that simulates the different steps of the two used approaches AHP and TOPSIS.
- Data Base to store the various data that will be used in the next steps of evaluation.

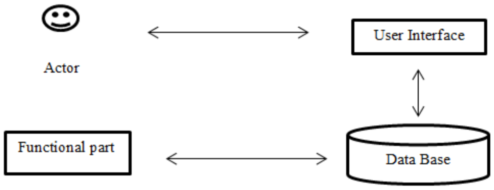


Fig. 9. Prototype architecture.

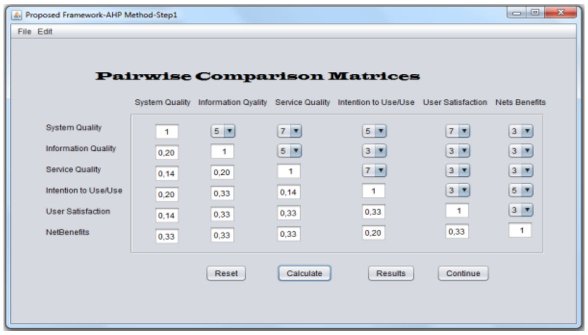
1) User Interface: After authentification throught a login and a password, the decision marker goes on to enter the data of the evaluation.

2) Functional part: The role of this part is to rank the different alternatives using AHP and TOPSIS methods according to the criteria chosen from the DeLone and McLean model (2003).The two used MCDM in this work requires a very important number of calculations specially when the number of criteria, sub-criteria and alternatives increases; for this reason, we present a prototype of software that implement our methodology process, it is developed in java language under Netbeans platform. The main interfaces will be presented in the following paragraphs.

The first step after authentication, the analyst fills the pairwise comparison matrices which represent several criteria. (Fig. 10 and 11) shows interfaces that lead the analyst to input the values of matrices respecting Saaty's pairwise comparison scale of AHP method. (Fig. 12) provides the weights of main and sub criteria using AHP equations.

The second step of the proposed prototype aims to implement the TOPSIS method, the first interface of this phase, lead the user to input the values of pairwise comparison of matrices which contain criteria and sub-criteria. (Fig. 13 and 14) shows the application of (10) (11) (12) (13) (14) and (15) based on the data of alternatives input in by the user to make calculations.

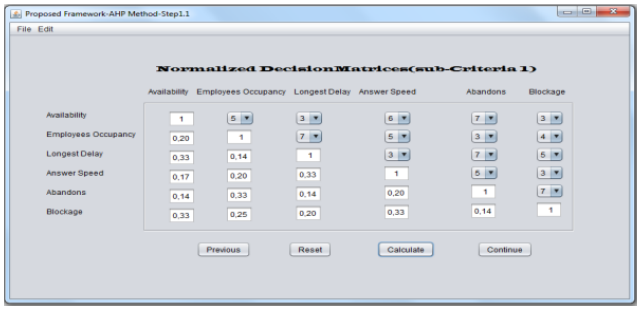
Finally, we display the results of alternatives ranking based on the calculations of the distance between positive and negative ideal solutions. Fig. 15 shows the list of the alternatives selected by the maker decision which can be printed.



Pairwise Comparison Matrices

	System Quality	Information Qyality	Service Quality	Intention to Use/Use	User Satisfaction	Nets Benefits
System Quality	1	5	7	5	7	3
Information Quality	0,20	1	5	3	3	3
Service Quality	0,14	0,20	1	7	3	3
Intention to Use/Use	0,20	0,33	0,14	1	3	5
User Satisfaction	0,14	0,33	0,33	0,33	1	3
NetBenefits	0,33	0,33	0,33	0,20	0,33	1

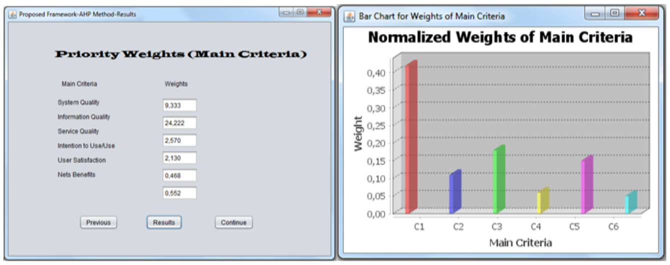
Fig. 10. Pairwise comparison matrix (main criteria).



Normalized DecisionMatrices(sub-Criteria 1)

	Availability	Employees Occupancy	Longest Delay	Answer Speed	Abandons	Blockage
Availability	1	5	3	6	7	3
Employees Occupancy	0,20	1	7	5	3	4
Longest Delay	0,33	0,14	1	3	7	5
Answer Speed	0,17	0,20	0,33	1	5	3
Abandons	0,14	0,33	0,14	0,20	1	7
Blockage	0,33	0,25	0,20	0,33	0,14	1

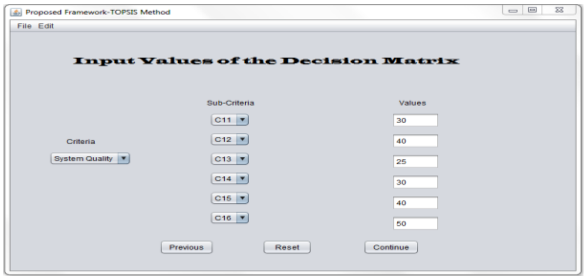
Fig. 11. Pairwise comparison matrix (sub-criteria).


Proposed Framework-AHP Method-Results
Previous
Results
Continue
Bar Chart for Weights of Main Criteria
Normalized Weights of Main Criteria
Weight
0,40
0,35
0,30
0,25
0,20
0,15
0,10
0,05
0,00
C1
C2
C3
C4
C5
C6
Main Criteria

Priority Weights (Main Criteria)

Main Criteria	Weights
System Quality	9,333
Information Quality	24,222
Service Quality	2,570
Intention to Use/Use	2,130
User Satisfaction	0,468
Nets Benefits	0,552

Fig. 12. Weights of main criteria.

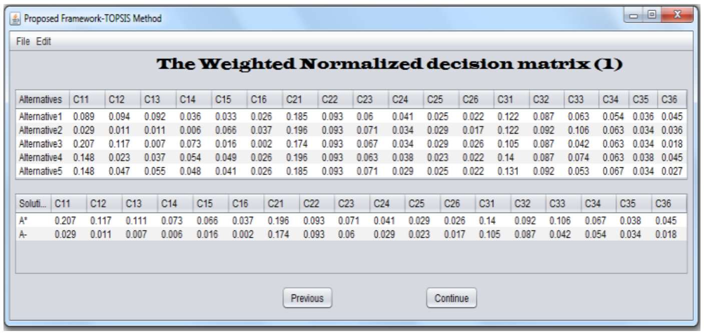


Input Values of the Decision Matrix

Criteria: System Quality

Sub-Criteria	Values
C11	30
C12	40
C13	25
C14	30
C15	40
C16	50

Fig. 13. First step of TOPSIS method.

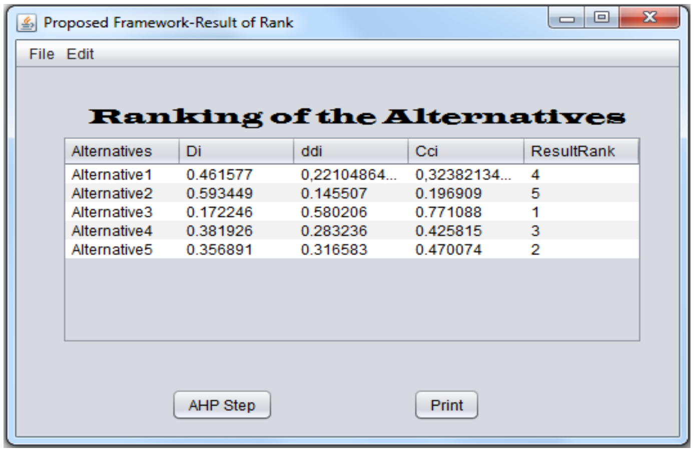


The Weighted Normalized decision matrix (1)

Alternatives	C11	C12	C13	C14	C15	C16	C21	C22	C23	C24	C25	C26	C31	C32	C33	C34	C35	C36
Alternative1	0.089	0.094	0.092	0.036	0.033	0.026	0.185	0.093	0.06	0.041	0.025	0.022	0.122	0.087	0.063	0.054	0.036	0.045
Alternative2	0.029	0.011	0.011	0.006	0.066	0.037	0.196	0.093	0.071	0.034	0.029	0.017	0.122	0.092	0.106	0.063	0.034	0.036
Alternative3	0.207	0.117	0.007	0.073	0.016	0.002	0.174	0.093	0.067	0.034	0.029	0.026	0.105	0.087	0.042	0.063	0.034	0.018
Alternative4	0.148	0.023	0.037	0.054	0.049	0.026	0.196	0.093	0.063	0.038	0.023	0.022	0.14	0.087	0.074	0.063	0.038	0.045
Alternative5	0.148	0.047	0.055	0.048	0.041	0.026	0.185	0.093	0.071	0.029	0.025	0.022	0.131	0.092	0.053	0.067	0.034	0.027

Soluti...	C11	C12	C13	C14	C15	C16	C21	C22	C23	C24	C25	C26	C31	C32	C33	C34	C35	C36
A*	0.207	0.117	0.111	0.073	0.066	0.037	0.196	0.093	0.071	0.041	0.029	0.026	0.14	0.092	0.106	0.067	0.038	0.045
A-	0.029	0.011	0.007	0.006	0.016	0.002	0.174	0.093	0.06	0.029	0.023	0.017	0.105	0.087	0.042	0.054	0.034	0.018

Fig. 14. Weighted normalized decision matrix (positive and negative solutions).



Ranking of the Alternatives

Alternatives	Di	ddi	Cci	ResultRank
Alternative1	0.461577	0,22104864...	0,32382134...	4
Alternative2	0.593449	0.145507	0.196909	5
Alternative3	0.172246	0.580206	0.771088	1
Alternative4	0.381926	0.283236	0.425815	3
Alternative5	0.356891	0.316583	0.470074	2

Fig. 15. Ranking of alternatives.

3) Data Base: to store data's evaluation. We chose to work with SQL Lite because the recorded information is not large. We used a SQLite database given the reduced size of the data.

D. Discussions

In the context of the application of MCDM methods, main criteria are chosen from The D&M model and sub-criteria from the literature to applied our research methodology. From the results, it can be observed that alternatives B_4 and C_3 consistently perform better than the rest. It should be noted that the weights of criteria differ between banking sector and construction industry sector.

Banks characterized by the nature of their operations and services that explain the high value of weights of system quality and information quality (Fig. 12) At the opposite of construction industry sector who rarely uses an information system which is clear by the weight of criteria that are very close. (Fig. 16) shows the separation measures of each alternative (5 alternatives of each sector) from positive ideal solution, it's clear that values of bank's alternatives are close this can be explained by the large use of IS within banks. In Construction industry, we observe a big difference between the values there are companies that use the IS and others that do not use it, in this sector the IS is not used in an equitable way. The same remark can be done concerning the separation measures of each alternative from negative ideal solution (Fig. 17).

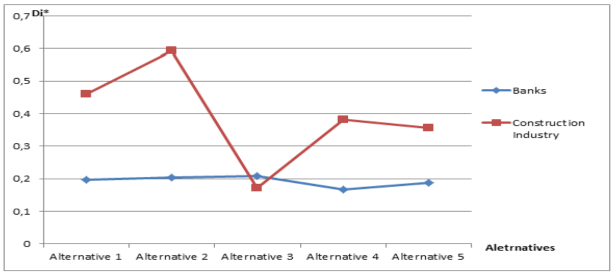


Fig. 16. Separation measures from positive ideal solution.

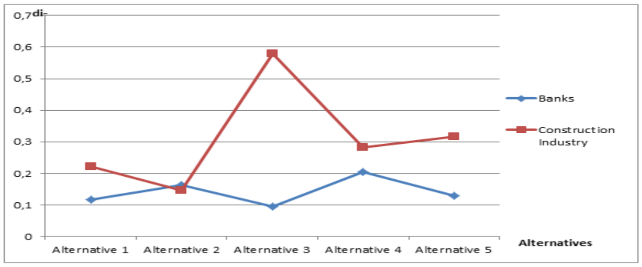


Fig. 17. Separation measures from negative ideal solution.

The relative CC_i of each alternative confirm the previous remarks (Fig. 18), bank's alternatives have close values due to the high use of information system within banks. Construction industry companies are very distant in terms of closeness coefficient values; we have company C_3 that use heavily IS and the others who use the IS partially.

We conclude that the proposed methodology and the hierarchical presentation of criteria can be used for IS evaluation no matter which sector is studied. The method evaluate IS independently of input values of the two used MCDM.

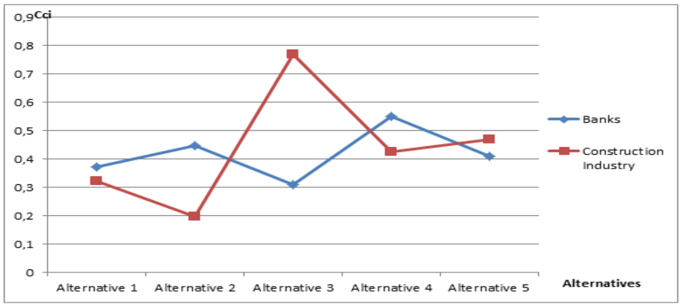


Fig. 18. Closness coefficient.

VI. Conclusion

Determining the success of information system and evaluating them is crucial for the sector's development. In this study, we were presented an application of our work methodology based on a hybrid MCDM process (AHP and TOPSIS methods). This methodology is tested on two sectors; banking and construction industry.

Five companies are chosen from each sector to select the most performing IS. Furthermore, a prototype of software implementing the work methodology is proposed. AHP method is used to determine the weights of main criteria and sub-criteria and TOPSIS method is utilized for ranking alternatives.

The software leads to select the best alternative among the others and simplify the calculations steps. When evaluating the ranking results, alternative B_4 with CC_i=0,551 and alternative C_3 with CC_i =0,771 are the best alternative respectively in construction industry sector and banking sector.

The proposed method enables decisions analysts to better evaluate their information system and provides more effective, complete and systematic decision support tool.

Our future work will treat more companies in both sectors to have a base of references, which will allow us to know the criteria to optimize according to sector's nature.

References

[1] O Nicolescu and I Verboncu, "Fundamentals of Organizational Management," Economics Tribune Publishing, Bucharest, 2006.

[2] W H DeLone and E R McLean, "Information Systems Success Measurement," Foundations and Trends in Information Systems, vol. 2, no. 1, pp. 1-116, 2016.

[3] S K Lee, J H Yu, "Success model of project management information system in construction," Automation in Construction, vol. 25, pp. 82-93, 2012.

[4] W H DeLone and E R McLean, "The Delone and McLean Model of Information Systems Success: a Ten-Year Update," Journal of Management Information Systems, vol. 19, no. 4, pp. 9-30, 2003.

[5] H R Weistroffer, C H Smith, and S C Narula, "Multiple criteria decision support software," Multiple Criteria Decision Analysis: State of the Art Surveys Series, Springer: New York, 2005.

[6] P McGinley, "Decision analysis software survey," 2012.

[7] F D Davis, "User acceptance of computer technology: a comparison of two theoretical models," Management science, pp. 982–1003, 1989.

[8] W H DeLone and E R McLean, "Information Systems Success: The Quest for the Dependent Variable," Information Systems Research, vol. 3, no. 1, pp. 60-95, 1992.

[9] G G Gable, D Sedera, and T Chan, "Enterprise systems success: a measurement model," in Proceedings of the 24th International Conference on Information Systems Association for Information Systems, Seattle, Washington, 2003, pp. 576–591.

[10] P Ifinedo and N Nahar, "Quality, Impact and Success of ERP Systems: A Study Involving Some Firms in the Nordic-Baltic Region," Journal of Information Technology, pp. 19–46, 2006.

[11] D Sedera and G G Gable, "Knowledge Management Competence for Enterprise System Success," The Journal of Strategic Information Systems, pp. 296–306, 2010.

[12] S Shang and B P Seddon, "Assessing and managing the benefits of enterprise systems: the business manager's perspective," Information Systems Journal, pp. 271–299, 2002.

[13] S A Kronbichler, H Ostermann, and R Staudinger, "A comparison of ERP-Success measurement approaches," Journal of Information Systems and technologie Management, pp. 281-310, 2010.

[14] D Latinopoulos, K Kechagia "A GIS-based multi-criteria evaluation for wind farm site selection. A regional scale application in Greece," Renewable Energy, 550-560, 2015.

[15] E K Zavadskas, Z Turskis, and S Kildiene, "State of art surveys of overviews on MCDM/MADM methods," Technol. Econ. Dev. Econ, pp. 165–179, 2014.

[16] M Velasquez, P T Hester, " An analysis of multi-criteria decision making methods ," International Journal of Operations Research, vol. 10, No. 2, pp. 56-66, 2012

[17] M M Kasim, H Ibrahim, and M S Bataineh "Multi-criteria decision making methods for determining computer preference index," Journal of ICT, pp. 137-148, 2010.

[18] T L Saaty, The analytic hierarchy process: New York: McGraw-Hill, 1980.

[19] T L Saaty, "Decision making with the analytic hierarchy process," Int. J. Services Sciences, pp. 83-98, 2008.

[20] C L Hwang and K Yoon, "Multiple attributes decision making methods and applications," in Berlin: Springer, 1981.

[21] A Daghouri, K Mansouri, and M Qbadou, "Assessing information system performance in banks based on multi-criteria decision making techniques," in International Conference on Communication, Management and Information Technology (ICCMIT 2018), Madrid, 2018.

[22] A Daghouri, K Mansouri, and M Qbadou, "Information System Performance Evaluation and Optimization using AHP and TOPSIS: Construction Industry case," in International Conference on Optimization and Applications, Maroc, 2018.

[23] K PILARCZYK, "Importance of Management Information System in Banking Sector," Annales Universitatis Mariae Curie-Sklodowska Lublin-Polonia, no. 2, 2016.

[24] M Georgescu and V Jeflea, "The particularity of the banking information system," in 7th International Conference on Globalization and Higher Education in Economics and Business Administration, GEBA, 2015, pp. 268-276.

[25] S Atul and S Upenda, "Management Information System in Construction Industry," INTERNATIONAL RESEARCH JOURNAL OF MULTIDISCIPLINARY STUDIES, vol. 2, March 2016.

[26] V K Bansal, "Application Areas Of Gis In Construction Projects And Future Research Directions," The International Journal of Construction Management, vol. 12, no. 4, pp. 17-36, 2012.

A Survey of Energy Aware Cloud’s Resource Allocation Techniques for Virtual Machine Consolidation

Asif Farooq
Department of Computer Science,
University of Sargodha, Lahore Campus,
Lahore, Pakistan

Tahir Iqbal, Muhammad Usman Ali,
Zunnurain Hussain
Department of Computer Science, Bahria University,
Lahore Campus, Lahore, Pakistan

***Abstract*—As the demand for cloud computing environment is increasing, new techniques for making cloud computing more environment-friendly are being proposed with an aim to convert traditional cloud computing into green cloud computing. A standout amongst the most imperative complications in cloud computing is streamlining of energy utilization because its importance is increasing rapidly. There are numerous strategies and algorithms used to limit the energy utilization in the cloud. Methods incorporate DVFS, UP-VMC, Utility based MFF, HCT, AVVMC, ACO, and ESWCT. In this survey, a review of energy-aware techniques is presented for making virtual machines more energy efficient in a cloud computing. Working on each technique is briefly explained. A comparative analysis is also given for comparing multiple efficient techniques with respect to performance metrics.**

Keywords—Cloud computing; energy aware; green cloud computing

I. Introduction

Cloud computing is defined as a large group of interconnected servers and computers with a primary objective of providing reliable services and custom-made computing environments to its users. As cloud computing grows the demand for resources required to support such a large network is also increasing. Hundreds of data servers are added daily to data centers around the world to support the growing needs of cloud computing which lead to power consumptions issues. The complete framework for cloud computing depends upon datacenters for data storage and processing. These data centers mainly consist of hundreds of thousands of servers which are unfriendly to the environment in terms of carbon emissions. According to [1] one data center can annually produce up to 170 million tons of carbon, which is estimated to increase up to 670 million tons in 2020 worldwide [2]. With a huge consumption of electricity, the operational costs are also high. The total amount of consumed energy is estimated up to 250 billion kilowatt-hours by a single datacenter [2]. Hence these factors make a datacenter unfriendly to the environment. To overcome such problems cloud’s resources must be managed in an optimal method. Allocation of resources plays an important role in maximizing the efficiency as well as reducing the power consumption of the entire system.

In Resource Allocation (RA), Available resources are assigned to needed cloud applications in such a way that it provides a level of satisfaction to its customers while ensuring the overall efficiency of the system. The RA should be beneficial to end users and economical to the service provider [13]. If resources are not managed precisely then it may starve cloud services. Thus energy efficient cloud resource allocation consists in identifying and assigning resources to each incoming user request in such a way that:

- The user requirements are completely fulfilled.
- Least possible number of resources are used.
- Data center energy efficiency is optimized.

Virtualization technology is adopted by data centers to overcome energy and resource efficiency. In virtualization process, each server’s resources is logically divided into a certain number of independent partitions. Each of these logical servers or partitions now becomes a Virtual Machine (VM), capable of running independent operating system and applications. Thousands of physicals sever can now run millions of VMs, through live VM migration techniques VMs can be transferred from server to server hence balancing the amount of VMs. This helps to switch a physical server in an idle (suspended) state if its resources are not being currently used up to a certain limit [3].

Enhancement of portability, manageability, and security are the main advantages of using VMs. Isolation can also be achieved, ensuring that VMs in a single physical server should not affect with each other. The virtualization process also blocks guest operating system of any VM from directly accessing the hardware resources of the physical server. Due to multiple overheads associated with VMs, regular monitoring is required to avoid performance degradation [4]. As multiple VMs depends on one single server, so the failure of one single server could crash hundreds of VMs. VM energy aware consolidation techniques are adapted for efficiently increasing cloud’s resource utilization. For effective utilization of data center resources two step are carried out:

1) Effective Placement of VMs in PMs.
2) Optimizing allocated resources using live migration techniques.

The primary goal for the placement of VMs is to increase the throughput of cloud's resources and to lower the power consumptions [3]. Two main types of placement algorithms are categorized:

- Power Based Placement Algorithms.
- Application QoS Based Placement Algorithms.

During live migration of a VM, the memory state is uninterruptedly transferred from one PM to another. Live VM migration is the key for improving energy efficiency and resource optimization. Live VM migration causes following problems for dynamic placement of VMs.

- To determine if a PM is over-loaded (which requires migration of VM(s) from one host to other.
- To determine when a host is under-loaded (switching a PM into suspended state by migrating all of its VMs to other PM(s).
- To determine which VM(s) must be transferred from an over-loaded PM.
- To determine new places for migrating VM(s) either from an over-loaded or under-loaded PM(s).

II. Literature Review

In [3] a dynamic VM consolidation utilization prediction approach UP-VMC is presented, both current and future resource utilization is handled for VMs consolidation. A regression-based prediction model is used for future resource utilization prediction. Authors also proposed a VM allocation algorithm to enhance the QoS while minimizing the number of migrations using prediction models. UP-VMC shows significant improvements in the reduction of energy and VM migrations. Another approach [4] based on deploying a self-managing VM placement solution that assigns VMs to PMs dynamically with respect to the utilization of resources. This proposed solution is based on utility functions with a main goal of maximizing the IaaS provider profit as well as reducing energy consumption cost. Provided results shows improvements as compared to existing heuristics based solution. In [5] authors proposed an Ant Colony Optimization (ACO) ACO metaheuristic-based server consolidation mechanism to overcome the issue of power consumptions as well as maximizing the resource utilization in large virtualized data centers. Authors compared proposed solution with existing methods and show the effectiveness of the proposed algorithm. In [6], authors proposed two algorithms (Honeybee algorithm and HCT) for VM placement. Both of these proposed algorithms are tested again many overload detection VM selection algorithms. Results are concluded based on energy consumption along with VM migration and SLA. Results conclude that the proposed HCT algorithm shows improvement in energy consumption, migration and SLA. Another approach [7] for energy-aware VM placement is proposed using a technique called Minimum Correlation Coefficient (MCC). Authors defined the VM placement problem with flexibility among bin sizes and prices and proposed a power-aware PABFD method. To adjust the trade-off between SLAV and the amount of energy consumed Fuzzy AHP method is adopted. This method delivers a satisfactory trade-off between energy consumptions and SLA violations. Two energy aware algorithms [8] are proposed (ESWCT) and (ELMWCT) for scheduling VMs using a workload-aware consolidation technique with a primary goal of reducing energy consumption in virtualized data centers. Multiples factors including imbalance utilization and resource utilization are deployed to compare against multiple scheduling algorithms. Results from both algorithms show good power savings and balanced resource utilization. In [9] an another algorithm for dynamic placement is proposed in which authors applied ant colony optimization (ACO) algorithm on multi-dimension bin packing (MDBP) algorithm. Results are compared with a traditional greedy algorithm (FDD). Both of these algorithm are implemented and validation in simulations. On average the new approach consume 4.1% less power as compared to FDD algorithm but comes a tradeoff with computational time, as ACO based algorithm took longer time for computation than FDD. In [10], author proposed another energy aware EEVS scheduling algorithm, which works by finding an optimal frequency to process VMs on a PM. Proposed approach also support DVFS and works by allocating VM to highest performance power ration PM. Results are simulated and show that EEVS algorithm can reduce up to 20% power consumption and increase 8 % of processing capacity in a best-case scenario. In [11] another nature-inspired ACO based algorithm of VM placement is proposed with a primary goal to get a non-dominated solution set that concurrently reduces the energy consumptions. The proposed ACO based VMPACS algorithm is compare with two single objective algorithm including FFD and SACO, obtained results shows that VMPACS can search solution space more efficiently to obtain a solution that use minimum number of servers and maximum resource utilization. This solution improved the overall efficiency hence less power is consumed. Two new approaches [12] are proposed related to VM placement issues, both of the proposed approaches are originated by modifying the traditional firefly algorithm with hierarchical cluster and meta-heuristics. Results related to power consumptions are deduced and compared with tradition HCT and honeybee algorithm, claiming

2% less energy consumption than original firefly algorithm and 12% less energy consumption than honeybee algorithm. Below table shows the techniques and algorithms that are selected to conduct this review paper (Table I). Each paper is selected on the basis of this paper's primary goal of energy efficiency.

TABLE I. An Overview of Energy Effective VM Consolidation Techniques

Techniques / Algorithms	**Year of Publish**	**Ref.**
Utilization Prediction aware VM Consolidation (UP-VMC)	2016	[3]
Self-managing utility functions	2016	[4]

HCMFF and MFF algorithm	2016	[12]
EEVS algorithm	2015	[10]
ACO metaheuristic-based server consolidation mechanism	2014	[5]
Honeybee Cluster Technique (HCT)	2013	[6]
Minimum Correlation Coefficient Approach	2013	[7]
ESWCT – ELMWCT algorithm	2013	[8]
VMPACS algorithm	2013	[11]
ACO-Based Approach	2011	[9]

III. Performance Metrics

The main goal of all of these approaches is to minimize the power consumption as much as possible either by reducing the amount of active PMs or VMs migrations in a manner that guarantees that SLAs are not violated. Hence the performances of listed techniques are measured on the following metrics:

A. *SLA Violations*

Service Level Agreement (SLA) is defined as a contract in-between cloud service provider and its customers. This agreement states the performance standards that the provider will obligated to meet. A number of performance and quality factors are established some of which include:

- The availability and uptime of services.
- Response time for applications.
- Notification Schedule for network changes.

For different types of customer infrastructures SLA may specify other parameters such as uninterruptable power supplies. SLA Violations (SLAV) is a performance metric independent form workload applied on any VM in Virtualized data centers. Total performance of VM is measured by obtaining SLAV in over-utilization (SLAVO) and in migration (SLAVM). Both of these evaluation metrics are independent of each other and carry equal importance in SLAV measurements.

$$\text{SLAV} = \text{SLAVO} \times \text{SLAVM}$$

B. *Energy Consumption*

Overall energy that is consumed by a physical resource during workloads is defined as energy consumption and is usually measured in kilo watts per hour (KwH). Consumption of PM usually depends upon the usages of CPU and memory during application workload. Although studies [9] have pointed that CPU often consumed more energy as compared to other PM resources. Therefore PM resources utilization is mostly represented by CPU utilization.

IV. Technical Review

In this Section, a brief overview for the proposed methodologies of each paper is presented. The working of algorithms is explained briefly to develop an insight for each paper.

A. *UP-VMC*

In [3] the proposed algorithm (UP-VMC) follows two steps procedure for VM Consolidation:

Step 1: Algorithm first start to target the VMs from overloaded and predicted overloaded PMs. Host should be consider overloaded or predicted overload as soon as any of the resources (CPU or Mem.) of PM is exceeded from a set threshold limit. Two regression based prediction models are used for resource utilization, Linear Regression and K-Nearest Neighbor Regression (K-NNR). These models predict the overloading of a PM upon the CPU and memory usage. To determine the relationship for PMs and VMs for current and future resource utilization following equations are used:

$$PUpde = \alpha + \beta Upde \qquad (1)$$

Where $PUpde$ and $Upde$ are capacity vectors for the predicted and current used PM respectively. α and β *are* regression coefficients.

$$PUv = \alpha + \beta Uv \qquad (2)$$

Similarly (2) predicts the resource utilization of VM.

Three selection policies were selected including Minimum Migration Time (MMT), Maximum Load (MaxL) and Minimum Load (MinL).

Step 2: In second step this algorithm target to eliminate the least-loaded PMs as much as possible. It transfers all VMs from under-loaded PM to most-loaded PMs While ensuring that the destination PM should not be over-loaded by the migrated VMs.

B. *Self-Managing Utility Function:*

In [4], author proposed self-managing policy, with a goal to maximize the profits by reducing energy consumption and SLA violations during VM placement. Equation (3) defines of the utility function $\left(Utility(a,t)\right)$ for self-managing policy as:

$$\mathrm{income(a,t)} - \left(\mathrm{EECost(a,t)} + \mathrm{EVCost(a,t)} + \mathrm{PDMCost(a,t)}\right) \qquad (3)$$

Where a represents a map for assigning selected VMs to PMs, (t) represents time period of each assignment, $\mathrm{income(a,t)}$ represents the total amount of hosting income from cloud's customers, $EECost(a,t)$ is the estimated amount of energy to be consumed and $EVCost(a,t)$ *is* the cost of SLAVs

For the development of utility based model five algorithms were presented. Following is an overview for each of the algorithms.

Algorithms-1 estimates the CPU utilization of PM under the workload of an assignment and records the utilization of each CPU of every PM that is linked to the given assignment.

Algorithm-2 calculates the expected amount of energy that is to be consumed in a given assignment in a time interval *t.*

Algorithm-3 calculates the violation cost of over-loaded PMs.

Algorithm-4 calculates the violation cost due to VM migration.

Algorithm-5 provides a pseudo code of an adopted algorithm that is modified with genetic algorithm for finding such VMs to PMs Assignments that maximizes the utility function.

C. AVVMC

In [5] AVVMC, a null solution along with a set of PMs and randomly selected VMs are assigned to each ant. After initialization a random ant is selected to choose a single VM applying probabilistic decision rule on all feasible VMs. The selected VM is then assigned next to its current PM.

Once every ant has a solution, best solution is picked based upon the value of objective function. For assigning a VM desirably to PM pheromone levels are associated with all VMs. Heuristics values are calculated dynamically for every assignment to balance out resource utilization. Equation (4) is used to initially assign pheromone levels and (5) is used to update pheromone levels.

$$\tau 0 = \ PE_{FFDL1Norm} \qquad (4)$$

where $PE_{FFDL1Norm}$ is the solution of Packing Efficiency of the FFD heuristic.

$$\tau_{v:p=(1-\delta)\ \tau_{v:p}+\ \delta\ \Delta\ \tau_{v:p}} \qquad (5)$$

Where *(δ)* indicates a decay parameter for pheromone and $\Delta\tau_{v:p}$ indicate pheromone reinforcement applied on each *v* - *p* pairs.

D. Honey Bee Cluster Technique (HCT)

In [6] a Honey Bee Cluster Technique (HCT) is proposed, the algorithm works in following steps:

1) Resources (CPU, Memory) are clustered using a technique called hierarchical clustering in which each cluster behaves as a single resource.

2) Assignment is categorized and honeybee parameters (*n, m, s, Iteration, α*) are initialized. Where *(n),(m),(s)* represents the number of employed bees, on-looker bees and scout bees respectively, *(iteration)* represents the max iteration number and *(α)* represents initial penalty parameter value.

3) A solution is constructed for employee Bee initialization. After that every employee bee locates a suitable VM for each task.

4) Evaluation of fitness function for each employed bee is calculated using (6).

$$\text{Fitness} = \frac{CT\ max}{CTi} + \frac{1}{PTi} \qquad (6)$$

where CT_{max} is total completion time of a task, CT_i is completion time of specific task and PT_i is the processing time for specific task.

E. PABFD-MCC

In [7], author presented an energy efficient architecture for IaaS layer along with a placement approach for VMs using a method called PABFD-MCC. The key idea behind this approach is to provide a suitable tradeoff between SLA violations and energy consumption, as the resource usage in a VM gets higher, the probability of host overloading is also increased along with a risk of SLA violations.

1) The correlation of CPU utilization among VMs is estimated, along with the CPU correlation between migrating VMs.

2) A best suitable host is selected for VM placement based on score assigned to each host, score is calculated on the basis of following factors:

- Power consumption of PM after VM allocation
- Correlation coefficient between migrating VMs and VMs running on PM.

F. ESWCT and ELMWCT

In [8] a two part algorithm is proposed, First part of algorithm is ESWCT which locates a placement for VM to get an effective and balanced host's resource utilization. Imbalance Utilization Value (IUVi) is calculated among multi dimension recourses of physical server. Equation (7) is used to define IUV_i as:

$$\frac{(CPU_i^u-IR_i^u)^2+(MEM_i^u-IR_i^u)^2+(NET_i^u-IR_i^u)^2}{3} \qquad (7)$$

Where IR_i^u is integrated resource utilization of a server (*i)* which is defined in Eq.8, $CPU_i^u\ \ MEM_i^u\ \ NET_i^u$ is the average usage of CPU, MEM and NET of server (*i)* respectively.

$$\frac{CPU_i^u+MEM_i^u+NET_i^i}{3} \qquad (8)$$

ESWCT works on following three steps:

1) Compute the capability of every component of each PM.

2) Get the component capability of VM.

3) Assign VM to PM with smallest IUV_i value.

ELMWCT is the second part of the algorithm, first it chooses the VM which are needed for migration, for this purpose node utilization threshold vector is introduced. Afterward VMs are allocated to PM using ESWCT algorithm.

G. Ant Colony Optimization (ACO) based Workload Placement

Another [9] ACO optimized VM placement algorithm is proposed for VM consolidation with a goal to reduce the number of active PMs and maximizes resource utilization. A multi-dimension bin packing workload placement problem is modeled for validating proposed solution. The algorithm works in following steps:

- Input parameters including a set of VMs and active PMs along with vectors including resource demand and resource capacity are given to the algorithm.
- Parameters are initialed and a pheromone trail is set to τ_{max} . After this, algorithm starts its iterations up to a defined number $(nbCycles)$ in every iteration each ant opens up a bin and starts building its solution.
- VMs are then assigned to PMs using probabilistic decision rule until there is no VM left or bin capacity is full.
- Once each ant finished building its solution the best solution is saved.
- Values of τ_{min} and τ_{max} are computed and pheromone levels are updated using pheromone update rule.

Following equations are used to for probabilistic decision rule and pheromone update, respectively.

$$P_v^i = {[\tau_{i.v}]^{\alpha} \cdot [\eta_{i,v}]^{\beta}} \Big/ {\sum u \in N_v [\tau_{u.v}]^{\alpha} \,.\, [\eta_{u,v}]^{\beta}} \qquad (9)$$

where $(\tau_{i.v})$ represents pheromone based desirability and $(\eta_{i,v})$ describes heuristics information.

$$\tau_{i,v} = (1 - \rho)\,.\,\tau_{i,v} + \Delta\tau_{i,v}^{Best} \qquad (10)$$

where $\left(\Delta\tau_{i,v}^{Best}\right)$ represents iteration's best item bin pheromone amount.

H. EEVS Algorithm

In [10], the proposed algorithm performs scheduling in three phases, VMs allocation, Updating VMs and cloud reconfiguration. To set an optimal frequency for each PM, DVFS technology is adopted. Set of VMs and PMs are given to algorithm as input and schedule of VM, energy consumption and processing time is the output from algorithm. Algorithm works in following steps:

- Sorts PMs in a decreasing order of optimal performance power ratio.
- Starts VM allocation for a specific time period.
- Optimal frequency is computed along with practical optimal frequency and assigned to each primary machine.
- Information for active PMs and running VMs are updated.
- After this, phase three for cloud reconfiguration begins.

The proposed methodology although consumes less amount of energy and processes more VMs than compared algorithm but there is some downsides to this approach. Two assumptions were made, VM migration and performance-power penalties of processor transition were ignored during the validation of this approach. Without these values this approach may not work well in practical cloud environments.

I. VMPACS

A nature inspired ACO approach [11] VMPACS is proposed, In this algorithm input parameters including a set of VMs, PMs with associated resource demand and the specific limit for resource utilization is given to the algorithm and a Pareto set P is output by the algorithm. The primary working VMPACS is almost identical to other nature inspired algorithm with a slight change for the use for pseudo-random proportional rule for VM placement. MGGA algorithm is selected for the comparison of VMPACS, as MGGA claims to be effective in multi objective VM placement problem. Following Pseudo-Random proportional rule is used by an artificial ant *(k)* to select a VM *(i)*.

$$i = \{_{s,}^{argmax_{u \in \Omega_k(j)} \{\alpha.\tau_{u,j} + (1-\alpha).\eta_{u,j}}} \qquad (11)$$

where α represents a control parameter for pheromone trail and q represents a uniformly distributed random number between 0 and 1, $\Omega_k(j)$ represents a set of VMs which qualifies for placement in PM (j).

VMPACS follows these steps to perform its operation:

- Parameters were initialized and value to all the pheromones trails are set to τ_0 .
- Each ant is set to receive all VMs request, and starts assigning VMs using pseudo-random proportional rule to active PMs. Once an ant finished building its path pheromone values are updated along with a global update of each Pareto set.

J. HCMFF and MFF

In [12] two energy effective approaches were proposed, Modified firefly algorithm (MFF) is implemented by modifying traditional firefly algorithms and hierarchical cluster based MFF (HCMFF) is implemented by modifying firefly algorithm with hierarchical clusters. Firefly approach uses male and female species of fireflies to populate its individuals. Both of these species can be assigned to different task, hence in proposed approach male fireflies represent PMs and females represent VMs. Both of these proposed algorithms were simulated and results were compared with traditional algorithms showing an effective reduction of energy consumption while maintaining SLA violations.

Following objective function is used to determine the attractiveness using the brightness of firefly.

$$x_j = PEnum_j \; X \; PEmips_j \; X \; VMbw_j \qquad (12)$$

where $PEnum_j$ is the number of processors allocated to VM_j, $PEmips_j$ is total number of instructions of all processors in millions per second for each VM_j,

MFF works in following steps:

- Once initialization of parameters an introductory population of fireflies was created.

- For all male and female fireflies light intensity is initialized, after this a firefly (*i*) is moved toward (*j*), with varying attractiveness of light due to distance, light intensity values were updates.
- Best solution is found by ranking fireflies.

HCMFF works with the same principal of MFF but with a slightly different approach.

By using a technique called hierarchical clustering, resources are clustered in terms of resource, bandwidth and memory.

- Every cluster behaves as a single resource in which VMs are categorized regarding requirements, after those firefly parameters were initialized.
- Firefly population is initialized and attractiveness of each individual is computed based on the objective function.
- Current solution is then found by ranking fireflies.

V. Comparative Analysis

To compare selected energy aware VM consolidation approaches, the above mentioned techniques are compared side by side on the basis of multiple factors. Number of VMs and PMs used to simulate results are different in each proposed methodologies. A detailed comparative analysis is shown in Table II.

TABLE II. Comparative Analysis of Energy Aware VM Consolidation Techniques

Technique	*Compared With*	*PMs*	*VMs*	*Resource*	*Assessment Model*	*Selection Policy*	*SLAV*	*Energy(KwH)*	*VM Migrations*
UP-VMC [3]	PUP-VMC VUP-VMC ACS-VMC SERCON MBFD MFFD	Load dependent	265	CPU Mem.	K-NNR	MMT	$6.5x10^{-6}$	68	1700
						MinL	$9.8x10^{-6}$	102	2800
						MaxL	$1.5x10^{-5}$	70	2400
Utility Based [4]	Heuristic Based	100	150	CPU	Genetic algorithm	Self-Managing	$2.5x10^{-5}$	90	1550
MFF [12]	Honey Bee	800	1052	CPU	Firefly algorithm	IQR-MC	$9x10^{-6}$	34.77	888
						LR-MMT	$1x10^{-4}$	35.33	867
						MAD-MMT	$1.1x10^{-4}$	33.93	855
HCMFF [12]	HCT	800	1052	CPU	Firefly algorithm	IQR-MC	$8x10^{-5}$	34.17	889
						LR-MMT	$6x10^{-5}$	35.09	815
						MAD-MMT	$1.6x10^{-4}$	32.91	873
EEVS [10]	MBFD	100	700	CPU	optimal frequency	random	-	10 (K-Watts)	-
AVVMC [5]	MMVMC VECTORGREEDY FFDL1Norm FFDVOLUME	100	1000	CPU MEM IO	Ant Colony System	Pseudo-random Proportional Rule	-	21 (K-Watts)	-
HCT [6]	Honey Bee	800	1000	CPU Mem.	Honey bee	IQR/RS	$1x10^{-4}$	34.29	852
						LR/MU	$8x10^{-5}$	36.85	869
MMC [7]	LR-MMT PAFDB	80	100	CPU	Minimum Correlation Coefficient	LR/MMT	$7.81x10^{-5}$	11.69	-
ESWCT& ELMWCT [8]	Random iV-Value iDAIRS iVectorDot	100	500	CPU	-	ELMWCT	-	23 (K-Watts)	-
VMPACS [11]	MGGA	Load dependent	200	CPU Mem.	ACO	Random proportional rule	-	11.75 (K-Watts)	-
ACO [9]	FDD	30	100	CPU	ACO	Probabilistic decision rule	-	131.41	-

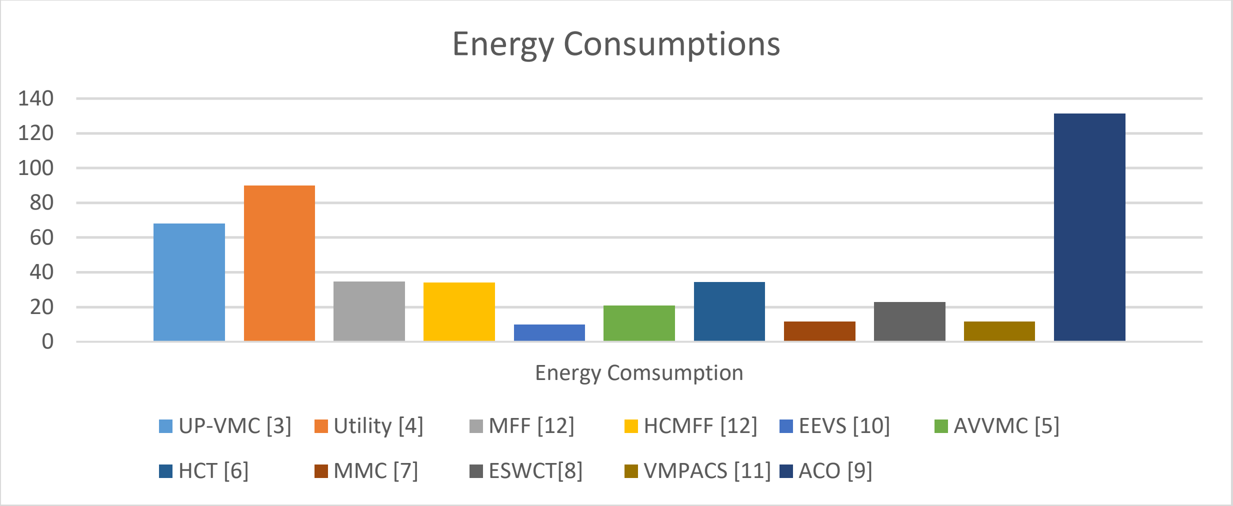


Fig. 1. Energy consumptions of multiple techniques in KwH.

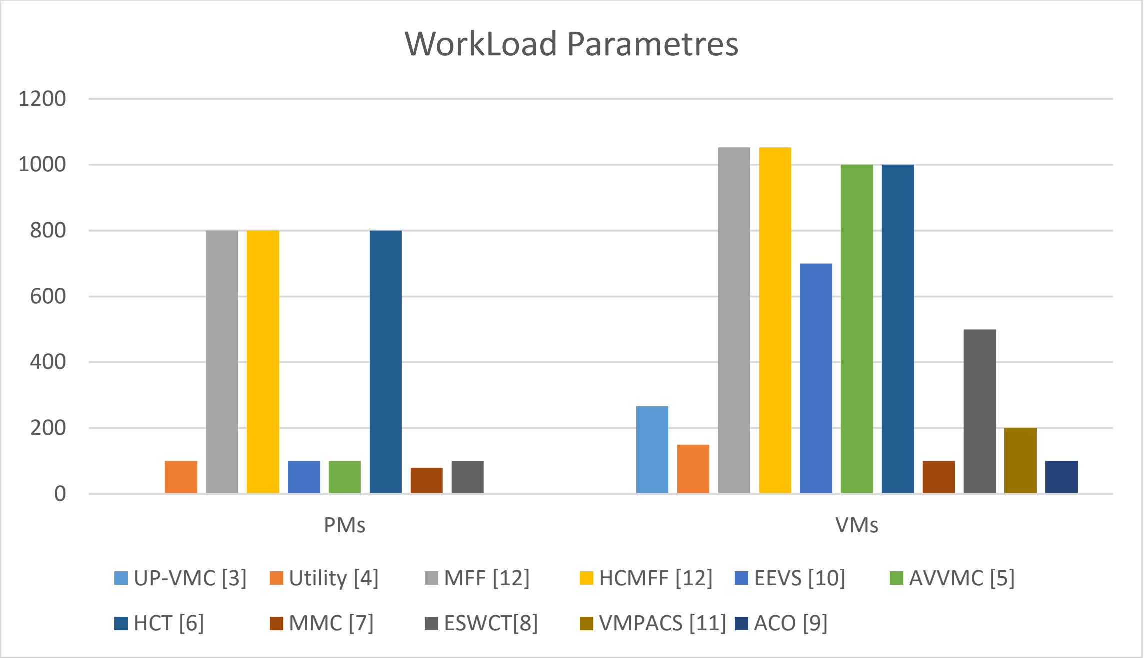


Fig. 2. Number of active PMs and VMs during workload testing.

Fig. 1 and 2 shows the amount of consumed energy by multiple energy aware techniques and the number of VMs and PMs used for the validation of results, respectively.

VI. Conclusion

As cloud computation applications are increasing daily, the number of servers used in cloud data centers are also increasing abruptly. Hundreds of thousands servers running in thousands of cloud datacenters around the world, consuming a large amount of energy. These vast power consumptions increase the volume greenhouse gases around the world. Hence a solution is required to reduce these energy consumptions to make cloud datacenters environmental friendly.

In this survey paper, multiples techniques for energy aware virtual machine consolidations are reviewed. All techniques have a primary goal of energy reduction with the minimum loss of cloud's applications efficiency. A technical analysis is given describing the proposed methodologies briefly, followed by a detailed comparative analysis.

Hence this survey paper will optimistically motivate future researchers to come up with smarter and secured optimal resource allocation algorithms and framework to strengthen the cloud computing paradigm.

References

[1] Shaden M.AlIsmail,Review of Energy Reduction Techniques for Green Cloud Computing,2016.

[2] Ankita Choudhary,A Critical Analysis of Energy Efficient Virtual Machine Placement Techniques and its Optimization in a Cloud Computing Environment,2015.

[3] Fahimeh Energy aware VM Consolidation in Cloud Data Centers Using Utilization Prediction Model, IEEE transaction on cloud computing, 2016.

[4] Abdelkhalik Mosa, Optimizing virtual machine placement for energy and SLA in clouds using utility functions, Journal of Cloud Computing: Advances, Systems and Applications, 2016.

[5] Md Hasanul Ferdaus et.al, Virtual Machine Consolidation in Cloud Data Centers using ACO Metaheuristic, 2014.

[6] Ajith Singh, cluster based bee algorithm for virtual machine placement in cloud data centre, Journal of Theoretical and Applied Information Technology,2013

[7] Negin Kord, An Energy Efficient Approach for Virtual Machine Placement in Cloud Based Data Centers,2013

[8] LI Hongyou, Energy Aware Scheduling Scheme Using Workload Aware Consolidation Technique in Cloud Data Centres,2013.

[9] Eugen Feller, Energy Aware Ant Colony Based Workload Placement in Clouds, 2011.

[10] Youwei Ding, Energy efficient scheduling of virtual machines in cloud with deadline constraint, 2015.

[11] Yongqiang Gao, A multi objective ant colony system algorithm for virtual machine placement in cloud computing,2013

[12] Esha Barlaskar, Energy efficient virtual machine placement using enhanced firefly algorithm, 2016.

[13] Zoha Usmani, A Survey of Virtual Machine Placement Techniques in a Cloud Data Center,2015

Ant Colony System for Dynamic Vehicle Routing Problem with Overtime

Khaoula OUADDI, Youssef BENADADA, Fatima-Zahra MHADA
Smart Systems Laboratory, Rabat IT Center
ENSIAS, Mohammed V University in Rabat, Madinat Al Irfane, BP 713, Agdal Rabat, MOROCCO

***Abstract*—Traditionally, in a VRP the vehicles return to depot before the end of the working time. However, in reality several constraints can occur and prevent the vehicles from being at the depot on time. In the dynamic case, we are supposed to answer the requests the same day of their arrival. Nevertheless, it is not always easy to find a solution, which ensures the service while respecting the normal working time. Therefore, allowing the vehicle to use additional time to complete their service may be very useful especially if we have a large demand with a limited number of vehicles. In this context, this article proposes a mathematical modeling with an Ant Colony System (ACS) based approach to solve the dynamic vehicle routing problem (DVRP) multi-tours with overtime. To test the algorithm, we propose new data sets inspired from literature benchmarks. The competitiveness of the algorithm is proved on the classical DVRP.**

Keywords—Dynamic vehicle routing problem (DVRP); multi-tours; mathematical modeling; hybrid; Ant Colony System (ACS); overtime

I. Introduction

Being defined more than 50 years ago, Vehicle Routing Problem (VRP) is one of the most classical combinatorial optimization problems. The main objective is to find the optimal path that can visit all nodes in question. For example, in a capacity VRP, these nodes are a set of customers that need to be served from a single depot with limited load capacity vehicles. Another example is the VRP with pickup and delivery, in which, the nodes can be a set of customers who will be served and from which the goods can be collected and a set of depots where vehicles get off. For each of these two examples we can define several sub-variants by adding some specifications to the problem. For example, we can find a problem with the constraint of time windows during which customers must be visited. Another example is the stochastic case where the requested quantity is unknown exactly before visiting requesting customer. Accordingly, the VRP become a very large class of problems that several studies have devoted themselves to review and classify [1], [2].

Among VRP variants, the dynamic VRP (DVRP) is relatively recent. In this variant, the information available at the beginning is incomplete and is subject to random variations over time. In other words, the starting solution is adjustable according to new data. In Fig. 1, we have a simple example of a DVRP where a vehicle has to serve a set of customers. Dynamic customers (E and F) are inserted in the new planned routes taking into account customers not yet visited (in dotted line).

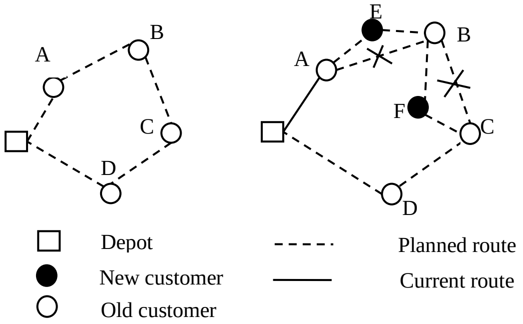


Fig. 1. Example of DVRP.

This concept, relatively new, has brought several advantages that are potentially beneficial for transportation companies; it can be helpful for companies to increase their competitiveness. As long as it allows companies to serve their customers on the same day of their requests and ensure, there is a better customer satisfaction. Besides, due to its flexibility and adaptability, DVRP is very useful in the emergency context where we need an immediate response to requests. Furthermore, the DVRP can handle dynamic travel time. This case is more present in urban areas where it is more difficult to predict network travel time because of the great congestion especially during peak hours.

All these advantages cannot be implemented without technological tools allowing real-time communication between the dispatcher and the driver. Fortunately, new communication and geographical location tools have allowed dispatchers and drivers to have a real time idea on the state of the network, and provide a real-time response to customers' requests.

That is why, Intelligent Transportation Systems [3] area set of platforms, each one is dedicated to a particular process. Among these platforms the Advanced Fleet Management Systems (AFMS) which are very useful in the DVRP case. They are specifically designed for dynamic or static business management of fleet while considering possible variations in the travel time on networks links.

As a result, since its first introduction by Psarftis [4] many variants of DVRP have been introduced and studied and literature studies have reviewed and classified them [5]-[8].

Despite all this number of variants treated, the concept of overtime still very little studied in the DVRP literature. This concept, which is widely used by transportation companies in general, is more needed in dynamic cases. Dynamically re-optimizing the current planning to insert new queries usually provides a solution that consumes more time by comparing it with a solution that considers all customers from the beginning. To verify this hypothesis we can just compare the results of the total traveled distance of the dynamic case in Kilby et al. [9] instances and the static case in those of Taillard et al. [10]. Even so, in this model, we must answer the maximum of requests the day of their arrival. If we have a limited number of vehicles, we will need more overtime especially in case of high demand. In this article, we introduce a new variant of DVRP with the concept of overtime. To this end, we will consider the case of a transportation company that has a homogeneous fleet of trucks and which responds dynamically to customers' request. Each truck can perform several tours during the day. A limited number of trucks are available to satisfy all customers 'orders. Consequently, it is not always easy to find a solution that ensures the service while respecting the normal working time. If necessary, the trucks are allowed to use overtime on the condition that it shouldn't exceed the maximum legal overtime. So, we have two objectives in this problem; Minimize the maximum overtime performed by trucks and minimize the total traveled distance. Thus, we propose a multi-objective mathematical model. To solve this problem we propose a hybrid Ant Colony System (ACS) algorithm

The rest of this paper is organized as follows: the second section presents a brief DVRP literature review. The third section presents the mathematical model. The methodology of resolution will be presented in the fourth section. Before concluding, we will present the numerical results in the fifth section.

II. DVRP: Literature Review

Going around DVRP literature we find all types of optimization methods known up to now, from the exact methods up to the metaheuristics. Otherwise, the literature addresses the DVRP according to four main perspectives: deterministic, stochastic, continuous and periodic. In the deterministic case we consider only known requests and deterministically respond to the dynamic ones. In other words, the current plan, takes into account the data actually known by the dispatcher. While in the stochastic case, we can consider stochastic data, such demand forecasts while elaborating a preliminary routing plan. In continuous optimization, the re-optimization starts at the arrival of each new request to get a new plan. While in the periodic one, the planning period is decomposed into time intervals. In this way, during each time interval, incoming requests are collected and inserted all at the end of the current interval or at the beginning of the next interval.

In order to have a global view of this research field, we will list some research works for the four main perspectives, and then we will exam those are related to the DVRP with overtime.

Let's start by works that have adopted deterministic continuous optimization. It started with Psaraftis [4], who is the first to introduce the dynamic DVRP. In this work, he used dynamic programming to deal with the dynamic dial-a-ride problem. The aim is to find the best route at each new demand. In the same perspective, Gendreau et al. [11] applied the Tabu search (TS) to solve the DVRP. As soon as a new request arrives, the algorithm saves its former results in the adaptive memory to insert the locations of the new requests. The TS was also applied, in a deterministic continuous way, by Chang et al. [12] on the DVRP with pickup, delivery and time windows.

As for the work done for the DVRP in a deterministic periodic manner, the first one we mention is the work of Chen and Xu [13] who have proposed an approach based on linear programming and dynamic column generation for the DVRP with time windows and infinite fleet. On their part, Hanshar et al. [14] decompose the working day into time intervals. Then they launch the optimization program at the beginning of the time interval. They propose a solution based on Genetic Algorithm (GA) for capacitated DVRP. The main contribution in their work was the way they represent a chromosome in dynamic optimizations. In the same perspective and using the Neighborhood Search Algorithm (NSA), Hong [15] solves the DVRP with hard time windows. It applies the withdrawal reinsertion mechanism of this algorithm to insert new queries in the already planned routes.

Another adaptation of the NSA to DVRP was proposed by Khouadjia et al. [16]. To conclude with the deterministic periodic manner, we will quote a very interesting work with ACS. It is the first application of the ACS on DVRP by Montemanni et al. [17]. In this masterpiece, the authors decomposed the planning period into time intervals. At the beginning of each time interval, the optimization program is launched to insert the incoming requests during the previous period in the planning of the rest of the day. The static problem is solved using an ordinary version of ACS. As for the other time intervals, the article applies the same algorithm with a modification of the initial rate of pheromone on the arcs of the network. As each sub-problem is potentially similar to its successor, a pheromone conservation mechanism is applied to put the weight on the arcs belonging to the previous solutions and thus reduces the execution time.

In a stochastic periodic manner, Hvattum et al. [18] have developed a heuristic approach whose principle is to divide the planning horizon into time intervals and to assign a set of promising queries to the vehicles at the beginning of each interval depending on their frequencies of occurrence in the possible stochastic scenarios. The algorithm then uses the Branch-and-Regret method to merge them in order to have a single optimal solution. Another example of the stochastic perspective, but this time in a continuous way, is that of Hemert and Poutré [19] who have used GA to solve a problem of collecting charges from customers and delivering them to a single central depot. The authors have introduced the notion of fruitful regions, where there are more probable potential customers. In the same perspectives, a more recent study of Schyns [20], aimed at optimizing the routing of refueling trucks in an airport, proposed an adaptation of the ACS to take into account the lack of visibility of the planning period and the hard time windows.

Our article will deal with the case of a deterministic periodic DVRP with overtime. After checking articles published in this area, we have found only one article that is the article of Gendreau et al. [21] which deals with a DVRP with pickup, delivery and time window by allowing the use of the overtime and without considering the capacity constraint. To solve this multi-objectives problem, the authors proposed NSA algorithm while adopting a deterministic continuous perspective.

III. Mathematical Model

The objective of this article is to solve a capacitated DVRP multi-tours that tolerate overtime. We have a homogeneous fleet of vehicles with limited capacity and a single depot. The vehicles leave the depot at the beginning of the day with a full capacity and return there to restock and start a new tour or to close their working day. In our case, we adopt the Montemmani et al. [17] approach, which divides the planning period into time intervals. Thereby, we collect all the incoming requests during a time interval, to insert them in the planning of the following period. In this way, the re-optimization is launched at the beginning of each interval and the found solution covers all the rest of the planning period.

We set T_f time limit to accept customers' requests. After this time, the incoming requests will be reported to be recorded in the next planning period. Thus, at the beginning of the planning period, we have a CVRP with a homogeneous fleet and a single depot. However, at the beginning of other time intervals, a vehicle that has already served one or more customers will have a less load as it will have a starting point other than the central depot. We will call this starting point fictitious depot. Therefore, we have a CVRP with heterogeneous fleet and several fictitious depots. In both cases, trucks are allowed to perform several tours. At the end of a time interval, if a truck is serving a customer, this later is considered as fictitious depots in the problem of the next time interval. Else, it will be on the road to a destination customer. In this case, this later is considered as fictitious depots in the problem of the next time interval. We present the mathematical model of one time interval. Thus, we put:

F: Set of depots

I: Set of Customers

F^*: Set of depots without central depot ($F\setminus\{0\}$)

I^*: Set of Customers and central depot ($I\cup\{0\}$)

K: Set of trucks

f: Index of depot (including fictitious depots and the central one)

i: Index of customers

k: Index of trucks

n: Maximum number of tours for a truck

0: Index of the central depot, $0 \in F$

d_{ij}: Cost (distance) of movement between i and j

t_{ij}: Travel time between i and j

Q_k: Remaining capacity of the truck k

Q: The initial capacity of trucks

T_l: Normal driving time remaining for the period l

T: Length of working period

αT: Maximal legal overtime

q_i: Quantity requested by the customer i

$$f_k=\begin{cases}1, \text{ if the truck k is initially stationned in fictitious depot } f\in F^* \\ 0 \quad \text{else}\end{cases}$$

We have a single decision variable:

$$x_{ij}^r=\begin{cases}1, \text{ if the customer j was visited after the customer i during the tour r} \\ 0 \quad \text{else}\end{cases}$$

To simplify the problem formulation, we assume that each vehicle can carry a maximum of n tours during the remainder of the day. Considering that we have K trucks in service, the maximum number of tours that can be achieved is nK. Excluding already served customers, we consider that a tour starts from the fictitious depot where the customer exists at the beginning of the current time interval.

Therefore, for every truck k, we choose to reserve n indices of tours that will be eventually assigned to it k,k+K,..,k+(n-1)K. Tours r with indices $\sum_{i=0}^{N}\sum_{j=0}^{N} x_{ij}^r = 0$, refers to tours that will not actually be performed by the trucks.

So we define S_k the set of n possible tours served by the truck k.

$$S_k=\{k+qK, q = 0..n\text{-}1\}$$

$$S_T=\bigcup_{k\in K} S_k$$

If a truck k actually makes n successive rounds, these rounds will have respectively the index k, k + K..., k +(n-1)K. The tour with the index k is the first to be performed, then k+K and finally the owner of the index k + (n-1)K. In this way, if the truck k is initially parked in a fictitious depot different from the central depot, the tour of index r = k must imperatively start from this fictitious depot. Any other tour starts from the central depot. So, if r$\epsilon\{1,..,K\}$ then, the tour r=k, starts from depot f>0. Else, r starts from the central depot. Now, we define the available capacity for the tour r as:

$$Q_r=\begin{cases}Q_k & \text{if } r\in\{1,..,K\} \\ Q & \text{else}\end{cases}$$

The travel time of a trunk k is the necessary time to serve all its tours. However the overtime for each vehicle is the additional time over the planning horizon whose vehicle needs to ensure its service. If the vehicle respects the constraint of time, the overtime is zero. It can be calculated for a trunk k as follow:

$$OT_k = \max\left(0, \sum_{r\in S_k}\sum_{i\in I\cup J}\sum_{\substack{j\in J\\ j\neq i}} t_{ij}x_{ij}^r \text{ -}T_l\right)$$

The solution overtime is the maximum overtime of all vehicles. It’s related to the last truck that returns definitively to the depot:

$$OT_l = \max_{0\leq k\leq K} OT_k$$

During each period of re-optimization, the mathematical model consists of minimizing the maximum overtime and the routing cost. If the problem has a solution that respects the constraints without needing overtime, then the objective in this case is to minimize the total traveled distance. Otherwise the objective will be to minimize the total overtime of the solution. The objectives of the problem can be formulated as follows:

$$\min \sum_{r\in S_T}\sum_{i\in I\cup J}\sum_{j\in J} d_{ij}\, x_{ij}^r$$

$$\min(OT_l)$$

Under the following constraints:

$$\textstyle\sum_{i\in I\cup F}\sum_{\substack{j\in I^*\\ j\neq i}} x_{ij}^r \leq Q_r, \forall r\in S_T \tag{1}$$

$$\textstyle\sum_{r\in S_k}\sum_{i\in I\cup F}\sum_{\substack{j\in I^*\\ j\neq i}} t_{ij}x_{ij}^r \text{ -}T_l \leq \alpha T\ ,\ k\in K \tag{2}$$

$$\textstyle\sum_{r\in S_T}\sum_{\substack{j\in I^*\\ j\neq i}} x_{ij}^r = 1, \forall i\in I\cup F^* \tag{3}$$

$$\textstyle\sum_{\substack{i\in I\cup F\\ i\neq j}} x_{ij}^r = \sum_{\substack{i\in I^*\\ i\neq j}} x_{ji}^r\ , \forall\ j\in I^*, \forall\ r\in S_T \tag{4}$$

$$\textstyle\sum_{r\in S_k}\sum_{i\in I\cup F}\sum_{\substack{j\in I^*\\ i\neq j}} x_{ij}^r \geq f_k,\ \forall k\in K \tag{5}$$

$$\textstyle\sum_{r\in S_T}\sum_{\substack{i\in I\cup F\\ j\neq i}} x_{ij}^r = 0 \tag{6}$$

$$\textstyle\sum_{i\in S}\sum_{\substack{j\in S\\ j\neq i}} x_{ij}^r \leq |S|\text{-}1,\ \forall r\in S_T,\ S\subset I,\ 2\leq |S| \leq\ |I|\text{-}1 \tag{7}$$

$$x_{ij}^r \epsilon \{0,1\},\ (i,j)\in I\cup F,\ r\in S_T \tag{8}$$

The first constraint (1) is made to comply with the remaining capacity of trucks in each tour. The second equation (2) restricts the overtime to a permitted maximum value. The third constraint (3) ensures that each customer is visited once and only once. The flow conservation constraint at the customer and the central depot level is carried out by (4); each truck that visits a customer must leave him after his delivery request and every truck that leaves the central depot must come back at the end of the working period. While the constraint (5) is made to ensure that trucks initially parked in a fictitious depot other than the central one must perform at least one tour. In this way we will be sure that these trucks return to the central depot at the end of the day. The sixth constraint (6) expresses that a fictitious depot can't be a destination. The constraint (7) prohibits the creation of sub-tours. Finally, the integrity constraints associated with decision variables are included in (8).

A study of Hassein and Rubinstein [22] has previously proved that an ordinary VRP is considered NP-hard when the set of customers contains more than 3 customers, so impossible to solve by an exact method. A meta-heuristic can therefore be used to solve this problem in order to find a good solution in a reduced time. The mathematical model just described is done to better describe the specificities of the problem and the objective functions.

IV. Resolution by the Hybrid ACS

A. Static Problem

To solve the problem that has just been described, we propose an approach based on the ACS. Firstly, we solve the problem of the beginning of the day. This problem, characterized by a single central depot and a homogeneous fleet, contains yesterday's requests arriving after the time T_f. The steps of our algorithm are described in Fig. 2.

First, we try to find a realizable solution that respect the normal time by minimizing only the first objective of the problem. If such solution isn't found we try to minimize the second objective without exceeding the maximal permitted overtime. Thus, the first step is affecting a positive value of τ_0 to each arc. Then, we place an ant on each customer. This allows for more diversification in the solutions found. The ants build, then, their tours. To move from one customer to another, the ant must select the clients that can visit without violating the constraints of the problem (remaining truck capacity and maximum return time). If no client responds to these constraints, the ant returns to the depot to begin a new tour. Otherwise, the ant chose the next customer to visit according to the following rule:

$$j = \begin{cases} \operatorname{argmax}_{u\in N_i^k}\left[\tau_{iu}(t).\eta_{iu}^{\beta}\right] & \text{if } q\leq q_0 \\ J & \text{if } q>q_0 \end{cases}$$

Where

q: A random variable uniformly distributed on [0,1]

q_0: A parameter in the interval [0,1]. It defines the balance diversification / intensification;

$\eta_{iu}=\frac{1}{d_{iu}}$: The visibility of the arc (i,u). It corresponds to the inverse of the distance d_{iu} between i and u.

$\tau_{iu}(t)$: The pheromone rate on arc (i,u) at the instant t.

β: The relative influence of the visibility.

N_i^k: The set of nodes that can be visited just after the position i by the ant.

$J\in N_i^k$: A randomly selected customer with probability:

$$p_{ij}^{k}(t)=\frac{\tau_{ij}(t).\eta_{ij}^{\beta}}{\sum_{u\in N_{i}^{k}(t)}\tau_{iu}(t).\eta_{iu}^{\beta}}$$

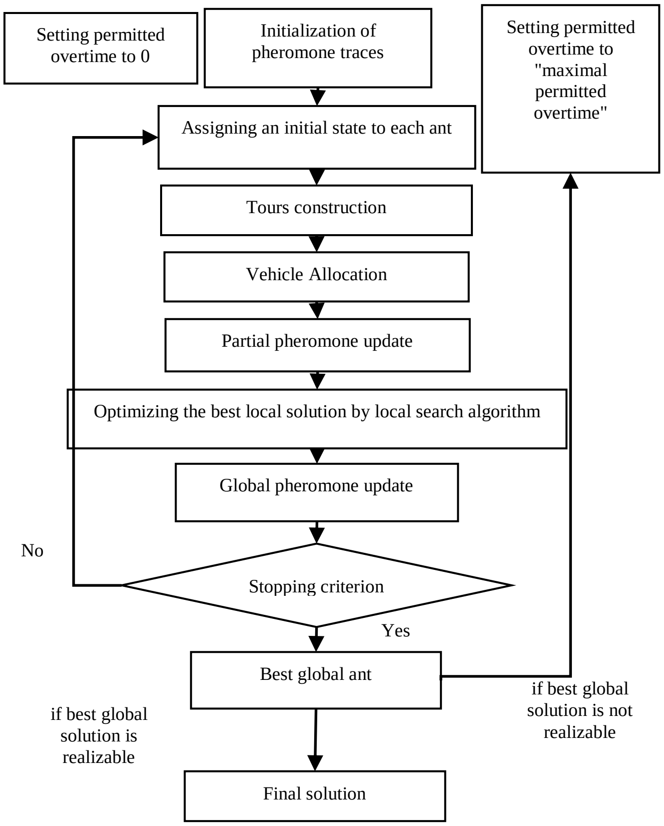

Setting permitted overtime to 0
Initialization of pheromone traces
Setting permitted overtime to "maximal permitted overtime"
Assigning an initial state to each ant
Tours construction
Vehicle Allocation
Partial pheromone update
Optimizing the best local solution by local search algorithm
Global pheromone update
No
Stopping criterion
Yes
Best global ant
if best global solution is not realizable
if best global solution is realizable
Final solution

Fig. 2. Steps of the algorithm.

After building a solution by an ant k, we allocate its tours to the vehicles in order to calculate the total overtime. The procedure of vehicle allocation is described in Fig. 4. Then, a local update of pheromone is done on arcs it visited according to the following formula:

$$\tau_{ij}(t+1)=(1-\rho)\tau_{ij}(t)+\frac{\rho.\tau_{0}}{1+OV_{k}}$$

Where

ρ: The evaporation factor set to a positive value less than 1. This factor is intended to avoid the unlimited accumulation of the pheromone traces on the edges of the graph.

OV_k: The overtime performed by the ant k.

Before applying the global update, the best solutions of the iteration undergoes a phase of optimization by a local search intra and inter tour. The global update is then carried out as follow:

$$\tau_{ij}(t+1)=(1-\rho)\tau_{ij}(t)+\rho.\Delta\tau_{ij}(t)$$

where, the arcs (i,j) belong to the best solution S of the iteration. $\Delta\tau_{ij}=\frac{1}{L_s}$ and L_s: is the total time of S.

At the end of each iteration, the solution of the best ant undergoes an optimization phase by local search algorithm that combines an intra and inter tour optimization (Fig. 3). This procedure is the same used by Ayadi and Benadada [23]. In this phase we look to optimize, only, the total traveled distance.

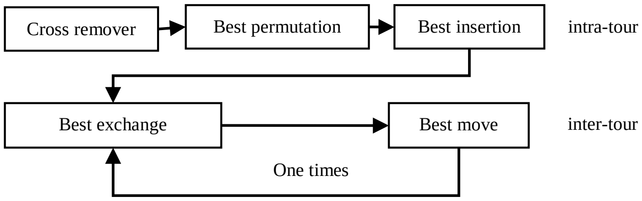


Fig. 3. Local search algorithm.

```
If (ToursNumber=<TrukNumber)
Assign a truck to each tour
End if;
Else
Rank the tours from the longest to the shortest;
Assign a truck to the TruckNumber tour;
While (there is "not assigned tours")
Rank the trucks according to the total length of the
assigned tours in an ascending order;
Assign the first free tour to the first truck
End while
End else
```

Fig. 4. Vehicle allocation procedure (Static problem).

In a first step, the crossovers are eliminated and the possible permutations, that decreases traveled distance within each tour, are sought. Then, the optimal permutations carried out. Afterward, a reinsertion operator comes to see if there is any better location for a customer in its tour. The optimal reinsertion is carried out. These three operations (elimination of crossovers, permutation and reinsertion of customers) form the intra tour local search. Subsequently, we look for the possibility of exchanging two customers of two different tours. Indeed, the optimal exchange is carried. Next, we try to find the possibility of moving a customer from one tour to another without violating problem constraints. The optimal displacement is chosen to be performed. These last two operations form the local search inter tour. Finally, the intra tour search is performed once, while the inter tour search is performed twice.

B. Dynamic Problem

In the dynamic case, we have the same steps described in the static case. However, the details are different since heterogeneous fleets and many fictitious depots, beside the central one, characterize the dynamic problem. The initialization of pheromone rate is done according to the pheromone rate retention mechanism proposed by Montemanni et al. [21]:

$$\tau_{ij}=\left(1-\gamma_{r}\right)\tau_{ij}^{old}+\gamma_{r}\tau_{0}$$

where,

τ_{ij}^{old}: The final value of the pheromone level on the arc (i,j) of the previous problem.

γ_r: A positive factor less than 1, which regulates the conservation of pheromone level.

τ_0: The initial constant of pheromone. It initializes arcs corresponding to the new clients.

Once pheromone traces are initialized, we place an ant on each fictitious depot. Each ant creates its tours by visiting customers one after one until the capacity or the remaining time don't allow to insert a new customer, in this case it returns to the central depot.

The ant starts the new tour from the nearest unvisited fictitious depot or from the central depot if all fictitious depots are visited and there are still unvisited customers. The move, from the central depot to the fictitious depot is done without deposing pheromone.

The transition rule depends on a parameter$0\le q_0\le 1$, which defines the balance diversification / intensification. In order to choose the next customer j, the ant k located in the customer i use the following rule:

$$j=\begin{cases}\operatorname{argmax}_{u\in N_i^k}\left[\tau_{iu}(t).\eta_{iu}^{\beta}\right] & \text{if } q\le q_0\\ J & \text{if } q>q_0\end{cases}$$

With

q_0 : A parameter in [0,1] that determine the balance diversification/ intensification

q: A random variable uniformly distributed on [0,1]

$\eta_{iu}=\frac{1}{d_{iu}}$: The visibility of the arc (i,u). It corresponds to the inverse of the distance d_{iu} between i and u.

$\tau_{iu}(t)$: The pheromone rate on the arc (i,u) at the instant t.

β: The relative influence of the visibility.

N_i^k: The set of nodes that can be visited just after the position i by the ant k.

$j\in N_i^k$: A randomly selected customer with probability:

$$p_{ij}^{k}(t)=\frac{\tau_{ij}(t).\eta_{ij}^{\beta}}{\sum_{u\in N_i^k(t)}\tau_{iu}(t).\eta_{iu}^{\beta}}$$

The update of pheromone is divided into two levels: a local update and a global one. The first is done after building a solution by an ant k. We modify the pheromone of arcs visited by this ant according to the following formula:

$$\tau_{ij}(t+1)=(1\text{-}\rho)\tau_{ij}(t)+\frac{\rho.\tau_0}{1+OV_k}$$

OV_k: The overtime performed by the ant k after affecting vehicles to its tours. The procedure of vehicle affectation is detailed in Fig. 5.

At the end of each iteration and before applying the global update, the best solution found in this iteration undergoes a phase of optimization by a local search intra and inter tour. To this end, we use the local search algorithm of the static problem.

```
Assign the tours starting from fictitious depot to corresponding truck
Rank the remaining tours from the longest to the shortest;

While (there is "not assigned tour")
        Rank the trucks according to the total length of the assigned tours in an ascending order;
        Assign the Firsttour to the Firsttruck
        Delete Firsttour from "not assigned tours"
end
```

Fig. 5. Vehicle allocation procedure (dynamic problem).

Finally, we must note that; if in a time interval of the problem we cannot find a solution without overtime, we will have one goal (minimize overtime) in the following time interval. In other words, we always try to minimize the total distance but once the algorithm ceases to find feasible solutions, the objective becomes the minimization of the overtime for all the rest of the planning period.

V. Computational Results

We have parameterized our algorithm according to the results found by Gambardella et al. [10] on the classical VRP. Thus, we set:q_0=0.9, β=1, ρ=0.1 Gambardella et al. [24] gave τ_0 the value of $\frac{1}{n*Cost(PI)}$ where Cost(PI) is the cost of a solution found by a greedy heuristic; In our case we fixed it in the static case to τ_0=0,1. This value was chosen by comparing its results with those of other values. In the dynamic case, we set $\tau_0=\frac{1}{n_d*Cost(PI)_p}$ where n_d is the number of clients of the current dynamic problem and $Cost(PI)_p$ is the cost of the previous problem; In this way, we reduce the execution time consumed by the greedy heuristic, especially since the current problem is similar to the previous one.

For the number of time intervals and the T_{co} we adopt the same parameter of Montemmanni et al. [17] (T_{co}=0.5*T and 25 time intervals). The algorithm stops after 200 iterations for all instances. It is coded in Java and executed on a machine with Intel Core i5 processor, 2.6 GHz, 8GB of RAM and Windows 8 as the operating system.

A. Static Case

The results of the static algorithm are already detailed in [25]. In this paragraph we give a reminder of these results.

We have tested our static algorithm on extracts data sets from Taillard et al. [10]. The problems used in these tests are: CMT1, CMT2, CMT3, CMT5, F134.

To avoid a zero in the denominator of η_{ij}, we have grouped customers who have the same coordinates in a single customer with a demand equal to the sum of the requests. Therefore, in F134, we have 132 customers and in CMT5, we have 199 customers. Table I presented our results referred to as OBM16, compared with those of Ayadi and Benadada [23] referred to, as AB13. While Table II presented the average of execution time of compared works, knowing that the AB13 time is normalized to our processor according to Geekbench benchmarks.

To qualify the solutions obtained, we use the Longest Trip Rate (LTR). It compares the time of the longest trip, which corresponds to the time of the truck that return the last to the depot, on the normal time horizon T:

$$\text{LTR}= \max_{k\in\{1..K\}} t(k) /T$$

With t(k) the time taken by the vehicle to make his tours.

TABLE I. NUMERICAL RESULTS OF THE STATIC CASE

Problem	T	M	AB13		OBM16	
			Distance	*LTR*	*Distance*	*LTR*
CMT1 z*=524,6	T_1	1	524,92	0,953	514.981	0.936
		2	536,56	0,998	523.768	0.991
		3	561,01	1,023	550.538	1.032
		4	547,1	1,027	536.329	1.050
	T_2	1	524,94	0,91	521.038	0.903
		2	530,79	1	512.242	0.941
		3	556,61	0,993	535.160	0.992
		4	546,43	0,985	538.937	1.003
CMT2 z*=835,2	T_1	1	835,77	0,953	842.325	0.960
		2	839,58	0,956	846.619	0.969
		3	836,18	0,964	872.351	1.043
		4	835,77	0,974	886.157	1.038
		5	839,71	0,996	845.569	1
		6	861,88	0,999	857.787	1.065
		7	885,57	1,034	869.800	1.119
	T_2	1	835,26	0,909	832.521	0.905
		2	835,26	0,912	856.500	0.934
		3	835,26	0,916	852.197	1.006
		4	835,77	0,959	875.851	0.963
		5	835,77	0,971	867.698	0.974
		6	842,28	0,997	870.636	1.003
		7	870,19	0,998	909.785	1.069
CMT3 z*=826,14	T_1	1	830	0,957	876.870	1.011
		2	828,74	0,97	865.569	1.005
		3	829,91	0,97	906.932	1.087
		4	828,74	0,982	870.609	1.027
		5	833,98	0,996	911.458	1.121
		6	867,72	1,008	986.338	1.198
	T_2	1	828,26	0,911	896.513	0.986
		2	828,26	0,913	891.445	0.991
		3	829,51	0,931	895.420	0.995
		4	829,54	0,969	871.22	0.999
		5	834,42	0,959	924.410	1.053
		6	836,56	0,994	921.07	1.082
F134 z*=1162,96	T_1	1	1166,96	0,956	1162.919	0.952
		2	1163,53	0,954	1142.409	0.945
		3	1162,97	0,964	1117.997	0.984
	T_2	1	1166,71	0,912	1137.630	0.889
		2	1163,53	0,911	1156.821	0.912
		3	1165,98	0,962	1143.229	0.950
CMT5 z*=1291,44	T_1	1	1313,22	0,968	1417.287	1.045
		2	1312,58	0,969	1479.293	1.092
		3	1313,21	0,969	1465.134	1.129
	T_2	1	1316,48	0,926	1469.505	1.034
		2	1380,38	0,972	1493.719	1.054
		3	1378,68	0,97	1467.871	1.091

TABLE II. AVERAGE OF EXECUTION TIMEIN SECONDS

Problem	Time AB13	Time OBM16
CMT1	39	11
CMT2	334	17
CMT3	3435	69
CMT5	5321	257
F134	6447	1246

B. Dynamic Casewithout Overtime

These results are already detailed in [26]. In this paragraph we give a reminder of these results.

In order to compare our algorithm with other works from the literature, we test it, first, on the benchmark of Kilby et al. [9]. To this end, we fix a single objective for the problem that is minimizing the total traveled distance. We consider five runs of the algorithm for each instance. Table III presents our results and those of Montemanni et al. [17]

TABLE III. NUMERICAL RESULTS OBTAINED BY OUR HYBRID ACS COMPARED TO THE ACS OF MONTEMANNI 2005

	Hybrid ACS			ACS Montemanni 2005		
	Best	*Averg*	*Time*	*Best*	*Averg*	*Time*
c50	685,81	708,64	0,20	631,3	681,86	4,10
c75	1077,39	1133,46	0,31	1009,36	1042,39	4,10
c100	1052,47	1078,13	1,93	973,26	1066,16	4,10
c100b	948,74	980,12	1,068	944,23	1023,6	4,10
c120	1279,32	1361,18	6,82	1416,45	1525,15	4,10
c150	1459,61	1535,27	4,564	1345,73	1455,5	4,10
c199	1827,81	1888,88	5,642	1771,04	1844,82	4,10
f71	316,6	321,05	4,336	311,18	358,69	4,10
f134	15675,0	16228,3	4,186	15135,5	16083,5	4,10
tai75a	1821,53	1916,60	0,398	1843,08	1945,2	4,10
tai75b	1555,78	1641,10	0,258	1535,43	1704,06	4,10
tai75c	1556,16	1632,73	0,662	1574,98	1653,58	4,10
tai75d	1514,16	1553,36	0,734	1472,35	1529	4,10
tai100	2225,45	2392,01	0,962	2375,92	2428,38	4,10
tai100	2384,79	2446,73	0,73	2283,97	2347,9	4,10
tai100	1662,35	1720,50	1,184	1562,3	1655,91	4,10
tai100	2008,47	2094,65	2,966	2008,13	2060,72	4,10
tai150	3368,62	3465,25	4,636	3644,78	3840,18	4,10
tai150	3082,21	3254,41	5,658	3166,88	3327,47	4,10
tai150	2842,62	2968,38	5,7	2811,48	3016,14	4,10
tai150	3253,3	3321,26	3,004	3058,87	3203,75	4,10
Total	51598,2	53642,2	55,95	50876,2	53794,0	86,24

The hybrid ACS outperforms the ACS Montemmani 2005 on 16 instances. The execution time is also lower than that of Montemanni 2005 with a percentage of 35% knowing that the Montemmani time presented in Table III is normalized to our processor according to Geekbench benchmarks.

C. Dynamic Case with Overtime

To test our algorithm, we have made some modifications on Kliby et al. [9] benchmark data set. Since, they are characterized by a big number of available trucks that is 50 for each instance, we can't have overtime with these instances.

We propose a data set for the DVRP with overtime using the twenty one basic Kilby problems; seven problems of Christofides et al. [27] (C), two of Fisher (F) [28] and twelve of Rochat and Taillard [29] (Thai). We use the same demands and truck capacities of the basic problems. Instances are generated by proposing several values of m (the number of available vehicle) and a restricted values of time horizon $T=[1,1*z^*/m]$, with z^* is the value of the best solution found by Rochat and Thaillard [29] for the VRP problem. The arrival time of customer requests is proportional to the arrival time set by Kilby et al. [9]. We set the maximum allowed overtime for each instance to one quarter of the normal time horizon. Time and distance are considered equivalent. The algorithm stops after 200 iterations for all instances. Table IV presents the value of T for m=1 and the size of each instance.

TABLE IV. DVRP INSTANCES

Problem	T	Size
CMT1	577	50
CMT2	919	75
CMT3	909	100
CMT4	1131	150
CMT5	1421	199
CMT11	114	120
CMT12	902	100
F71	266	71
F134	12979	134
Thai75a	1780	75
Thai75b	1479	75
Thai75c	1420	75
Thai75d	1502	75
Thai100a	2245	100
Thai100b	2134	100
Thai100c	1547	100
Thai100d	1739	100
Thai150a	3361	150
Thai150b	3000	150
Thai150c	2595	150
Thai150d	2910	150

For each instance, we tested m between 1 and 5. In Table V we write, just, the instances to which we found at least one feasible solution (which respects the maximum time plus the maximum permitted overtime). For each instance three runs of hybrid ACS are considered.

T is the normal maximum time while m represents the number of used vehicles. We consider the average of the total distance traveled and the overtime as well as the minimum of these two values obtained during the three executions. Time represent the average of the execution time in (min, second). A feasible solution is any solution that does not exceed the normal time plus the maximum overtime allowed. Instances denoted by * are instances in which only one feasible solution has been found, while those denoted by ** are instances in which two feasible solutions have been found. For both CMT4 and CMT5, we don't find any feasible solution. Therefore, they are not noted on the table. For the other instance, we found three feasible solutions.

TABLE V. NUMERICAL RESULTS OF DYNAMIC CASE WITH OVERTIME

Problem	**T**	**m**	**Average**		**Best**		**Time**
			Distance	***Over***	***Distance***	***Over***	
CMT1	577	1	629,95	52,95	625,38	48,38	0,66
	289	2	673,85	53,87	658,15	40,79	0,42
	192	3	672,24	36,28	652,52	26,37	0,37
	144	4*	674,09	36	674,09	36	0,3
CMT2	919	1	1082,73	163,73	992,85	73,85	0,44
	459	2	1079,1	81,5	1064,08	73,69	0,43
	306	3	1092,22	64,306	1069,12	58,87	0,38
	230	4	1060,74	48,81	1046,83	45,17	0,44
CMT3	909	1	1015,32	106,32	988,62	79,62	3,97
	454	2	1018,86	66,10	983,59	46,13	4,71
	303	3	1064,43	59,99	1015,42	40,79	3,17
	227	4	1047,00	44,77	1021,17	39,7	2,81
	182	5*	1114,19	44,77	1114,19	44,77	2,36
CMT11	1146	1**	1429,11	286,07	1426,16	280,16	20,37
CMT12	902	1	1015,01	113,01	1007,87	105,87	3,45
	451	2	1054,03	83,32	1049,76	75,67	4,25
	301	3	1038,55	53,44	982,75	29,51	2,45
	225	4	1012,49	49,17	998,63	45,81	2,56
F71	266	1**	312,63	46,63	304,84	38,84	4,55
	133	2	322,26	28,61	314,07	24,16	2,16
F134	12979	1	15474,94	2682,94	15085,66	2293,66	188,51
	640	2*	15007,73	1428.80	15007,73	1428,8	15,46
Thai75a	1780	1	1982,87	202,87	1876,06	96,06	0,49
	890	2	2035,67	134,89	1990,81	108	0,36
	593	3**	2097,32	119,66	2086,23	108	0,4
Thai75b	1479	1	1540,21	61,45	1478,29	0	0,92
	740	2	1579,59	52,25	1540,43	32,02	0,46
	493	3	1659,39	66,96	1613,21	53,84	1,015
	370	4**	1679,67	76,07	1658,09	64,73	0,46
Thai75c	1420	1	1568,35	148,35	1473,16	53,16	0,40
	710	2	1686,46	141,26	1651,26	118,7	0,32
	473	3*	1617,17	70,33	1617,17	70,33	0,44
	355	4*	1712,65	80,5	1712,65	80,5	0,53
Thai75d	1502	1	1761,29	259,29	1619,28	117,28	1,14
	751	2	1610,88	65,91	1572,01	35,31	1,19
	501	3	1738,87	90,37	1665,06	64,2	1,20
	375	4**	1806,92	93,37	1761,58	93,03	1,34
Thai100a	2245	1	2513,64	268,64	2487,71	242,71	1,48
	1123	2	2683,37	221,12	2621,6	192,8	1,32
	748	3	2531,46	106,9	2377,98	50,53	2
	561	4	2557,75	106,87	2462,64	75,99	1,85
	449	5**	2653,21	99,13	2626,6	92,09	2,13
Thai100b	2134	1	2522,88	388,88	2443,42	309,42	1,85
	1067	2	2633,28	253,25	2584,29	229,49	1,39
	711	3	2462,98	118,73	2388,63	95,14	1,46
	533	4**	2528,62	106,62	2497,6	96,23	2,24
Thai100c	1547	1	1633,16	86,16	1567,05	20,05	2,656
	773	2	1692,32	80,75	1505,88	0	3,95
	516	3	1820,90	95,44	1717,18	60,27	2,45
Thai100d	1739	1	2102,63	363,63	2017,65	278,65	3,766
	869	2*	2156,86	216,7	2156,86	216,7	3,05
Thai150a	3361	1	3881,88	520,88	3772,71	411,71	5,71
	1680	2	4065,08	360,43	3887,27	274,96	4,343
	1120	3	3904,39	206,4	3819,79	164,81	5,92
	840	4	3956,27	173,6	3934,36	172,36	7,64
Thai150b	3000	1**	3436,23	436,23	3426,57	426,57	13,34
	1500	2	3405,98	212,1	3281,88	155,48	8,92
	1000	3*	3612,9	230,52	3612,9	230,52	8,4
Thai150c	2595	1	3086,49	491,49	3006,83	411,83	12,11
	1297	2*	3190,45	312,39	3190,45	312,39	9,29
Thai150d	2910	1	3416,54	506,54	3323,09	413,09	11,17
	1455	2	3459,54	289,57	3431,4	271,49	5,70

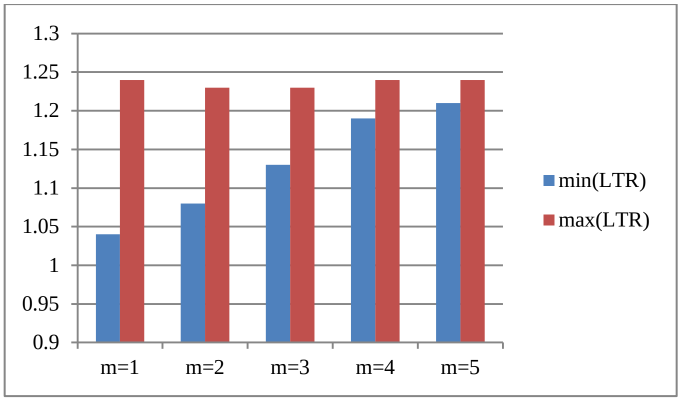


Fig. 6. Max and min of the LTR value depending on m.

The graph of Fig. 6 shows the minimum and maximum value of the LTR according to m.

A note to make is that the more the m increases, the more the minimal LTR increases. However, the maximum value is not significant compared to m. This can be justified by the fact that the maximum value of the LTR for m = 1 was obtained for the CMT11 instance which has no feasible solution for the other value of m. If we exclude this instance, the maximum value becomes 1.20 (Thai100d). Adopting this remark we can conclude that the LTR increases with m. This is also justifiable by the fact that we could not have a feasible solution for the big values of m.

Highlighted results are results where the minimum value of the total distance traveled and the minimum value of the overtime do not match the same solution. We have five instances where there is no such correspondence on 62instances. This gives a percentage of 3,1%. Thus in 96,9% a solution that minimizes the overtime minimizes also the distance. This being stated, we can conclude that the two objectives of our problem are proportional on 96,9%

D. *Practice Use*

For an industrialist what counts from all what is said is having an optimal or near optimal solution in a practical time. Our algorithm was able to give a near optimal solution for the DVRP with overtime in an execution time that does not exceed 6.50 (min, second) on average. To take advantage of these results, we are working on a software project that runs the same algorithm but with adaptable and comfortable interfaces for managers and industrialists while allowing to have results for a static problem in case of need.

VI. Conclusion

To conclude, this article introduces a new variant of the DVRP that is the multi-tours DVRP with overtime. The article gives a mathematical model of the problem with a hybrid ACS resolution. The results of the static and dynamic algorithm are competitive, comparatively to other works from literature. To test our algorithm on the dynamic case with overtime, we have proposed new benchmarks inspired from the very famous ones. Results have shown that the two objectives of the problem are proportional on 96,9%.

Several avenues exist for future works; a future goal is to work with another metaheuristics on the same problem to compare its results with those given by the hybrid ACS. Another research direction includes introducing new constraints such customer time windows, as well as considering stochastic data while providing the first planning of the day.

References

[1] Braekers K., Ramaekers K. & Van Nieuwenhuyse I., "The vehicle routing problem: State of the art classification and reveiw".Computers and Industrial Engineering, 99, 300-313, (2016)

[2] Eksioglu B., Volkan Vural A., & Reisman A., "The vehicle routing problem: A taxonomic review", Computers and Industrial Engineering, 57(4), 1472-1483, (2009)

[3] Crainic T.G., Gendreau M., & Potvin J.-Y., " Intelligent freight-transportation systems : assessment and the contribution of operations research", Transportation Research Part C: Emerging Technologies, 17 (6), 541–557, (2009)

[4] Psaraftis H., "A dynamic-programming solution to the single vehicle many-tomany immediate request dial-a-ride problem", Transportation Science, 14 (2), 130–154, (1980)

[5] Pillac V., Gendreau M., Guéret C., & Medaglia A.,"A review of dynamic vehicle routing problems", European Journal of Operational Research, 225(1), 1-11, (2013)

[6] Ferrucci F., Bock S., &Gendreau M. "A pro-active real-time control approach for dynamic vehicle routing problems dealing with the delivery of urgent goods ". European Journal of Operational Research, 225(1), 130-141, (2013)

[7] Hu, T. "Evaluation Framework for Dynamic Vehicle Routing Strategies Under Real-Time Information". Transportation Research Record, 1774(1), 115-122, (2001)

[8] Tirado G, Hvattum L, Fagerholt K, Cordeau J. " Heuristics for dynamic and stochastic routing in industrial shipping ". Computers & Operations Research, 40(1), 253-263, (2013)

[9] Kilby P., Prosser P.and Shaw P. "Dynamic VRPs:A study of scenarios". APES-06-1998, University of Strathclyde, U.K, (1998)

[10] Taillard, E. D., Laporte, G.and Gendreau, M."Vehicle routing with multiple use of vehicles", Journal of Operationeal Research Society, 47(8), 1065-1070, (1996)

[11] Gendreau M., Guertin F., Potvin J.-Y., Taillard E., "Parallel tabu search for realtime vehicle routing and dispatching ", Transportation Science, 33 (4), 381–390, (1999)

[12] Chang M.S., Chen S., Hsueh C., "Real-time vehicle routing problem with time windows and simultaneous delivery/pickup demands", Journal of the Eastern Asia Society for Transportation Studies, 5, 2273–2286, (2003)

[13] Chen Z., Xu H., "Dynamic column generation for dynamic vehicle routing with time windows", Transportation Science, 40 (1), 74–88, (2006)

[14] Hanshar F.,& Beatrice M., "Dynamic vehicle routing using genetic algorithms", Applied intelligence, 27(1),89-99, (2007)

[15] Hong, L."An improved LNS algorithm for real-time vehicle routing problem with time windows", Computers & Operations Research, 39(2), 151-163, (2012)

[16] Khouadjia M., Sarasola B., Alba E., Jourdan L., and Talbi E."A comparative study between dynamic adapted PSO and VNS for the vehicle routing problem with dynamic requests", Applied Soft Computing, 12(4), 1426-1439, (2012)

[17] Montemanni R., Gambardella L.M., Rizzoli A.E., Donati A.V., "Ant colony system for a dynamic vehicle routing problem", Journal of Combinatorial Optimization,10 (4), 327–343, (2005)

[18] Hvattum L.M.,Lokketangen A.,Laporte G., "Solving a dynamic and stochastic vehicle routing problem with a sample scenario hedging heuristic ", Transportation Science, 40 (4), 421–438, (2006)

[19] Hemert J. and Poutré J. "Dynamic routing problems with fruitful regions: models and evolutionary computation", in Xin Y. Edmund K.

José A. Jim S. Juan J. John A. Jonathan E. Peter Tiňo P. Ata Kabán A. and Hans-Paul S (eds), Parallel Problem Solving from Nature - PPSN VIII, Berlin, Springer, 692-701, (2004)

[20] Schyns M., "An ant colony system for responsive dynamic vehicle routing", European Journal of Operational Research, 245(3),704-718, (2015)

[21] Gendreau M., Guertin F., Potvin J., & Séguin R., "Neighborhood search heuristic for dynamic vehicle dispatching problem with pick-ups and delivries", Transportation Research Part C (14), 157-174, (2006)

[22] Hassin R., Rubinstein S., "On the complexity of the k-customer vehicle routing problem", Operations Research Letters (33), 71-76, (2005)

[23] Ayadi R., Benadada Y.,"Memetic algorithm for a Multi-Objective vehicle routing problem with multiple trips", International Journal of Computer Science and Applications, (IJCSA),(10), 72– 91, (2013)

[24] Gambardella,L.M., Taillard, E. and Agazzi, G."MACS-VRPTW: a multipleant colony system for vehicle routing problems with time windows", editors New Ideas in Optimization, McGraw-Hill, 63-76, (1999)

[25] Ouaddi K., Benadada Y., Mhada F.Z., "Multi period dynamic vehicles routing problem: Literature review, modelization and resolution", The 3rd IEEE International Conference on Logistics Operations Management (GOL'16), DOI: 10.1109/GOL.2016. 7731706, (2016)

[26] Ouaddi K.., Benadada Y., Mhada F.Z., "Ant Colony system approach for dynamic vehicles routing problem multi tours", Proceedings of the 2nd international Conference on Big Data, Cloud and Applications, 109, (2017)

[27] Christofides, N., Mingozzi, A. and Toth, P.,"The vehicle routing problem", Combinatorial Optimization. Wiley, Chichester, 315-338, (1979)

[28] Fisher, M,"Optimal Solution of Vehicle Routing Problems Using Mimimum k-trees". Operations Research, 42(4), 626-646, (1994)

[29] Rochat, Y. and Taillard, E.,"Probabilistic Diversification and Intensification in Local Search for Vehicle Routing". Publication CRT-95–13, (1995)

Multi-Class Breast Cancer Classification using Deep Learning Convolutional Neural Network

Majid Nawaz, Adel A. Sewissy, Taysir Hassan A. Soliman
Faculty of Computer and Information, Assiut University

***Abstract*—Breast cancer continues to be among the leading causes of death for women and much effort has been expended in the form of screening programs for prevention. Given the exponential growth in the number of mammograms collected by these programs, computer-assisted diagnosis has become a necessity. Computer-assisted detection techniques developed to date to improve diagnosis without multiple systematic readings have not resulted in a significant improvement in performance measures. In this context, the use of automatic image processing techniques resulting from deep learning represents a promising avenue for assisting in the diagnosis of breast cancer. In this paper, we present a deep learning approach based on a Convolutional Neural Network (CNN) model for multi-class breast cancer classification. The proposed approach aims to classify the breast tumors in non-just benign or malignant but we predict the subclass of the tumors like Fibroadenoma, Lobular carcinoma, etc. Experimental results on histopathological images using the BreakHis dataset show that the DenseNet CNN model achieved high processing performances with 95.4% of accuracy in the multi-class breast cancer classification task when compared with state-of-the-art models**.

Keywords—Breast cancer classification; Convolutional Neural Network (CNN); deep learning; medical image processing; histopathological images

I. Introduction

Breast cancer is a major public health issue because it is the most common cancer in women and the leading cause of cancer death worldwide. Indeed, nearly one in seven women will be affected by this pathology during its existence, the risk increasing with age [1]. In addition, worldwide studies in 2012 reported 522,000 deaths from breast cancer in the same year, an increase of 14% over 2008 [2]. The development of massive breast cancer screening has led to earlier diagnosis and rapid management with a significant improvement in survival rate. The treatment and analysis of medical images is a rapidly expanding area where the problem of automatically searching for information contained in medical images is urgently needed. Indeed, the great diversity of medical imaging devices, the difficulty of interpretation of these images as well as their large number, generates tedious work for those who must interpret them [3]. In order to process this large volume of information, doctors are currently turning to the use of systems to assist in the analysis and interpretation of these images. This analysis aims to facilitate the diagnosis made by the practitioner and to make it as accurate and reliable as possible [4]. However, and in contrast to advanced technology in the medical sector, breast cancer analysis remains a real public health problem and a very sensitive topical research topic to address. Mammographic imaging is one of the most commonly used modalities [5]. This tool that we are interested in this paper has become an indispensable tool for any clinical examination related to breast cancer. In the field of computational medical imaging, methods of deep convolutional neural networks (CNN) [6] have proved successful for the hierarchical unsupervised learning of imaging features of increasingly complex data directly from raw images, allowing to discover the relevant characteristics, instead of extracting features defined a priori by the user.

A selection of variables can be done in an integrated way with the learning of the characteristics, and this, both on the raw data and on the learned characteristics [6], [7]. Similarly, the supervised classification can also be integrated into the same architecture with the two previous steps to optimize and automate the process [8], [9]. Studies have compared the conventional multi-step computational imaging methods with deep learning methods, and showed a better classification accuracy and mortality prediction with deep learning methods in the case of screening mammograms breast cancer [8]. Deep learning refers to advanced statistical learning methods organized in multiple layers, to extract representations of data on multiple levels, and whose layers are not predefined by the user but learned directly from the data by the algorithm, thus mimicking human neuronal functioning [10].

It has been successfully applied to various pathologies and modalities, including the use of convolutional networks (CNN) that exploit large databases for the extraction of relevant descriptors and segmentation [11]. The main challenge of cancer automatic aided diagnosis systems is dealing with the inherent complexity of histopathological images. To deal with this, we choose to use a powerful convolutional neural network of multiclass classification problem. We propose to use the DenseNet model [12], one of state of the art in the image recognition competition ImageNet [13]. The DenseNet is built for natural images processing but we modified it to deal with histopathology images for breast cancer classification using transfer learning.

The results obtained using transfer learning on the proposed custom model surpasses the current best performance, for all of the resolutions in the benchmark dataset. The remainder of this paper is divided into four sections. After introducing, related works on breast cancer classification are reviewed in Section 2. Section 3 presents the proposed CNN model for multi-class breast cancer classification. Experiments, results and comparison with popular CNNs models are detailed in Section 4. Finally, this paper is concluded in Section 5.

II. Related Works

Research related to the detection of breast cancer has increased during the last decade. Much work has been directed towards the detection of the presence of cancerous tissue in the breast and the classification of tumors. Some researchers have preferred to design aided diagnosis systems based on Content based image retrieval techniques that would have the advantage of offering radiologists images available in a medical image database, whose content is known and which would be similar to image request for which the radiologist would have doubts. However, this approach also raises problems of search time and adequate similarity measurement between the request image and those contained in the database. For example, Tourassi et al. [14] proposed a content search system for tumors detection that makes use of the expert knowledge present in the different mammograms that make up the image database. To achieve this, they first use the matching template to find among the images in the database those that are similar to the ROI request presented by the user of the system. In order to determine whether the query ROI contains a tumor (of any kind) or only healthy tissue, they proposed a decision measure that effectively combines several similarity measures on the best matches. For their part, Alto et al. [15] preferred extracting descriptors of texture, shape and sharpness of the edge. Those related to pixel intensity, shape and texture were merged by Tao et al. [16], in order to find the tumors similar to that contained in the ROI query and classify it as benign or malignant. For better visual similarity, Zheng et al. [17] proposed a system that provides further interaction with the user; system in which, the latter is asked to evaluate the nature of the spiked tumor of the query image so that the system only looks for matches with similar degrees of speculation. This work was subsequently improved by removing from the search base the ROIs that gave the worst similarity scores [18].

Moreover, several works have tried to find in the image databases, mammary tumors with similar characteristics such as shape, contour and pathology. For example, the shape, intensity and texture descriptors were combined using a suitable weighting system, also exploiting the user interaction to optimize the quality of the proposed matches [19]. Narvaez et al. [20] have proposed a method that begins by merging the shape and texture descriptors extracted on the two incidences of the breast to find the best matches, which images are then used to annotate the query ROI. Liu et al. [21] have on their side introduced an image search based on a hash function to produce a diagnosis for the tumors contained in the ROIs queries. More specifically, a hash function inspired by graph theory and named anchor [22] was used to compress two descriptors, namely the SIFT histogram and the GIST into binary codes; finally the search for similarity was made in the Hamming space. Moayedi et al. [23] developed an automatic classification of tumors in mammograms. They exploited three approaches for texture determination, namely texture analysis based on contour let transforms, as well as geometric characteristics representing the orientation, the zone and the center of the tumor, and the statistical descriptors obtained from the co-occurrence matrix using successive improvement learning (SEL) and weighted SVM, vector support based fuzzy neural networks (SVFNN) and kernel-based SVM classifier for localization as well as classification of tumors into malignant and benign. In [24], the authors carried out a quantitative approach for the classification of tumors of mammography based on the descriptors of the texture. Indeed, they extracted a set of texture descriptors on 130 mammograms, under different configurations and scales. In addition, multivariate analysis of variance (MANOVA) was applied to the construction of additional subsets of statistically independent texture descriptors. Thus, a texture signature is attributed to both malignant and benign tumors. The authors used linear and nonlinear classifiers for the classification stage, consisting of Linear Discriminant Analysis (LDA), Least Square Minimum Distance (LSMD), K nearest neighbors (k-NN), the function Radial Basis (RBF) and Multilayer Perceptron (MLP), Artificial Neural Networks (ANN), as well as Support Vector Machines (SVM). The authors asserted that texture descriptors extracted at large scales are richer in content than texture descriptors extracted at small scales. They achieved a tumor classification rate of 83.9% using the Support Vector Machines (SVM). Guo et al. [25] proposed a Multilayer Perceptron (PMC) as a classifier for diagnosing breast cancer. As a first step, a variable selection step is performed on the data using Genetic Programming (GP). The variables are then assigned to the entry of the PMC to evaluate the classification performance. The neural network converged with an average classification rate of 96.21%. The results obtained by the authors put forward the ability of the GP method to transform the information by reducing the dimensionality of the variable space, and to define the relationship between the data in an automatic way. These last two properties help to improve the accuracy of the classification. The recognition rates obtained are interesting. That said, the classifier was of the binary type with a class of malignant cases and a class of benign cases.

The works done in the literature [26], on deep learning through Convolutional Neural Networks among others, has opened the way to an automatic representation approach based on a non-supervised descriptors extraction, i.e. independent of any human intervention that could affect these performances. Deep CNNs often have a number of parameters so large that it cannot reasonably be trained without a very large dataset. Medical imaging datasets are often not sufficiently large to train a deep CNN model from scratch adequately. Thus, the usage of transfer learning in medical imaging has been explored. Transfer learning aims to transfer knowledge between large source and small target domains [27]. For CNNs, this is often done by pre-training a CNN model with the source dataset, then re-training parts of the model with the target dataset. In [28], the authors adapted the popular CNN AlexNet to classify breast cancer tumors from histopathological images on BreakHis dataset [29]. They proposed a sliding widow mechanism to extract random patches for the training strategy. They reached an average classification rate of 79.85%. Han et al. [30] proposed a breast cancer multi-classification framework using class structure-based deep convolutional neural network model (CSDCNN). It has particular feature learning manner using prior knowledge of class structure on histopathological images. The structured deep learning model has reached remarkable performance with 93.2% of average accuracy on

BreakHis dataset. Nuh et al. [31], distinguished cell and non-cell samples in breast tumors images using Convolutional Neural Networks with different spatial patches. The classification accuracy was estimated at 86.91% and 86.17% for 5x5 and 7x7 sub-window sizes respectively. In [32], the authors presented CNN classifier for visual analysis of invasive ductal carcinoma tissue regions in malignant breast tumors images. The proposed framework yielded higher performance compared to random forest classifier with 84.23% detection accuracy. Hafemann et al. [33] have shown, for histopathological images, that Convolutional Neural Network outperforms traditional textural descriptors. Besides, the traditional approach to extract appropriate features for classification tasks in pathological images requires considerable efforts and effective expert domain knowledge, frequently leading to highly customized solutions, specific for each problem and hardly applicable in other contexts [34].

III. Proposed CNN Model for Multi-Class Breast Cancer Classification

A Convolutional Neural Network (CNN) is feedforward neural network introduced by Kunihiko Fukushima in 1980 [30] and improved by Yann LeCan et al. in 1998 [35], [36]. A CNN is composed of 6 types of layers: an input layers, a convolutional layer, a non-linear layer, a pooling layer, fully connected layer, and an output layers. Fig. 1 illustrates a traditional CNN architecture.

Convolutional Neural Networks (CNN) are one of the most remarkable approaches of deep learning, in which multiple layers of neurons are formed in a robust manner. They have shown that they are capable of demonstrating an impressive generalization capability on large data sets with millions of images [37], [38]. These results come mainly from the particular architecture of CNNs that takes into account the specific topology of tasks related to the field of computer vision that exploit two-dimensional images. Other dimensions can also be taken into account when it comes to color images with multiple channels.

To train a CNN we determine the mapping function using the feedforward operation and we optimize the loss function using retro propagation techniques in particular, the gradient decent algorithm. The CNN that we choose for the task of breast cancer classification is not a traditional CNN model. DenseNet [12] is a CNN model which they replace convolution non-linear and pooling layers with dense blocks and transition layers using the original CNN layers except the first convolutional layer. Fig. 2 presents the original DenseNet model with three dense blocks and two transition layers.

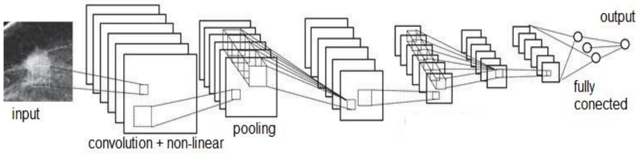


Fig. 1. Convolutional neural network.

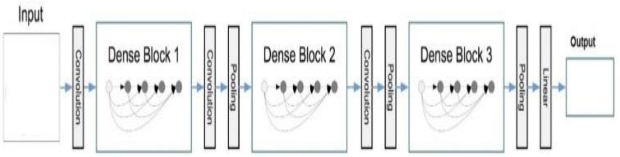


Fig. 2. DenseNet with three dense blocks architectures [12].

The dense block proposed by DenseNet contains convolution and non-linear layers. Also, they apply some optimization techniques like dropout and batch normalization. In addition, in the dense block proposed by DenseNet, outputs from the previous layers are concatenated instead of using the summation. So, assume that an input image has the shape of (28, 28, 3), in which three represents the RGB color space. First, we spread image to initial N channels and receive the image (28, 28, N). Every next convolution layer will generate k features, and remain the same height and width. The feature concatenation process is illustrated by the Fig. 3. If we assume that we have N= 24 and K= 12 we will receive the image with same dimension, but with plenty of features (28, 28, 48).

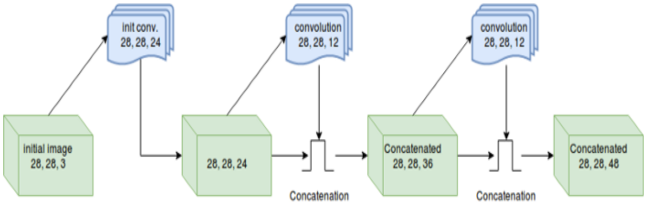


Fig. 3. DenseNet concatenation process inside the dense block.

To reduce the size, DenseNet uses transition layers. These layers contain convolution with kernel size = 1 followed by 2x2 average pooling with stride = 2. It reduces height and width dimensions but leaves feature dimension the same. The transition layer is presented in Fig. 4. As a result, if the input is an image with shape (28, 28, 48), we receive an output image with shapes (14, 14, 48).The DenseNet scale naturally to hundreds of layers, while exhibiting no optimization difficulties. Thus, that makes DenseNet one of the most powerful models in image recognition tasks.

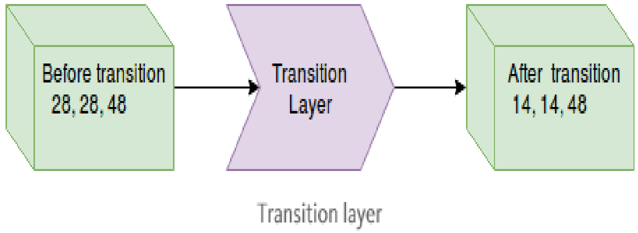


Fig. 4. Transition layer.

As mentioned above we will modify the DenseNet model to deal with histopathology images to build a breast cancer classifier using transfer learning. Our custom-made model is inspired by DenseNet, and contains four dense blocks and three transition layers to classify breast cancer tumors. The proposed CNN model is presented in Fig. 5. The DenseNet is built to deal with natural image and non-microscopic images. To solve this, we use kernel of 7x7 sizes for the first convolutional layer to detect small variation and substance in the image and extract more important features. Also, the size kernel of the convolutional layers in dense blocks is reduced to deal with the complex structure of the histopathology images. An average pooling layer with a 7x7 kernel size and stride 2 is used before the fully connected layer to fix the feature map connected to this layer. In addition, we configure the softmax layer for the eight classes of BC histopathological images instead of the 1000 classes of the ImageNet dataset [13].

Transfer learning is defined as fine-tuning CNN models pre-trained from natural image dataset to medical image tasks. Learning from clinical images from scratch is often not the most practical strategy due to its computational cost, convergence problem, and insufficient number of high quality labeled samples. A growing body of experiments has investigated pre-trained models in the presence of limited learning samples. We initialized weights of different layers of our proposed network by using pre-trained model on ImageNet. Then, we employed last layer fine-tuning on BreakHis cancer images dataset. Therefore, the ImageNet pre-trained weights were preserved while the last fully connected layer was updated continuously. The first convolutional layer of the network is then un-frozen, and the entire network is fine-tuned on the BreakHis training data. The advantage of the DenseNet is feature concatenation that helps us to learn the features in any stage without the need to compress them and the ability to control and manipulate that features. This technique helps us to avoid the parameter number explosion so we reduce the training process complexity and eliminate the over fitting problem.

IV. Experiments and Results

In our experiments we use the BreakHis dataset [29] for training and testing. It contains 7909 microscopic biopsy images of benign and malignant breast tumors. Images are acquired, by a Microscope System coupled with Digital Color Camera, in RGB True Color Space using four magnifying factors: 40X, 100X, 200X, and 400X. The image distribution is summarized in Table I.

TABLE I. Image Distribution in the BreakHis Dataset

Magnification	Benign	Malignant	Total
40X	652	1370	1995
100X	644	1437	2081
200X	623	1390	2013
400X	588	1232	1820
Total of Images	2480	5429	7909

Both benign and malignant breast tumors in BreakHis dataset are sorted into four distinct subtypes. Lobular carcinoma (LC), Ductal carcinoma (DC), Papillary carcinoma (PC) and Mucinous carcinoma (MC) are the types of malignant breast tumors. For the benign tumors, the types are Fibroadenoma (F), Adenosis (A), Tubular adenoma (TA) and Phyllodes tumor (PT). Fig. 6 shows examples of the breast cancer subclasses.

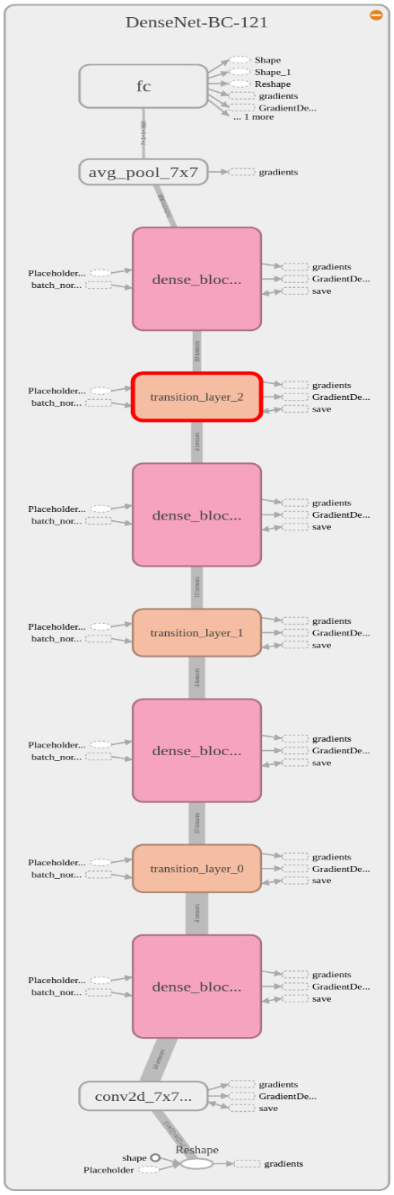


Fig. 5. DenseNet model from the Tensorboard tool of Tensorflow.

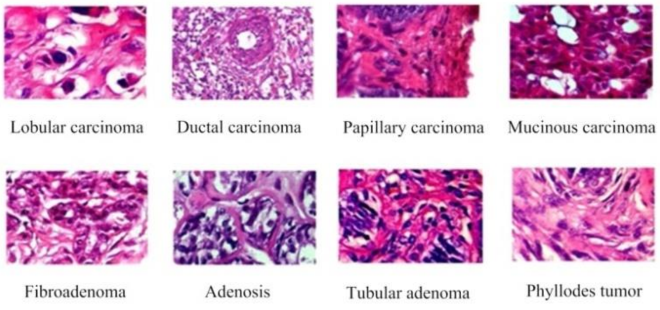


Fig. 6. Examples of histopathological images of breast cancer subclasses from BreaKHis dataset.

To develop the model, we use the Tensorflow deep learning framework [39] and the Nvidia digits tools [40]. The model is trained and tested using the MSI Pro Series desktop equipped with an Intel i7 processor and an Nvidia Geforce GTX960 GPU.

Following the experimental protocol proposed in [29], the dataset is divided into 70% for training set and 30% for validation set. When discussing medical images, there are two ways to report the results. In the first one the decision is patient-wise, therefore, the recognition rate is computed at the patient level. Let N_p be the number of histopathological images of patient P. For each patient, if N cancer images are correctly classified, the patient score and the global patient recognition rate are defined as in equation 1 and equation 2 respectively.

In the second case the recognition rate is computed at the image level. Let N_t be the number of histopathological images of the testing set. If N_r cancer images are correctly classified, then the recognition rate at the image level is represented in (3).

$$\text{Patient score} = \frac{N}{N_p} \qquad (1)$$

$$\text{Patient recognition rate} = \frac{\sum \text{patient score}}{\text{total patient number}} \qquad (2)$$

$$\text{Image recognition rate} = \frac{N_r}{N_t} \qquad (3)$$

A. *Training*

The proposed CNN model aims to treat the high resolution images generally used for the histopathological classification of breast cancer. The DenseNet model is modified to extract fully global feature from the histological images and use them in the training process using transfer learning. In this case we resize all the images to 224x224x3 RGB color space.

After obtaining the weights of the model pre-trained on ImageNet, transfer learning is done in the following steps. First, the fully connected layer has randomly initialized weights. We freeze the convolutional layers of the network, and only train the fully connected layer using the BreakHis training dataset. The fully connected layer is trained from scratch on the features extracted from the fixed convolutional layers. The first convolutional layer of the network is then un-frozen, and the entire network is fine-tuned on the BreakHis training data. This involves re-training the CNN, starting from the retained weights, and using a very small step size.

To train the model we use the Adam optimizer to minimize the loss function [41]. Adam optimizer is a gradient descent algorithm with an adaptive momentum that computes adaptive learning rates for each parameter [42]. Fig. 7 represents the total loss minimization during the training process. The training process took 11 hours and the total loss achieves a minimum of 0.3424.

B. *Testing*

After training our model we use the dataset reserved for validation to test the model. Table II reports the accuracy of our model with the different magnification factors of the BreakHis dataset in both image level and patient level. Our model shows best performance with high multi-classification accuracy. It achieves, respectively, 95.4% of average accuracy of the image level and 96.48% accuracy of the patient level for all magnification factors.

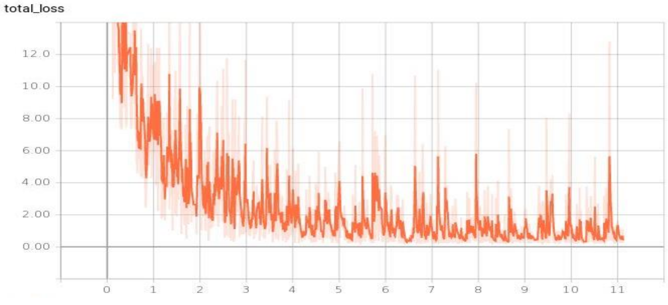


Fig. 7. Total loss minimization.

TABLE II. THE MODEL MULTI-CLASSIFICATION ACCURACY WITH THE DIFFERENT MAGNIFICATION FACTORS OF THE BREAKHIS DATASET

Magnification factors	**40x**	**100x**	**200x**	**400x**	**Average**
Image level Accuracy (%)	93.64	97.42	95.87	94.67	95.4
Patient level Accuracy (%)	94.23	97.86	96.35	95.24	96.48

To provide a proper performance evaluation, we compare the results of the proposed model with the most powerful CNNs in the histopathological breast cancer images multi-classification. The performance comparison with state of the art models in Table III confirms that our model reached the highest multi-classification accuracy. AlexNet [26], the state of the art in the visual image recognition competition ImageNet (ILSVRC12), yielded 83% of detection accuracy in the histopathological images binary classification (benign and malignant) [28]. However, for the multi-classification task, it achieves about 80 % of accuracy.

TABLE III. COMPARISON WITH SOME POPULAR CNNS IN THE MULTI-CLASS BREAST CANCER CLASSIFICATION

Accuracy (%)	**Model**	**Magnification Factors**				
		40x	***100x***	***200x***	***400x***	***average***
Image level	***LeNet [43]***	46.4	47.34	46.5	45.2	46.36
	AlexNet [28]	86.4	75.8	72.6	84.6	79.85
	CSDCNN[30]	92.8	93.9	93.4	92.9	93.25
	DenseNet (our's)	93.64	97.42	95.87	94.67	95.4
Patient level	***LeNet [43]***	48.2	47.6	45.5	45.2	46.62
	AlexNet [28]	74.6	73.8	76.4	79.2	76
	CSDCNN[30]	94.1	93.2	94.7	93.5	93.87
	DenseNet (our's)	94.23	97.86	96.35	95.24	96.48

The CSDCNN [30] is convolutional neural network proposed by Zhongyi Han et al. for breast tumors detection and multi-classification; it achieves about 94% accuracy. LeNet [43] is a traditional CNN used for the handwritten character recognition and achieving remarkable accuracy. However, on the histopathological images, its performance was considerably inferior, achieving about 47% multi-classification accuracy [28].

Compared to the mentioned powerful CNN models, our proposed model achieved the highest multi-classification accuracy with about 96% of average accuracy of the image level. Histopathology tumor detection and classification into multi-classes would play a key role in breast cancer diagnosis, reduce the heavy workloads of pathologists and establish the appropriate therapeutic approach by doctors.

V. Conclusion

In the context of classification, deep convolutional neural networks (CNNs) have been widely proven in the scientific and industrial community. In this work, we investigated the performance of a deep neural network model on a classification task related to breast cancer detection. The modification applied to the DenseNet model proves that deep learning model used in natural images processing can achieves high performance in medical images processing. In our case we achieve about 96% of accuracy in the multi-class breast cancer classification task and that outperform human expert in the diagnostic domain. The performance achieved can be improved if we provide more data using larger datasets.

References

[1] R. L. Siegel, K. D. Miller, and A. Jemal, “Cancer statistics, 2015,” CA: a cancer journal for clinicians, vol. 65, no. 1, pp. 5–29, 2015.

[2] IARC (2013). Globocan 2012: estimatied cancer incidence, mortality and prevalence worldwide in 2012.

[3] Y. Zheng, “De-enhancing the dynamic contrast-enhanced breast MRI for robust registration”. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pp:933-941 (Springer, 2007).

[4] Y. ZHENG, W. Benzheng, L. Hui, “Measuring sparse temporal-variation for accurate registration of dynamic contrast-enhanced breast MR images”. Computerized Medical Imaging and Graphics, 2015, vol. 46, p. 73-80.

[5] S. R. Lakhani (ed.). “WHO Classification of Tumours of the Breast”. International Agency for Research on Cancer, 2012.

[6] Q. Zhang, Y. Xiao, W. Dai, J. Suo, C. Wang, J. Shi, “Deep learning based classification of breast tumors with shear-wave elastography”. Ultrasonics 2016;72:150–7.doi:10.1016/j.ultras.2016.08.004.

[7] K. Sohn, G. Zhou, C. Lee, H. Lee, “Learning and Selecting Features Jointly with Pointwise Gated Boltzmann Machines.”, 2013, p. 217–25.

[8] J. Cheng, D. Ni, Y. Chou, J. Qin, C. Tiu, Y. Chang, “Computer-Aided Diagnosis with Deep Learning Architecture: Applications to Breast Lesions in US Images and Pulmonary Nodules in CT Scans”. Sci Rep 2016, 6:24454. doi:10.1038/srep24454.

[9] G. Carneiro, L. Oakden-Rayner, A. Bradley, J. Nascimento, L. Palmer, “Automated 5-year Mortality Prediction using Deep Learning and Radiomics Features”, Chest ComputedTomography 2016.

[10] Y. LeCun, Y. Bengio, G. Hinton, “Deep learning”. Nature 2015;521:436–44. doi:10.1038/nature14539.

[11] O. Ronneberger, P. Fischer, and T. Brox, “U-net : Convolutional networks for biomedical image segmentation,” in International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 234-241, 2015.

[12] G. Huang, Z. Liu, K. Q. Weinberger and L. Van der Maaten, (2017, July). Densely connected convolutional networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (Vol. 1, No. 2, p. 3).

[13] O. Russakovsky, D. Jia, S. Hao, Imagenet large scale visual recognition challenge. International Journal of Computer Vision, 2015, vol. 115, no 3, p. 211-252.

[14] G. D. Tourassi, R. Vargas-Voracek, D.M. Catarious Jr and C. E. Floyd, Computer-assisted detection of mammographic masses: A template matching scheme based on mutual information. Medical Physics, Vol 30, No.8, pp: 2123–2130, 2003.

[15] H. Alto, R. M. Rangayyan and J. L. Desautels, Content-based retrieval and analysis of mammographic masses. Journal of Electronic Imaging, Vol 14, No.2,:023016– 023016, 2005.

[16] Y. Tao, S. Lo, M.T. Freedman and J. Xuan, A preliminary study of content-based mammographic masses retrieval. In Medical Imaging, 65141Z. International Society for Optics and Photonics, 2007.

[17] B. Zheng, A. Lu, L. A. Hardesty, J. H. Sumkin, C. M. Hakim, M. A. Ganott and D. A. Gur, A method to improve visual similarity of breast masses for an interactive computer-aided diagnosis environment. Medical Physics, Vol 33, No.1, pp: 111–117, 2006.

[18] S. Park, R. Sukthankar, L. Mummert, M. Satyanarayanan and B. Zheng. Optimization of reference library used in content-based medical image retrieval scheme. Medical Physics, Vol 34, No.11, pp: 4331–4339, 2007.

[19] C. Wei, Y. Li and P. Huang, Mammogram retrieval through machine learning within bi-rads standards. Journal of Biomedical Informatics, Vol 44, No.4, pp: 607–614, 2011.

[20] F. Narvaez, G. Diaz and E. Romero, Multi-view information fusion for automatic bi-rads description of mammographic masses. In SPIE Medical Imaging, 79630A. International Society for Optics and Photonics, 2011.

[21] J. Liu, S. Zhang, W. Liu, X. Zhang and D. Metaxas, Scalable mammogram retrieval using anchor graph hashing. In 2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI), pp; 898–901, 2014.

[22] W. Liu, J. Wang, S. Kumar and S. Chang, Hashing with graphs. In Proceedings of the 28th international conference on machine learning (ICML-11), pp; 1–8, 2011.

[23] F. Moayedi, Z. Azimifar, R. Boostani, S. Katebi, Contourlet-based mammography mass classification using the SVM family. Computers in Biology and Medicine, Vol 40. pp. 373–383, 2010.

[24] F. Li, J. Wang, B. Tang, D. Tian, Life grade recognition method based on supervised uncorrelated orthogonal locality preserving projection and K-nearest neighbor classifier. Neurocomputing, Vol 138. pp. 271–282, 2014.

[25] G. Hong and N. Asoke, Breast cancer diagnosis using genetic programming generated feature. Pattern Recognition, 2006, vol. 39, no 5, p. 980-987.

[26] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks”. In Advances in neural information processing systems, 1097–1105 (2012).

[27] Z. Weiming, W. Henry, Y. Fung, C. Zhenghao, Z. Seid Miad, L. Zhicheng, and C. Yuk Ying, Using Transfer Learning with Convolutional Neural Networks to Diagnose Breast Cancer from Histopathological Images. ICONIP 2017, Part IV, LNCS 10637, pp. 669–676, 2017. https://doi.org/10.1007/978-3-319-70093-9_71.

[28] F.A. Spanhol, L.S. Oliveira, C. Petitjean, C. and L. Heutte, “Breast cancer histopathological image classification using convolutional neural networks”. In International Joint Conference on Neural Networks (2016), p. 2560-2567.

[29] F.A. Spanhol, A. Fabio., L.S Oliveura., C. Ptitjean. A dataset for breast cancer histopathological image classification. IEEE Transactions on Biomedical Engineering, 2016, vol. 63, no 7, p. 1455-1462.

[30] Z. Han, B. WEI, Y. ZHENG, “Breast cancer multi-classification from histopathological images with structured deep learning model”. Scientific reports, 2017, vol. 7, no 1, p. 4172.

[31] N. Hatipoglu and B. Gokhan, Classification of histopathological images using convolutional neural network. In : Image Processing Theory, Tools and Applications (IPTA), 2014 4th International Conference on. IEEE, 2014. p. 1-6.

[32] R. Cruz , A. Basavanhally, A. Gonzalez et al. Automatic detection of invasive ductal carcinoma in whole slide images with convolutional neural networks. In : Medical Imaging 2014: Digital Pathology. International Society for Optics and Photonics, 2014. p. 904103.

[33] L. G. Hafemann, L. S. Oliveira, and P. Cavalin, “Forest species recognition using deep convolutional neural networks”. In International Conference on Pattern Recognition, 2014, pp. 1103–1107.

[34] T. H. Vu, H. S. Mousavi, V. Monga, U. A. Rao, and G. Rao, “Dfdl: Discriminative feature-oriented dictionary learning for histopathological

image classification," in Proceedings of the IEEE 12th International Symposium on Biomedical Imaging (ISBI). IEEE, Apr. 2015, pp. 990–994.

[35] Y. LeCun, B.Boser, J.S. Denker, D. Henderson, R.E. Howard, W. Hubbard, and L.D. Jackel. "Backpropagation applied to handwritten zip code recognition". Neural Computation, 1989.

[36] Y. LeCun, L. Bottou, Y. Bengio, and P. Haner. "Gradient-based learning applied to document recognition". Proceedings of the IEEE,86:2278–2324, 1998

[37] K. Simonyan and A. Zisserman, Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.

[38] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. Imagenet large scale visual recognition challenge. International Journal of Computer Vision, 115(3), pp:211–252, 2015.

[39] Tensorflow a deep learning framework https://www.tensorflow.org

[40] Nvidia digits system for deep learning model implementation on Nvidia GPU https://developer.nvidia.com/digits

[41] D. Kingma, and J. L. Ba, (2015). "Adam: a Method for Stochastic Optimization". International Conference on Learning Representations, 1–13.

[42] I. Sutskever, J. Martens, G.E. Dahl and G.E. Hinton, (2013). On the importance of initialization and momentum in deep learning. ICML (3), 28, p:1139-1147.

[43] Y. LeCun, "Comparison of learning algorithms for handwritten digit recognition". In International conference on artificial neural networks. vol. 60, 53–60 (1995).

A Comparative Study of Engineering Students Pedagogical Progress

Khalid Mahboob
Dept. of Computer / Software Engineering
Sir Syed University of Engineering & Technology
Karachi, Pakistan

Danish Ur Rehman Khan
Dept. of Electronics Engineering
N.E.D University of Engineering & Technology
Karachi, Pakistan

Syed Abbas Ali
Dept. of Computer & Information System Engineering
N.E.D University of Engineering & Technology
Karachi, Pakistan

Fayyaz Ali
Dept. of Software Engineering
Sir Syed University of Engineering & Technology
Karachi, Pakistan

***Abstract*—Students' pedagogical progress plays a pivotal role in any educational institute in order to pursue imperative education. Educational institutes, Universities, Colleges implement various performance measures in order to keep analyzing and tracking progress of students to cultivate benefits of education in a better way. There are several data mining techniques to apply on education in order to build constructive educational strategies and solutions. This study aims to analyze and track engineering under graduate student's records to judge quality education, student motivation towards learning, and student pedagogical progress to maintain education at high quality level and predicting engineering student's forthcoming progress. Different engineering discipline students' (of three different cohorts) data have been analyzed for tracing current as well as future pedagogical progress based on their sessional (pre-examination) marks. In this research, the classification techniques by k-nearest neighbor, Naïve Bayes and decision trees are applied to evaluate different engineering technologies student's performance and also there are different methodologies that can be used for data classification.**

Keywords*—*Pedagogical progress; classification; k-nearest neighbor; Naïve Bayes; decision trees; engineering students

I. Introduction

Following higher education is a challenging stage for students as well as educational institutes to deal with huge amount of data. There are various applications used in educational environment in the form of archives, images, blogs, audios, videos, artifacts, scientific documents, meta-data or online hyperlinks or in various other formats [1]. Educational institutes with variety of educational data like student attendance records, examination records, fees records, personal information, etc., entails to be managed and tracked time to time. Therefore, data mining techniques have been used to discern and extract certain patterns that are potentially expedient in the domain of education at any level. Thus, educational data mining can be regard as an interdisciplinary field that assists methods of extracting useful information from enormous sets of data [2]. The advancements in the field of educational data mining have made it possible to perform academic analysis in an innovative ways to focus on educational institutes' efficacy and to reduce student retention [3].

For the successful adoption of educational data mining, it is very necessary to have wide-ranging of pedagogical data so that the various data mining techniques can be applied on that to enhance learning process after analyzing students' academic data to monitor pedagogical progress with improvements, to predict and improve retention of student at an early stage, and to analyze the chances of failures or mistakes by the students in a learning system [4]. Predicting and enhancing students learning has become more challenging in order to improve student's grades and thus benefit educational institutes in adaptation of different learning strategies [5].

Educational data mining focuses on quantitative and qualitative analysis of data necessary to use many techniques based upon multiple disciplines like machine learning, artificial intelligence, expert system, pattern matching, decision support system and it involves the uses of these techniques in an effective ways like if a student/learner may intend to improve learning skills by acquiring e-learning method. Similarly, if a teacher/instructor may require identifying students learning performance by analyzing student academic or prelim records [3]. Any educational institute or university may uses data mining methodologies to determine that how the results of students can be enhanced along with taking attention to reduce student retentions [3]. There have been already many data mining methods applied in the field of education at multiple levels. So many parameters to analyze and track student pedagogical progress have been undertaken using data mining techniques at various levels: at a training system level to predict whether selected or specific knowledge/skills are mastered, at a course level or degree level to predict whether a student will be capable to pass a course or a degree, or to predict their marks [5]. The present study aims to initially identify and investigate the students score based upon pre-final or prelim marks throughout the academic session and finding ways to improve student performance.

Educational data mining deals with several data classification techniques such as Decision Trees, Naïve Bayes, K-Nearest Neighbor, Neural Networks, Support Vector Machines, Quadratic classifiers, and many others [6], [7]. The information derived from applying these techniques can be used for tracking and monitoring student's pedagogical progress in various disciplines to probe students' performance in different courses [6]. This paper is planned as follows: Section II explains the related works. Section III briefly describes the different data mining classification techniques. Section IV defines the data preprocessing and methodology. The following section i.e. Section V presents the results and discussion used in this case study. The conclusion elaborates on useful findings, discourses them and presents future works.

II. Related Works

Much of the work has been done on educational data mining applying different techniques on different levels of education using different tools. This section contributes on the related work in the domain of education using data mining techniques:

Diego Buenaño Fernández et al. [1] have compared three open source tools (Weka, Rapid Miner, and Knime) on different academic records of students. Analysis based on three engineering programs (Network and Telecommunications Engineering, Computer and Information Systems Engineering and Electronics and Information Networks Engineering.) of a University. Most of the relevant tasks were consisted on the data pre-processing. Particularly, the K-Means algorithm is deployed on different attributes of academic data with the applicability of four specific algorithms: ChiSquaredAttributeEval, FilteredSubsetEval, GrainRatio-AttributeEval and OneRAttributeEval. The result shows that the three tools were very similar in working in terms of precision. Results obtained after detailed cluster analysis that can help teachers or instructors to guide students in a course with suitable measures on time.

Raheela Asif et al. [4] presented a case study based on predicting student academic performance at the end of university degree of the degree program at early stage which can help universities to emphasis on skilled students and to initially detect with low educational accomplishment and find effective ways to support them. The four academic cohorts' data comprising 347 undergraduate students have been extracted by using different classifiers. Artificial neural networks, decision trees, k-nearest Neighbor, naive Bayes, and rule induction classification techniques have been used in a study. The study has shown the possibility of predicting fourth year graduation performance at university with pre-university marks and initial university two year marks. The accuracy on analyzing the datasets was satisfactory. Further five courses were evaluated to predict the performance of graduation that did not lead to a better accuracy.

Nawal Ali Yassein et al. [5] have proposed patterns in the available student and courses records to predict students' performance. Statistical package for social sciences (SPSS) and data mining tool (clementine) were used for experimentation purpose. The research comprises of two parts: first to find out the factors related to course success rate, second to determine predictors based on student performance. Both classification and clustering techniques were used to analyze different features that can affect student performance in a course(s).

Brijesh Kumar Baradwaj et al. [6] the classification – decision trees induction method is utilized to evaluate students' academic performance at the end of semester examination to identify early dropouts and to find out students' who might need attention and counseling analyzing diverse academic parameters like student attendance, assignments, class tests, seminar, etc.

T.Archana et al. [8] surveyed on focusing student's performance prediction by improving performance and by increasing student retention in order to increase the quality of education and can be valid in different environments.

Kalpesh Adhatrao et al. [9] have rendered a system to predict the performance of students based on their preceding performances under the concept of data mining classification. ID3 (Iterative Dichotomiser 3) and C4.5 classification algorithms have been applied to predict the performance of fresh students generally and individually.

Ashish Dutt et al. [10] provided the applicability and usability of a clustering method in the context of educational data mining consisting over three decades systematic literature review. The crucial benefit of clustering algorithm towards data analysis is that it provides relatively an explicit schema of learning ways of students specified a number of variables like completing learning tasks on time, groups learning, class learner behavior, classroom decoration and student learning motivation. Clustering can provide relevant intuitions to variables that are applicable in splitting the clusters.

Raheela Asif et al. [11] have predicted student performance using different data mining techniques based on pre-university marks and examination marks of initial years at university. Two cohorts of Civil Engineering technology data were analyzed and various classifiers were applied on that data with a reasonable accuracy. The decision trees were used as an indicator to detect the courses with low performance in order to give warning to students earlier in the degree program.

Raheela Asif et al. [12] performance of students' progress has been investigated by analyzing the data of two immediate cohorts and applying k-means clustering. Each student was characterized by 4-tuple with his/her average remain the same, or either increase or decrease when comparing to their preceding years. Different ranges of accuracies obtained by using different classifiers.

Raheela Asif et al. [13] two aspects of under-graduate students have been studied were: firstly to predict the students' academic accomplishment of four year study programme at the end and secondly analyzing typical progresses and combining them with the results of predictions. The results drawn were the possibility of prediction of graduation performance using pre-university and initial two years university marks only with a reasonable accurateness. Few courses were also put into focus as an indicator regarding good or bad performances with respect to low, intermediate and high marks.

Surjeet Kumar Yadav et al. [14] Decision tree algorithms have been applied on students' previous academic data to produce a model that can be helpful to predict the students' academic performance in order to detect early drop outs of students. CART algorithm among others classification algorithms disclosed the best results for data classification.

Raheela Asif et al. [15] investigated the academic performance of students by applying X-means clustering technique analyzing the data of two immediate cohorts. Shifting of marks from high to low or low to high (or vice versa) in different academic years for both cohorts has been observed. It has been reported that there is a possibility of using one cohort in order to predict the performance of the succeeding cohort with varying accuracies using different classifiers.

III. Data Mining Approaches and Techniques

Data mining is a computational study of data processing which has been successfully functional in many areas that intend to extract useful knowledge from that data [5]. There are various techniques of data mining that are operate-able on large volumes of data to find out hidden patterns and their relationships helping in decision making for different applications such as Artificial Intelligence, Business Management, Decision Support, Machine Learning, Market Analysis, and Statistical and Database Systems [6], [9]. Likewise, several data mining algorithms and techniques are used in knowledge discovery from large databases such as Association Rules, Classification, Clustering, Decision Trees, Genetic Algorithm, Nearest Neighbor methods, Regression, etc. [6].

Among others, classification is a data mining technique, particularly, which plots data into predefined classes or groups [5], [9]. Classification is specifically used for predicting the unknown class label of data objects [16]. This is considered as the most commonly applied technique [6]. This approach often employs Classification (IF-THEN) Rules, Decision Trees or .Neural Networks [2]. In classification, the accuracy of the classification rules are estimated by using training data sets [6]. The classification can works on different training data sets by constructing a model or classifier. Building a classifier or model is the initial step in the learning phase. The classification algorithms are used in building the classifier with the set of parameters essential for proper discrimination [2], [6].

There are many classifiers to implement data mining methods in order to perform in a better way can also apply in the education domain [4]. There are some classifiers which outperforms better than others. Here is the summary of three famous classification techniques, i.e. decision trees, Naïve Bayes and k-nearest neighbor have been opted for this study.

A. Decision Trees

A decision tree is a non-cyclic tree structure which consists of root node, connecting branches and internal nodes (leaf nodes) [2], [4]. Each leaf node corresponds to an attribute denotes a test on it and holds the class label whereas each branch from a sequential path denotes the test outcome. The node at the topmost of the tree called the root node which represents the entire datasets [2], [4]. The tree always starts with the single node containing training datasets [16]. If the tuples in a dataset belongs to the same class then the node turns into a leaf, labeled with that corresponding class [16]. Otherwise, an attribute selection method is used to determine the splitting criterion. Such a method may use a heuristic or statistical measure (e.g., information gain or gini index) to select the best way to separate the tuples into individual classes [16].

B. Naïve Bayesian Classifier

In terms of machine learning, Naive Bayes classifier is a kind of simple probabilistic model to solve problems controlled by strong independence assumptions [17]. It's highly scalable and fast to train data very efficiently in a supervised learning situation with high accuracy in numerous applications [4], [17]. We have a set of unknown tuples (instances), embodied by an n-dimensional vector, $\boldsymbol{X} = (x_1, x_2, \dots., x_n)$, with the probabilities of instances $P(C_i|X)$(where i is possible outcome of a class). The posterior probability can be decomposed as:

$$P(C_i|X) = \frac{P(C_i)\ P\ (X|C_i)}{P(X)}$$

C. k-Nearest Neighbors Algorithm

The k-nearest neighbors' algorithm (k-NN) is regard as non-parametric method in the field of pattern recognition to classify records based on similarity measures learning [4], [18]. Two records or objects are measured by the distance between them based on the likeness of two records [4]. The output is considered as a class membership [18]. A record or object remains classified by a majority vote of its neighbors, or in other words, the k records with the minimum distance to the anonymous record with k is a positive integer and typically small [4], [18]. If k = 1, then the record or object is merely assigned to the class of that single nearest neighbor [18]. Hence, the k-NN algorithm is simplest among the machine learning algorithms.

IV. Data Preprocessing and Methodology

In the field of data mining, data preprocessing is a crucial step to deal with incomplete, noisy and inconsistent data [19]. Data preprocessing includes various tasks such as data cleaning, data integration, data transformation, data reduction, data discretization, etc. to continuously formulate data in a consistent and accurate style. For our case study, the data has been collected and preprocessed of three different cohorts (two class sections each) of three different engineering disciplines (Computer Engineering, Software Engineering, and Electronic Engineering) of the SSUET, Pakistan. The students' pedagogical progress is analyzed by taking single (core) course sessional marks of different technologies. The research focuses on three comparative studies in order to track and analyze engineering student's pedagogical progress are: comparative analysis of a performance of three different courses (use to teach in different engineering technologies), comparative analysis of gender wise performance in each technology and comparative study between two sections students' performance in a particular course.

Overall 290 undergraduate engineering students enrolled in academic batches 2011 Computer Engineering with 102 students (Section D and E), 2014 Software Engineering with 94 students (Section A and B) and 2017 Electronic Engineering with 94 students (Section C and D) has been comparatively analyzed for academic progress in their particular course in different semesters. Only pre-examination (sessional) marks of students have been used in this study. Different variables as arbitrating parameters have been selected to measure students' academic progress. Different parameters and response variables varying according to technology and course are mentioned in a Table I for reference.

TABLE I. SESSIONAL VARIABLES IN DATASET

Variable	Values	Technology	Course	Possible Results
Gender	{Male, Female}	C.E, S.E, E.E	RDBMS, DS&A, OOP	
Mid-Term Marks	{12-15, 11-7, <=6}	C.E	RDBMS	{Excellent, Average, Poor}
	{15-20, 10-14, <=9}	S.E, E.E	DS&A, OOP	
Test / Presentation Marks	{4-5, 3, <=2}	C.E	RDBMS	{Excellent, Average, Poor}
Assignment Marks	{4-5, 3, <=2}	C.E	RDBMS	{Excellent, Average, Poor}
Lab Performance	{4-5, 3, <=2}	C.E	RDBMS	{Excellent, Average, Poor}
	{15-20, 10-14, <=9}	S.E, E.E	DS&A, OOP	
Project Demonstration	{8-10, 7-5, <=4}	C.E	RDBMS	{Excellent, Average, Poor}
Bonus Marks	{Yes, No}	C.E, S.E	RDBMS, DS&A	
Quiz/Test Marks	{7-10, 5-6, <=4}	S.E, E.E	DS&A, OOP	{Excellent, Average, Poor}
Total Marks	{32-40, 24-31, <=23}	C.E	RDBMS	{Excellent, Average, Poor}
	{35-50, 34-25, <=24}	S.E, E.E	DS&A, OOP	

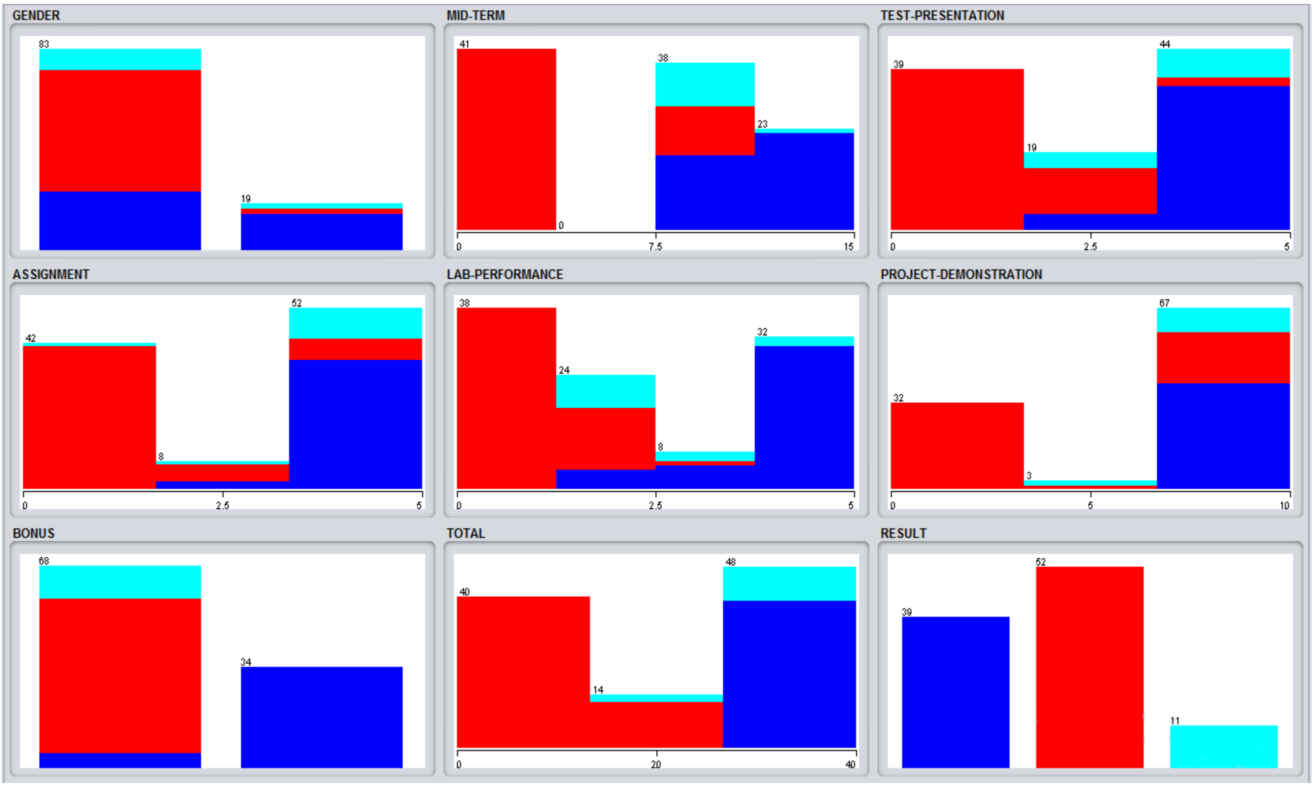


Fig. 1. Overall attributes visualization of RDBMS.

The course name acronym are RDBMS stands for Relational Database Management System, DS&A stands for Data Structures and Algorithm, and OOP stands for Object Oriented Programming. The preprocessing using Weka tool

[20] to analyze three different courses of different technologies has been presented below in the form of graphical visualization and statistical analysis of the attributes as mentioned in Table I. For the course RDBMS and technology Computer Engineering, 83 instances of male and 19 instances of females have been analyzed in the dataset. The preprocessing results clearly shows in Fig. 1 that 39 instances were Excellent progress (visualized in a dark blue bar), 11 were Average progress (visualized in a cyan bar), and 52 were Poor progress (visualized in a red bar) mined from the Result attribute.

For the course DS&A and technology Software Engineering, 59 instances of male and 35 instances of females have been analyzed in the dataset. The preprocessing results clearly shows in Fig. 2 that 69 instances were Excellent (visualized in a dark blue bar), 12 were Average (visualized in a cyan bar), and 13 were Poor (visualized in a red bar) mined from the Result attribute.

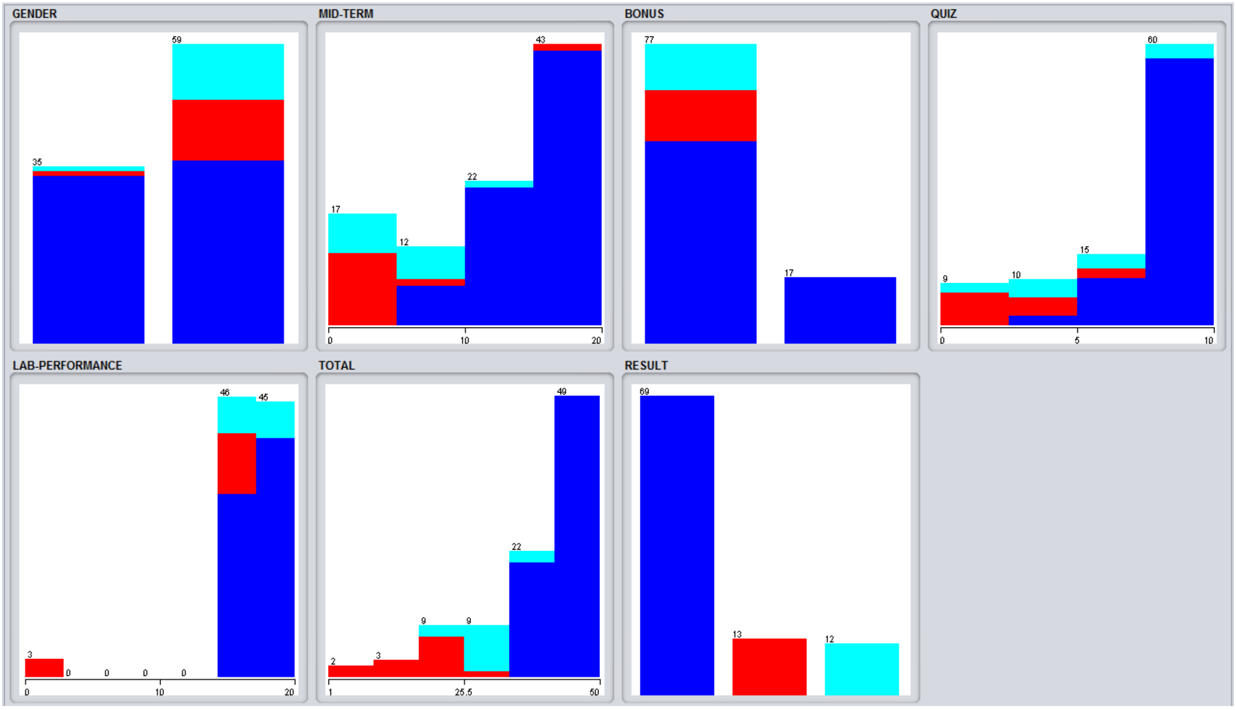


Fig. 2. Overall attributes visualization of DS&A.

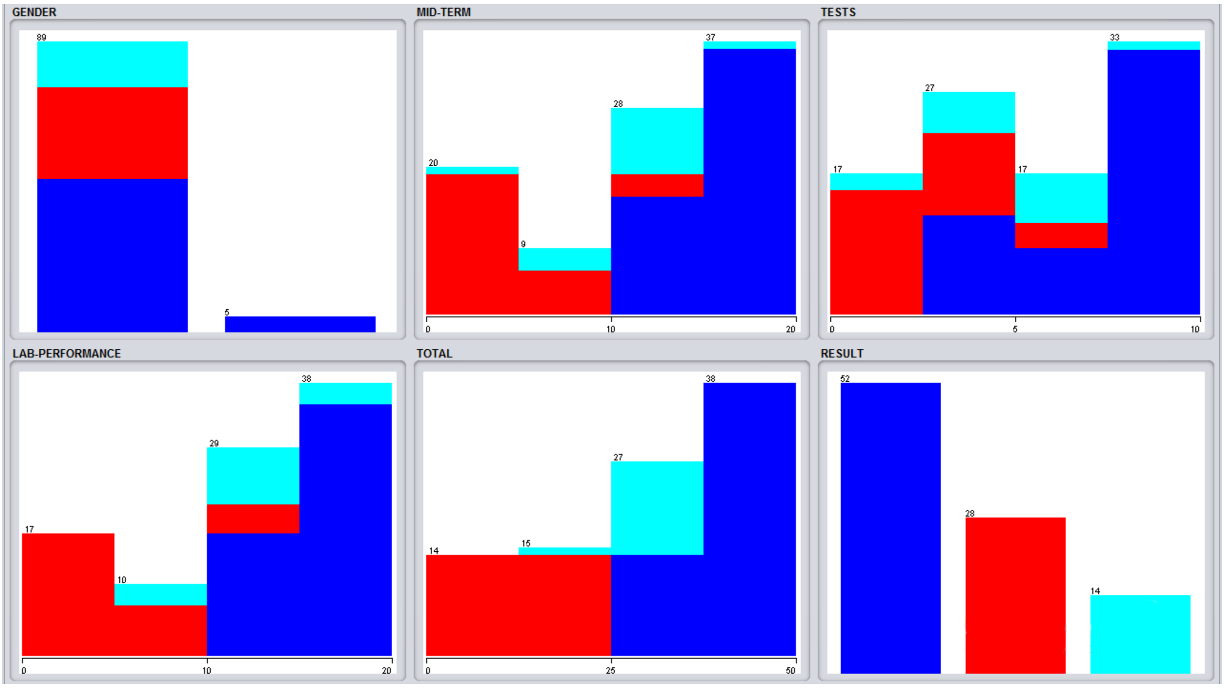


Fig. 3. Overall attributes visualization of OOP.

For the course DS&A and technology Software Engineering, 59 instances of male and 35 instances of females have been analyzed in the dataset. The preprocessing results clearly shows in Fig. 2 that 69 instances were Excellent progress (visualized in a dark blue bar), 12 were Average progress (visualized in a cyan bar), and 13 were Poor progress (visualized in a red bar) mined from the Result attribute.

For the course OOP and technology Electronic Engineering, 89 instances of male and 5 instances of females have been analyzed in the dataset. The preprocessing results clearly shows in Fig. 3 that 52 instances were Excellent progress (visualized in a dark blue bar), 11 were Average progress (visualized in a cyan bar), and 52 were Poor progress (visualized in a red bar) mined from the Result attribute.

The Weka (Waikato Environment for Knowledge Analysis) tool (Knowledge Explorer) is the useful GUI (Graphical User Interface) tool with the collection of major machine learning algorithms coded in java [1], [14], [20]. It basically contains applications for data pre-processing, association rules, classification, clustering, regression and visualization [1]. Engineering under-graduate students of three different cohorts of three different technologies with three different courses have been studied comparatively to trace their pedagogical progress using sessional (pre-examination marks) in particular courses. The tool Weka was employed to carry out study using three different classification algorithms Decision trees J48, Naïve Bayes, and K-Nearest Neighbor [20]. The reason of particularly using these three algorithms for classification beside other algorithms in a study is that they give better results and represent rules which can be simply interpretable by humans and therefore can be used in making decision rules [4].

V. Results and Discussion

The course wise results per technology and per gender of accuracy, kappa, mean absolute error, root mean squared error, relative absolute error, and root relative squared error with different classifiers are organized in Table II. The kappa statistic basically measures the agreement of prediction with the true class -- 1.0 indicates complete agreement. The Mean Absolute Error simply measures the average degree of the errors in a set of estimates, without considering their direction. It simply measures accuracy for continuous variables. The Root Mean Squared Error is a quadratic scoring rule which mainly measures the average degree of the error. Relative values are simply ratios, and have no units. The ratios are commonly expressed as fractions (e.g. 0.762), as percent (fraction x 100, e.g. 76.2%), as parts per thousand (fraction x 1000, e.g. 762 ppt), or as parts per million (fraction x 106, e.g. 762,000 ppm).

Fig. 4 clearly indicates the results of comparative study of three courses and per gender per three different technologies in terms of accuracy using three different classifiers for data classification. From Table II and Fig. 4, J48 and k-NN achieves the highest accuracies for the course RDBMS whereas k-NN achieves the highest accuracy when analyzing the results based on gender in the same course. In the same manner, k-NN has maximum accuracy for the course DS&A as compared with the others two classifiers and for gender analysis as well. Likewise, J48 and k-NN attains highest accuracies for the course OOP whereas k-NN was best in accuracy than the rest of the two classifiers for gender analysis.

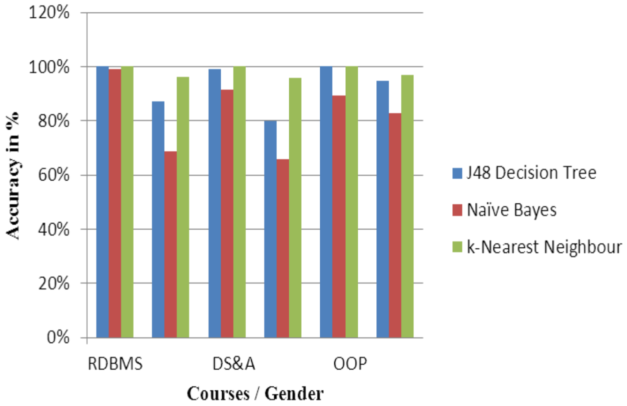


Fig. 4. Comparison of classifiers accuracy.

For J48 algorithm, the decision trees are obtained based on above study results shown in Fig. 5 to 9.

By observing the decision trees, the students' progress in different courses of different engineering technologies can be analyzed and tracked in order to find out academic strengths and weakness. The tree in Fig. 5 indicates that the students who scored above 31 marks have excellent progress with 39 students, those who scored above 23 and less than equal to 31 have average progress with 11 students, and those who scored below and equal to 23, they have a poor progress with 52 students in a course RDBMS.

TABLE II. CORRELATION RESULTS COURSE WISE PER TECHNOLOGY PER GENDER

Course / Gender	Classifier	Accuracy		Kappa statistic		Mean absolute error		Root mean squared error		Relative absolute error		Root relative squared error	
RDBMS (CE)	J48 Decision Tree	100 %	87.2549 %	1	0.4289	0	0.1893	0	0.3077	0 %	61.6848 %	0 %	79.019 %
	Naïve Bayes	99.0196 %	68.6275 %	0.9832	0.3401	0.0297	0.3081	0.1081	0.5389	7.6244 %	100.3648 %	24.5272 %	138.3908 %
	k-Nearest Neighbour	100 %	96.0784 %	1	0.8592	0.0089	0.0542	0.0108	0.1553	2.2722 %	17.6433 %	2.4624 %	39.8842 %
DS&A (SE)	J48 Decision Tree	98.9362 %	79.7872 %	0.975	0.589	0.0131	0.2934	0.0809	0.383	4.5336 %	62.6884 %	21.4719 %	79.2323 %
	Naïve Bayes	91.4894 %	65.9574 %	0.8116	0.3542	0.0576	0.354	0.1833	0.5338	19.9468 %	75.6293 %	48.6552 %	110.4211 %
	k-Nearest Neighbour	100 %	95.7447 %	1	0.911	0.0114	0.0637	0.0127	0.1671	3.96 %	13.6103 %	3.3748 %	34.5606 %
OOP (EE)	J48 Decision Tree	100 %	94.6809 %	1	0	0	0.1007	0	0.2244	0 %	92.3716 %	0 %	99.9141 %
	Naïve Bayes	89.3617 %	82.9787 %	0.8277	0.272	0.0611	0.1778	0.178	0.3557	15.6583 %	163.018 %	40.3704 %	158.3742 %
	k-Nearest Neighbour	100 %	96.8085 %	1	0.558	0.0092	0.0382	0.0114	0.1266	2.3657 %	35.0721 %	2.5962 %	56.3469 %

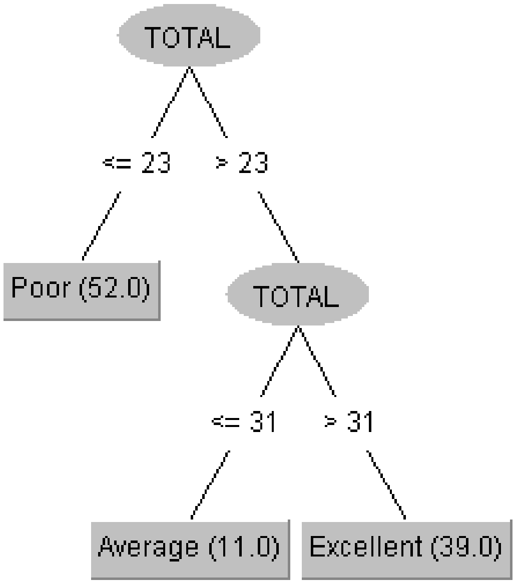


Fig. 5. Decision tree produced for RDBMS.

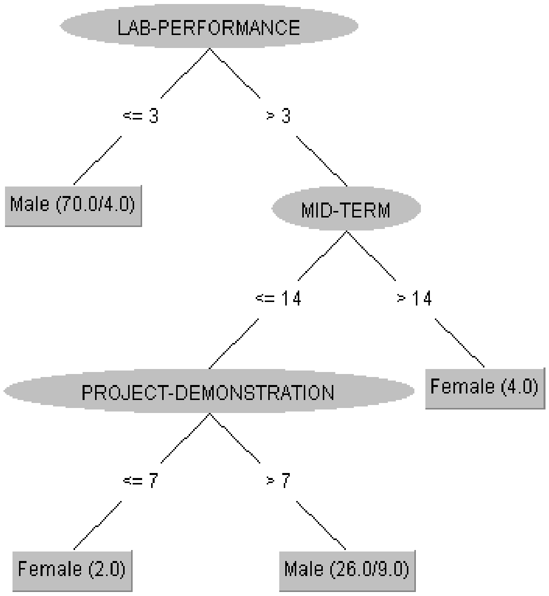


Fig. 6. Decision tree produced for gender (in lab performance and project demonstration).

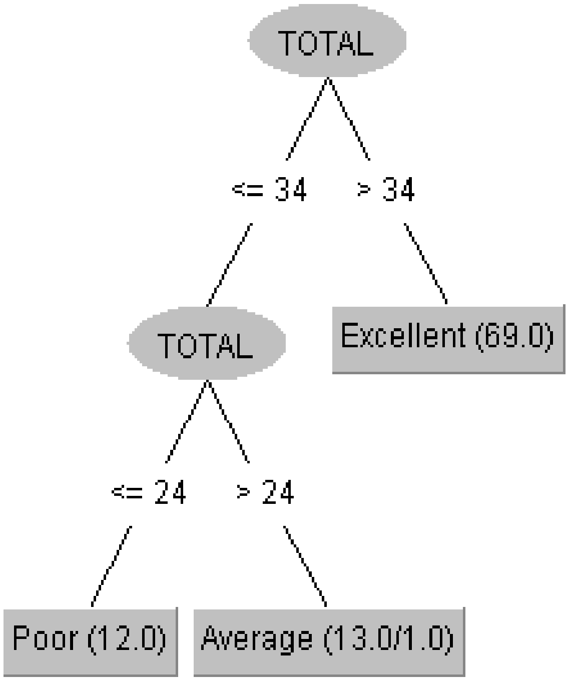


Fig. 7. Decision tree produced for DS&A.

The tree in Fig. 6 indicates that particularly in lab-performance and in a mid-term, female students' progress is better than male students whereas male students were comparatively better than female students in a project demonstration of Computer Engineering technology.

The tree in Fig. 7 indicates that the students who scored above 34 marks have excellent progress with 69 students, those who scored above 24 and less than equal to 34 have average progress with 13 students (one misclassified record), and those who scored below and equal to 24, they have a poor progress with 12 students in a course DS&A.

The tree in Fig. 8 indicates that particularly in a quiz and mid-term, female students' progress is comparatively better than male students whereas male students were better than female students in a lab performance and bonus marks of Software Engineering technology.

The tree in Fig. 9 indicates that the students who scored above 34 marks have excellent progress with 52 students, those who scored above 20 and less than equal to 34 have average progress with 14 students and those who scored below and equal to 20, they have a poor progress with 28 students in a course OOP.

The tree obtained for gender wise comparison in a course OOP was an incomplete tree with 1 number of leave(s) and size of a tree for only male(s) students consisting of 5 incorrectly classified instances for all female students due to false positives instances classified in a confusion matrix. False positives can be defined as a class of instances with number of instances predicted positive that are actually negative.

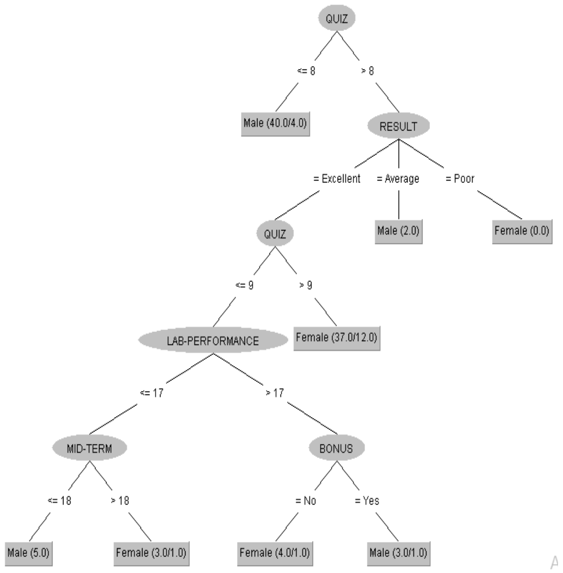


Fig. 8. Decision tree produced for gender (in a quiz and lab performance).

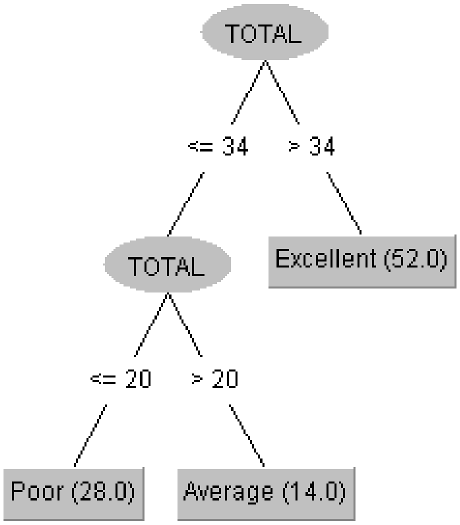


Fig. 9. Decision tree produced for OOP.

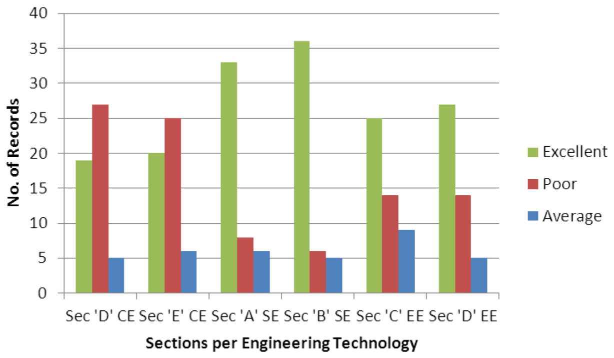


Fig. 10. Comparison of section wise per engineering technology.

Fig. 10 shows the results of comparative study of two sections per engineering technology with respect to excellent, poor or average progress. In a computer engineering technology, majorly students' progress displays poor in both sections in above figure but section 'D' shows further poor progress than section 'E'. In a software engineering technology, major students' progress contribution is excellent but section 'B' shows more excellent progress than the other section as shown in a figure. In an electronic engineering technology, most of the students are shown excellent progress but section 'D' is comparatively improved than section 'C'.

VI. Conclusion and Future Directions

The present study aims the significance, scope and techniques of data mining in the domain of education is addressed in a multiple education disciplines and technologies at higher education level interestedly. One of the useful and widely used data mining methodologies is classification. There are many classification algorithms but three mostly used classification algorithms have been used in this study. In this study, only a single (core) course sessional or pre-examination marks of three different cohorts belonging to computer, software and electronic engineering have been analyzed and followed to determine students pedagogical progress in their corresponding engineering fields and their learning attitude towards particular course(s) for the preparation of final examination based on their pre-examination marks. Further, gender wise and section wise progress has also been studied for each stated engineering technologies. The highly influencing sessional variables have been identified as a criterion of awarding and judging students' academic progress in a course(s) and applying data mining classification techniques to implement high potential data mining applications at higher education level, referring to the optimal manipulation of data mining approaches and techniques to deeply analyze and track the engineering student's pedagogical progress throughout the academic session. The study basically employs the classification techniques with three different classifiers: J48 Decision Trees, Naïve Bayes and K-NN to classify attributes affecting students' progress in their core course(s) for the betterment of academic stakeholders' assistance and regulation in order to improve academic progress which is the main objective of study. Essentially, as mentioned earlier, these three algorithms have been selected for this study because they perform better than any other technique and are easily interpretable. Among the three algorithms, J48 and K-NN gives better results in terms of accuracy.

In future, some more courses with sessional or pre-examination marks as well as final examination marks of more engineering technologies for students' academic evaluation can be deemed in order to refine and embody continuously smooth pedagogy and learning process. Thus, the study helps and guide students to improve their academic performance and reduce failure in a course(s) by taking appropriate actions and to increase retention for the semester examination.

References

[1] Fernandez DB, Lujan-Mora S, "Comparison of applications for educational data mining in Engineering Education," 2017 IEEE World Engineering Education Conference (EDUNINE). 2017.

[2] www.tutorialspoint.com. (2018). Data Mining Tutorial. [online] Available at: https://www.tutorialspoint.com/data_mining/. [Accessed 21 Feb. 2018].

[3] Essay UK Free Essay Database. (2018). Essay: Educational Data Mining (EDM) - Essay UK Free Essay Database. [online] Available at: http://www.essay.uk.com/essays/information-technology/essay-educational-data-mining-edm/. [Accessed 21 Feb. 2018].

[4] Asif R, Merceron A, Pathan MK, "Predicting Student Academic Performance at Degree Level: A Case Study. International Journal of Intelligent Systems and Applications," 2014Aug;7(1): pp. 49–61.

[5] Yassein NA, Helali RGM, Mohomad SB. "Predicting Student Academic Performance in KSA using Data Mining Techniques," Journal of Information Technology & Software Engineering. 2017;07(05).

[6] Kumar B, Pal S, "Mining Educational Data to Analyze Students Performance," International Journal of Advanced Computer Science and Applications. 2011;2(6).

[7] En.wikipedia.org. (2018). Statistical classification. [online] Available at: https://en.wikipedia.org/wiki/Statistical_classification#Algorithms/. [Accessed 23 Feb. 2018].

[8] Archana T, Gandhi UD, "Prediction of Student Performance in Educational Data Mining - A Survey. International," Journal of Pharmacy & Technology. 2016Sep;8(3): pp. 17757–63.

[9] Adhatrao K, Gaykar A, Dhawan A, Jha R, Honrao V, "Predicting Students Performance Using ID3 and C4.5 Classification Algorithms," International Journal of Data Mining & Knowledge Management Process. 2013;3(5):39–52.

[10] Dutt A, Ismail MA, Herawan T, "A Systematic Review on Educational Data Mining," IEEE Access. 2017Jan17;5: pp. 15991–6005.

[11] Asif R, Hina S, Haque SI, "Predicting Student Academic Performance using Data Mining Methods," International Journal of Computer Science and Network Security. 2017May;17(5):187–91.

[12] Asif R, Merceron A, Pathan MK, "Investigating Performances' Progress of Students," pp.116–23.

[13] Asif R, Merceron A, Ali SA, Haider NG, "Analyzing undergraduate students performance using educational data mining," Computers & Education. 2017;113: pp. 177–94.

[14] Yadav SK, Bharadwaj B, Pal S, "Data Mining Applications: A comparative Study for Predicting Student's performance," INTERNATIONAL JOURNAL OF INNOVATIVE TECHNOLOGY & CREATIVE ENGINEERING. 2012Feb;1(12): pp. 13–9.

[15] Asif R, Merceron A, Pathan MK, "Investigating performance of students," Proceedings of the Fifth International Conference on Learning Analytics And Knowledge - LAK 15. 2015.

[16] Agarwal S, "Data Mining: Data Mining Concepts and Techniques," 2013 International Conference on Machine Intelligence and Research Advancement. 2013.

[17] En.wikipedia.org. (2018). Naive Bayes classifier. [online] Available at: https://en.wikipedia.org/wiki/Naive_Bayes_classifier [Accessed 13 March. 2018].

[18] En.wikipedia.org. (2018). K-nearest neighbors algorithm. [online] Available at: https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm [Accessed 13 Mar. 2018].

[19] En.wikipedia.org. (2018). Data pre-processing. [online] Available at: https://en.wikipedia.org/wiki/Data_pre-processing [Accessed 05 Apr. 2018].

[20] Cs.waikato.ac.nz. (2018). Weka 3 - Data Mining with Open Source Machine Learning Software in Java. [online] Available at: http://www.cs.waikato.ac.nz/ml/weka/ [Accessed 07 Apr. 2018].

Student's Opinions on Online Educational Games for Learning Programming Introductory

Roslina Ibrahim, Nor Zairah A. Rahim, Doris Wong H. Ten, Rasimah C.M Yusoff, Nurazean Maarop, Suraya Yaacob
Advanced Informatics School
Universiti Teknologi Malaysia
Jalan Sultan Yahaya Petra, 54100, Kuala Lumpur, Malaysia

***Abstract*—Use of educational games is an approach that has potential to change the existing educational method. This is due to games popularity among younger generation as well as engagement and fun features of games compared to conventional learning method. In addition, games are among the most widespread media amongst younger generation or so-called "digital natives" apart from movie, music and internet technology. Game play activities is an important issue to be thoroughly understood due to the facts that many of them are addicted to game play activity. In contrast, conventional learning approaches are not interesting enough to the younger generation. Thus, integration of games technology into education is potentially believed to increase student interest and motivation to learn. This study developed and evaluates an online educational game for learning Programming Introductory course at a university in Malaysia. A total of 180 undergraduate students from computer and engineering background participate in the study. Findings shows that about 80% of students have positive attitude towards the games with around 84% of them find that the games is a fun way to learn, at the same time, an average of 80% agreed that the game provide them with opportunity to learn. Furthermore, about 75% of the students agreed that the game make them able to do self-assessment for Programming course. It was interesting to find that almost 85% of the student said that they will want to use educational games as their future learning approach. Despite many more evidence will be needed especially in Malaysia context, this study is important to rationalize that games can be one of the new learning approaches in the future.**

***Keywords*—Educational games; programming introductory; undergraduate; games evaluation**

I. Introduction

The advent uses of video and computer games have gained enormous interest among researcher to study how it is possible to be use as one of the learning approach [1]-[4]. It was stated that the growth of educational games studies is very rapid especially in the last decade [5]. Many studies were conducted from numerous perspectives including educational games design, games development, learning effectiveness and retention, how students learn by using the games and so on [6]-[9]. This, among others is due to games characteristics that are immersive, challenging, fun, engaging and highly motivated [10]-[12]. Those criteria are not easily found in any other conventional learning approach. Besides, [5] have associated game play as one of activities for digital native – generation X and millennial whose grow up with the technology. Games also have been said to have many potential learning benefits for 21st century skills such as communication skills, high order thinking skills, problem solving and able to prepare new generation for new kind job of challenges as well as new skills [13]. Furthermore, games are suggested to be able to cover diverse learning principles as suggested in [14] that revealed that good games incorporated as many as 16 learning principles including:

- Identity: Players will have a character that they need to play as someone in a game to accomplish any game task.
- Interaction: The games provide substantial interaction with players. It will become idle if player does not interact with it.
- Challenge and Consolidation: Games can offer a set of challenge or problems and require the players solve it. Difference with school system, poorer students sometimes doesn't get enough opportunity to consolidate while good students do not get ample challenge to solve.
- Well order problems: Problems or hurdles in games are design to be in order so that the earlier ones will lead to the next problems.
- Pleasantly frustrating: Failure is an options in games that provides challenge for players to accomplish. Games are also avoiding player feeling humiliated if they fail the challenges. School, on the contrary does not really provide failure or learning by mistake as provided in games.
- Explore, Think Literally, And Rethink Goals: Games can inspire players to carefully discover the options, think and used such discovery to think in order to achieve one's goals.

Games and play can be an effective learning environment not just because it is fun but also due to features including immersive, require the players to make recurrent and vital decisions, adjust to each player individually, contains clear goals, and involve a social network [15]. In another study, [16] suggested that students in technical universities faces the problem of low motivation, and further added that games can be an effective way to improve student motivation and learning.

Due to potential of educational game as the new learning approach, interesting features of games and interest of new generation towards games and computer technology, it is important to have thorough study on how games can help students to learn and how students perceived game as their learning approach.

II. Past Studies on Educational Games in Malaysia

Based on literature gathered, many studies on educational games were done in developed countries especially in US, UK and European countries [17]-[20]. USA, for example have its summit on educational games as well as many associations, conferences and websites to handle issues regarding using games for educational purposes [21]. In Asia, most studies also were done in China, Taiwan, Korea and Japan [22]-[27]. However, Malaysia has rather a limited number of papers and research regarding games based learning. Fortunately, the activity has started and growing number of studies were found in literatures. Besides, Malaysia needs to have more study and evidences about the use of educational games in our educational environment in order to establish our very own educational policy on the use of these recent technologies [28].

Generally, there are two types of educational games studies conducted in Malaysia so far, one is development of games and testing its effectiveness towards student, the other is development of new framework or tools in assisting games development and evaluation process. This paper mostly covers studies that use games and its evaluation on the students. Study by [29] integrated educational games within a courseware application. The study designed a courseware that combines games and storytelling for 7 and 8 years old school students with the topic “Morality”. Students found the courseware fun and interesting with good usability features. The game inside the courseware was also found effective in delivering learning content and increased their motivation to learn.

Study by [30] reviewed and proposed a framework for evaluation of educational games from user experience perspectives. The framework can be used by game designers to design and evaluate user experiences of educational games. Study by [31] proposed and evaluate a model to measure the determinants of games actual use. Data were collected among students from a public university in Malaysia. Several factors were found to influence the use of games including enjoyment, behavioural control, subjective and attitude. Meanwhile [32] study the associations between game play and academic achievement among form 1 students in a Malaysia secondary school. It was found that about 75% respondents of were active gamers. Students spent an average of 8.4 hours a week to play games. Interestingly, it was also found that game play activity is weakly associated with student academic achievement. Another study done to enhance students creative perception through teaching of games development [33]. It was found that group of students who develop games shows significant different compared to group who use ordinary teaching method.

Another study done by [34] about educational game for learning Islamic Education among primary school students. Samples were taken from 50 primary school students and 3 teachers. The study developed an adventure games prototype titled “Adventure with Ibrahim” based on National Primary School Curriculum (KBSR) with the topic “Morality”. A usability test was using several usability criteria namely student concentration on the games, thoughtful of the quizzes given, knowledge and favourites parts on the games and student reaction about the games. Interestingly, majority of the student agree to use games in all topics of the subject because of games fun features. All teachers are also agreed if the more topics were taught by using the game.

A study done by [35] on the background and game experience of a Malaysian primary and secondary school students. It was found that almost 100% of male student are playing computer games while about 90% of female said so. It was also found that Malaysian students are acquainted with most of the genres of computer games including strategy, role playing games, action, adventure, multiplayer online games, simulation, puzzle and trivia as well as game for learning. Computer is the most popular platforms (80%) followed by Sony PlayStation 1 and Sony PlayStation 2 (41%), and mobile phone (33%). Generally, both male and female students accept the idea of using computer games as their learning medium. In addition, it was found that game play encourages social skills among the students.

A study investigated the potential of educational games to learn History subject [1]. Data was collected at a secondary school in Bangi Malaysia. It was found that more than 90% of the students have experience in playing computer games. As a matter of fact, more than 30% of the student plays game more than 3 hours per week. About the popular platforms., around 60% students plays game using console followed by using computer, handheld and other devices. The most popular genres are adventure games (more than 60%), trailed by fighting games, puzzles and sports games. Asking about the reason why they play the games, about 70% says they play game for fun while more than 60% because they need to fill up free time, fantasy features in games (46%), followed by adventurous features and challenge. Most student think that the advantage of game plays are fun, improving electronic skills, calming themselves down, fill up free time as well as inducing their creativity. He suggested that educational games have high potential to be used as the substitute approach to the subject considered as boring such as history subject.

Meanwhile in another study, [36] have reviewed studies on history games based learning and games features that are most effective in promoting engagement and supporting the process of learning history subject. It was found that gaming experience, learning experience, adaptivity and usability are the important features in supporting good user experience in learning history using games.

Study by [37] argues that ordinary courseware is missing in providing support for students to evaluate their achievements. To overcome that, she proposed a mini educational computer board game prototype to learn Science for Kindergarten student. Courseware content, presentation style, user interfaces and courseware elements of multimedia were used for evaluation. It was found that more than 95%

students feel that the games help them in their learning. All students were found to be more motivated to learn the subject while about 90% says the games generate their interest to study. Student also likes the games fun activities, styles, user interface and easy to use features. Teacher also found the games is interesting compared to ordinary ways of teaching. The games were found interesting enough in providing interactive and fun learning environment for the kids.

Another study use computer games to introduce programming subject to children between age 5 to 7 years old [38]. It was suggested that use of games might help the children to understand the basic concept of programming given that it is a complex subject. Furthermore, educational games provide a unique opportunity for integrating the cognitive, affective, and social aspects of learning. The finding indicates that students found the games as very interesting and they do grab basic concept of programming data types (integer, float and character).

TABLE I. PAST STUDIES ON EDUCATIONAL GAMES AND ITS EFFECTIVENESS

Author	**Genre/ sample**	**Subject**	**Findings**
[3]	Simple games	Computer Science	Increase motivation
[29]	2D game	Islmic Ethic	Increase motivation
[40]	Action/ N=1274	Reading and Maths	Increase student motivation and technology literacy
[41]	Simulasi/ N=96	Physics	Students achievement increase
[42]	Adventure N=130	Computer basic	Highly motivated and increase achievement
[43]	N=32	Road safety and fire	Kids learn the intended concepts
[44]	*RPG*	Computer Science	Motivates students
[37]	Chess	Moral	Increase motivation
[34]	Adventure	Moral	Increase motivation to learn
[45]	N=100	Physiology	Increase interest and knowledge self learning and fun
[46]	Mini Games N=65	Culture	Game provide opportunity to learn
[45]	N=100	Physiology	Increase interest and knowledge self learning and fun
[38]	Mini 2D 5	Computer Science	Kids love it
[47]	word	Computer Science	High interest to learn
[48]	Adventure N=60	History	Increase student knowledge and achievement to learn history

A study on why educational games matters to the future of education [39] proposed a framework of educational games that shows the relationship between educational game engagement and motivation to clear goals, rules, feedback, playability and control. It analyse the game attributes and its correlation with the theory of play that form the basis of educational games. It was found that games can motivate and encourage learners as well as able to provide powerful learning environments especially for kids. In some case, it can encourage students to learn on certain subjects that they may not have interest at all.

On different note, [28] did a review on use of computer games for learning indicate that games have a great prospective as teaching and learning tools. The amazing elements in computer games such as fun, interactive, curiosity, challenging, fantasy, competition, clear goal, and encouraging feedback made games as one of the highly potential and flexible technology for learning. He further added that 21st century skills such as adaptability, risk-taking, self-direction, interactive communication, planning and managing are the skills that is difficult to teach; therefore, a proper games design may provide an improved way in teaching those skills. Games are also fast and receptive as well as able to handle large amount of content and can be easily updated. Table I shows past studies on educational games and its effectiveness.

III. Research Methodology

The methodology in this study was divided into two parts which are the online game design and development, and game evaluation. The following section discuss in details for both process.

A. Game Design and Development

The online educational game was design based on elements that an educational should have in order to make an effective game. Many studies suggested that educational game must integrate both games design elements and educational elements [46], [49]-[53]. Educational games elements discuss in those studies must be carefully chosen to ensure suitability with different genre and requirement for an educational game. In this case, we have come out with the educational games design elements adapted from those studies (Fig. 1).

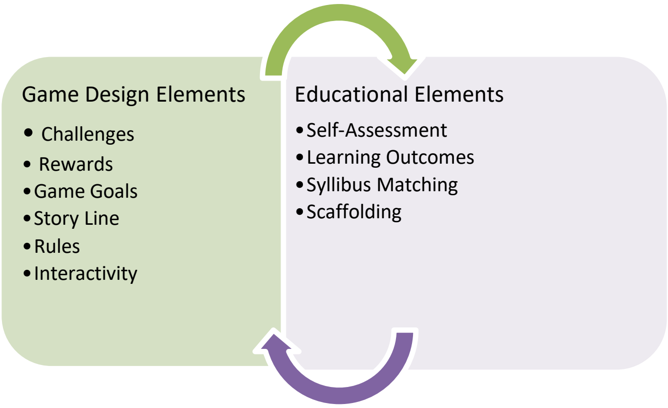


Fig. 1. Educational games design elements.

Challenges is referring to things need to be overcome by the player. It was stated in [54] that good games must integrate suitable challenges to the target players as well as the games feature. The challenges must be carefully design with the precaution that it is neither too easy nor too hard. Challenge scan be between player versus computer or player versus other players. Rewards is referring to things that the player will get after passing certain challenges or achieving certain games goals. Among the examples are games score, health status and things that can be used to proceed further in the games. Rewards can be given during the game play session or at the end of the session [50]. Game goal is the main thing need to be achieved by the player of any games. It is the final point of playing the games. Goals can engage player to the games because it connects the player to their inner motivation to achieve things and feeling satisfied about it.

Story line is the flow of the game event that connects the game cycles into complete sequences. Good games need to have good story that relate the whole things together [55]. Apart from that, games also need to contain rules to be adhering by the players similar as physical games such as football, golf and so on. It's a set of procedure that must be following in order to proceed within the same levels or to the next game levels. Interactivity refers to interaction between the players with the games characters, environment and feedback or message during the gameplay session.

This online game designed to include educational elements as listed in Fig. 1. Self-assessment refers to integration of educational content into the assessment form. It can be in many forms depending on the games type. The game is also design to achieve certain learning outcomes of a topic or course. A course or topic normally have its own learning outcomes ready, designers can adapt the content so that it will meet the outcomes of a topic content or things to be achieved by the player by the end of the game session. Another important educational element is syllabus matching [52] whereby the games content should follow an authentic syllabus that match the learners need. This includes their school syllabus or any knowledge that they are acquiring. Feedback is another important element in education whereby learners will be given the information about what they have done or achieved. Feedback can help learners learn their mistakes or confirming their knowledge gained during the gameplay session.

Using the elements discuss above; we have design and developed an online educational games prototype called ROBO-C purposely for learning C Programming Introductory. It contains four game modules together with a note module. The game modules are Find Me, Go Grow, Hangman and Fix the Bot. Each game has its own levels with different presentation of learning content. The game design went through several processes including content design, games genre design, content matching to games and levels and also notes design. The game content focusing on Programming topic called as Looping. This includes Introduction to Looping, Types of Looping, *continue* and *break* Statement, *do-while* Statement, *for* Statement and *while* Statement. Fig. 2 and 3 shows the game screen shot.

Fig. 3 shows the screen for one of the game module which is Go Grow. This module has 3 levels with 15 self-assessment questions provided in each level. The flower tree will grow into a nice flowering plant depend upon the correct answer and score the player will get. It also has Hints button to helps player get the clue for every question. ROBO-C Prototype was hosted in local server for testing and data collection purpose.

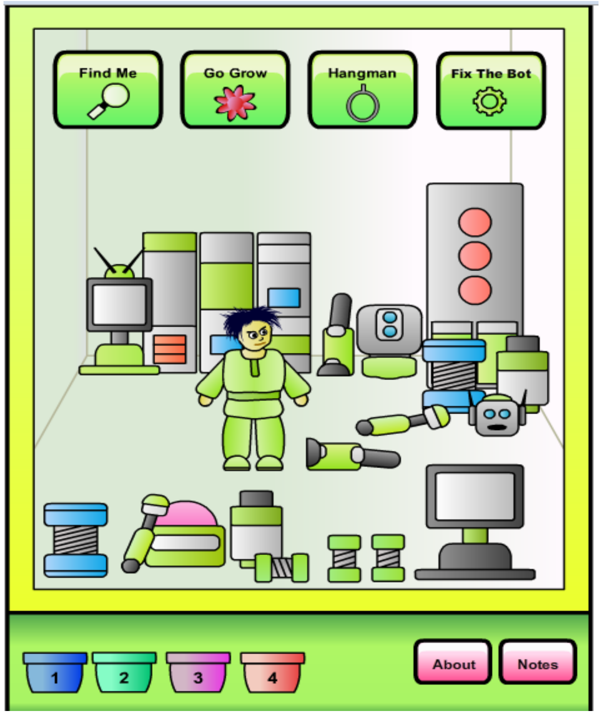
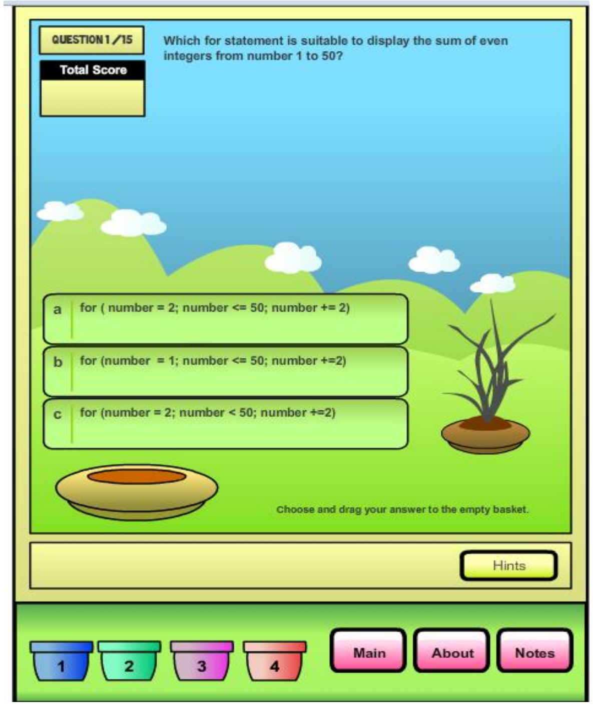


Fig. 2. ROBO-C starts screen.



Which for statement is suitable to display the sum of even integers from number 1 to 50?

a for (number = 2; number <= 50; number += 2)

b for (number = 1; number <= 50; number +=2)

c for (number = 2; number < 50; number +=2)

Choose and drag your answer to the empty basket.

Fig. 3. Interface for game module go grow.

B. Game Emprirical Evaluation

Game evaluation was done using instrument developed from analysis of past studies on acceptance theories [56], [57] [24], [58]. It was then modified to meet the educational games technology. Six components were used for game evaluation purposes which are Ease of Use [56] Usefulness [56], [57], [59] Attitude [56], [57], Self-Efficacy [57], Anxiety [24], [57] Enjoyment [60] and Intention to Use [56]-[58].

Using above as the references, the following text describe the meaning of each construct. Usefulness is defined as how the respondents think that the application can help them to perform the task they are doing. Ease of Use is defined as how easy the application can be use and learn by them. Attitude is overall affective reaction towards the use of application while self-efficacy describes the judgement of one's ability to use the application. Anxiety defined as emotional fear, apprehension and phobia felt by individuals in using the application. Enjoyment describe as state of mind or an individual trait. Intention to Use defines as the plan that individual have in using the application.

A total of 27 items were derived and modified from the studies and design into Likert' scale of 5 scales ranging from 1 (strongly disagree) until 5 (strongly agree). Respondents were asked to play the game for about an hour or until they are ready to answer the survey question. 180 students took part in the data collection session with all of them play the game for about an hour and fill up the question afterwards. All of them stay until the session finished. All respondents do not have any formal experience using educational games in their study.

IV. Data Analysis and Result

Data were analysed using statistical packages IBM SPSS 20 for both reliability analysis and descriptive analysis. The respondents consist of 180 undergraduates student from IT (62%) and Engineering (38%) background with 88 male (49%) and 93 female (51%) students. All respondents owned personal laptops and have access to internet in campus. The survey includes several items about game habits among the students. The items are i) experience with online game play, ii) Reasons to play games, iii) Preferred games genre and iv) Agreement of using game to learn. Result and discussion for game habits are shown in Fig. 4, 5, 6 and 7, respectively.

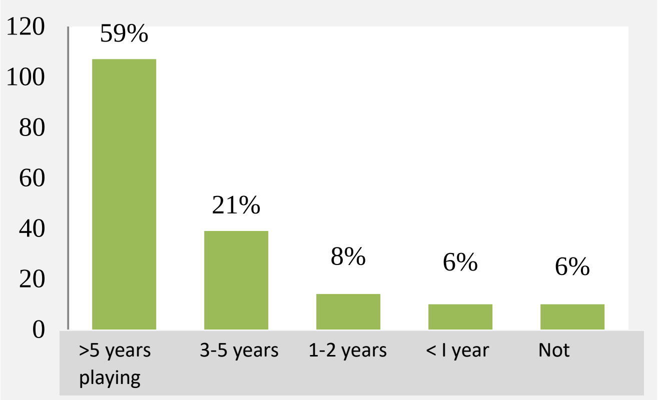


Fig. 4. Online game experience among students.

Fig. 3 shows result for online game play experience. The figure is about the frequency of respondents and their game experience in years. Almost 60% or more than of the students have played game for more than 5 years with only about 10% play in for less than a year or not playing. This shown that the students have vast experiences with online gameplay.

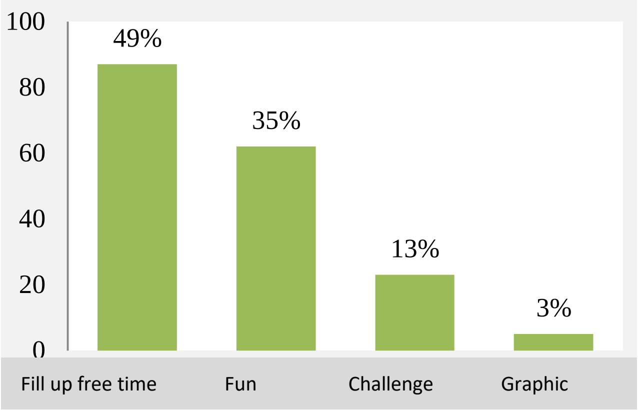


Fig. 5. Reasons to play games.

Most students play games to fill up their free times (49%) while 35% of them play because games are fun, the rest of them play game because it is challenging and have nice graphic.

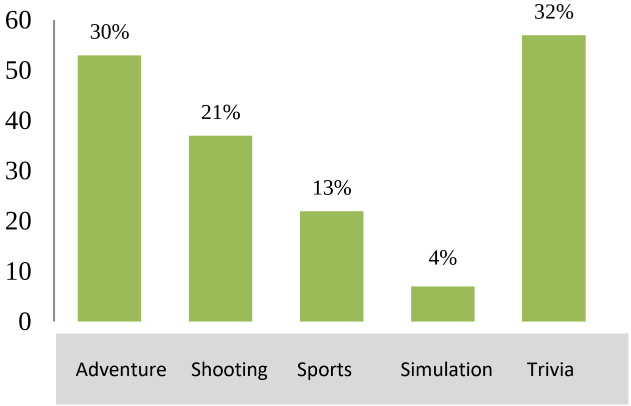


Fig. 6. Preferred games genre.

Fig. 5 shows the most preferred games genre among students. The most popular genres are trivia games (board, card, mini games and the like) and adventure games. Shooting is also the popular genre followed by sports and simulation games. Trivia games can be popular because it is easy to play and have a short time to finish the games. It is also easy to open due to small file size and does not need require lots of preparation in order to play the games.

On the student's agreement to use games for learning purposes, Fig. 6 shown the findings. More than 80% of them agree to use game for their learning with equal number between male (48%) and female (52%) students. The rest of the students are not sure about it (12%) while 5% are not agree to use games for learning. Refer Fig. 7.

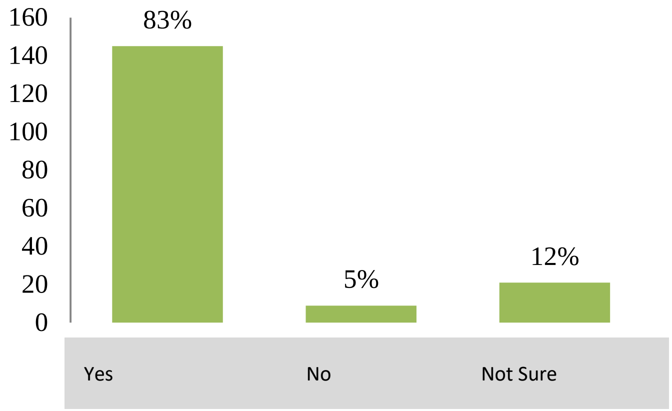


Fig. 7. Agreement to use game for learning.

The following paragraph discusses the findings for descriptive analysis of the items. Table II shows number of items for each component together with value for Cronbach Alpha reliability analysis. All constructs have value of more than 8 except for self-efficacy and enjoyment. Overall value for all constructs is .849.

TABLE II. NUMBER OF ITEMS AND RELIABILITY VALUE

Components	*Item Code*	*No. of Items*	*Cronbach Alpha Value*
Usefullness	Use	4	.842
Ease of Use	EoU	4	.845
Attitude	Att	4	.815
Self-efficacy	SE	4	.605
Anxiety	Anx	4	.887
Enjoyment	Enj	4	.735
Intention to Use	ITU	3	.811
Overall Items		**27**	**.849**

The following paragraph present the descriptive result of study based on constructs and items. Table III shows the constructs and items use in this study. Total of 27 items with 7 constructs were used for data collection.

Table IV shows the result based on mean value together with percentage for every item. Items range from strongly disagree (1) until strongly agree (5). The mean value and percentage of response is presented in the table. For construct Usefulness, almost 75% found the game is useful in their study while about 76% agreed the game can increase their learning productivity. About similar percentage also agreed the games can help them to learn the subject more quickly. Similarly, result for Ease of Use is also positive with overall of 75% respondents agreed that the game is easy to learn. Despite the game is new to them, most of the student agreed that they can use and interact with the game by themselves (more than 70%).

TABLE III. NUMBER OF ITEMS AND RELIABILITY VALUE

Code	**Items**
Construct: Usefulness	
Use1	I find the educational game useful in my study.
Use2	Using the educational game enables me to learn the subjects more quickly.
Use3	Using the game increases my learning productivity.
Use4	If I use the educational game, I will increase my chances of getting good grade.
Construct: Ease of Use	
EoU1	My interaction with the games is clear and understandable.
EoU2	It would be easy for me to become skillful at using the games.
EoU3	I would find the games easy to use.
EoU4	The game is easy to learn.
Construct: Attitude	
Att1	Using the game is a good idea.
Att2	The games make C Programming Subject more interesting.
Att3	Learning with educational games is fun.
Att4	I like learning with the educational games.
Construct: Self Efficacy	
I can learn the subject using the games…	
SE1	If there was no one around to tell me what to do as I go.
SE2	If I could call someone for help if I got stuck.
SE3	If I had a lot of time to learn the content of the games.
SE4	If I had just the built- in help facility for assistance.
Construct: Anxiety	
Anx1	I feel apprehensive about using the games
Anx2	It scares me to think that I could lose a lot of information using the games by hitting the wrong key.
Anx3	I hesitate to use the games for fear of making mistakes I cannot correct.
Anx4	The game is somewhat intimidating to me.
Construct: Enjoyment	
Enj1	When using the educational games, I will not realize the time elapsed.
Enj2	Using educational games will give enjoyment for my learning.
Enj3	Using educational games will stimulate my curiosity.
Enj4	Using educational games will lead to my exploration.
Construct: Intention to Use	
ITU1	I intend to play educational games in the future.
ITU2	I predict I will play educational games in the future.
ITU3	I plan to play educational games in the future.

Students were found to have a very positive attitude towards the use of educational game with almost 85% agree that learning with educational game is fun and is a good idea (81%). This is a positive finding about the idea of using games as their future educational medium. Observation during data collection session also shows that students are highly attached to the computer screen playing the game throughout the session. They are also enthusiast to use the game to learn the subject. For self-efficacy, students were found able to use the games without much help from outside. More than half agreed that they will be able to learn the subject better if given more times to play the games. Overall students were found having acceptable level of self-efficacy despite the newly introduce application to them.

Anxiety is the fear that user may have while using the application. Generally, students were found not apprehensive in using the game, they were also not fear about making any mistake while using the game and they found that the game is not at all intimidating to them. Students are also enjoying in using the game with more than 75% found the game will provide enjoyment in their learning and stimulate their curiosity. This is a good indicator for learning since the game promotes fun aspect as well as make the students curious about the subject the are learning. Conventional way of learning is somewhat boring to these so-called digital native generations [10]. Therefore, use of educational game can be the learning approach of the future due to the fun features of games technology.

TABLE IV. RESULT OF DESCRIPTIVE ANALYSIS

Item Code	Mean Value	Strongly disagree	Disagree	Not Sure	Agree	Strongly agree
Use1	3.91	-	3.3%	22.1%	55.8%	18.8%
Use2	3.87	-	2.2%	25.6%	55.0%	17.2%
Use3	3.88	-	3.9%	20%	60%	16.1%
Use4	3.78	-	3.3%	32.2%	47.8%	16.7%
EoU1	3.81	-	4.4%	25.6%	54.4%	15.6%
EoU2	3.81	-	2.8%	26.1%	58.9%	12.2%
EoU3	3.91	-	2.8%	21.7%	57.2%	18.3%
EoU4	3.93	-	2.8%	20.0%	58.3%	18.9%
Att1	4.03	-	2.8%	17.8%	57.8%	23.3%
Att2	3.95	-	2.8%	20 %	56.7%	20.6%
Att3	4.08	-	1.7%	14.4%	57.8%	26.1%
Att4	3.98	-	3.3%	21.1%	50%	25.6%
SE1	3.58	0.6%	7.8%	36.1%	44.4%	11.1%
SE2	3.60	0.6%	7.2%	31.1%	53.9%	7.2%
SE3	3.60	-	8.9%	29.4%	54.4%	7.2%
SE4	3.59	1.1%	5.6%	33.9%	51.7%	7.8%
Anx1	2.47	13.9%	41.7%	27.8%	16.7%	-
Anx2	2.51	13.9%	40%	30%	13.9%	2.2%
Anx3	2.44	17.2%	37.8%	30.6%	12.8%	1.7%
Anx4	2.12	27.8%	41.1%	22.2%	8.9%	-
Enj1	3.56	0.6%	8.3%	35%	46.7%	9.4%
Enj2	3.85	-	2.8%	23.9%	58.9%	14.4%
Enj3	3.89	-	1.1%	22.8%	61.7%	14.4%
Enj4	3.93	-	2.2%	20.6%	58.9%	18.3%
ITU1	4.02	-	1.1%	17.2%	60%	21.7%
ITU2	4.02	-	-	16.7%	65%	18.3%
ITU3	4.02	-	1.1%	17.2%	60%	21.7%

V. DISCUSSIONS AND CONCLUSION

This study developed and evaluates an online educational game for self-learning of programming concept. From the survey conducted, it was found that students show a highly positive attitude towards the game despite not much of exposure to such technology prior his study. The game is also found useful for them despite the lack of content due to prototype version. Students were also thinking that the game is easy to use even though they were only introduced to the game for about 30 minutes to one hour. This is a good sign that the students are rather well versed with the games technology with more than half of them have experience of more than 5 years with games.

Among the interesting findings is student's attitude towards game. They agreed that using the game is a good idea and making the learning of subject more interesting. The game is also fun and generates their interest to learn. At the same time, they also think that they can use the game by themselves without much help needed to assist them. Students were also found not feeling anxious in using the game. They can play the games without much fear to lose information or making mistakes. Programming subject is important for any computer science students as well engineering students, but it was found as a boring subject and many students struggle to learn the subject in many studies, therefore some enjoyment element is needed in order to make more interesting and fun.

Game is one of the technology full of fun aspect, thus we proposed games for learning. For enjoyment aspect, students agreed games provide enjoyment for their learning as well stimulates their curiosity to learn. Game were also found lead them to more exploration. This is a good sign of games and prompt more studies for better conclusion. In general, students have a very good intention to use games in the future with more than 80 per cent stated agree to use the application for their learning. Therefore, game is among the application for new way of learning, however many more studies needed to ensure all aspects of educational game design, development and effectiveness were thoroughly investigated and known.

Future works will be conducting evaluation of the games usability features as well as its effectiveness in improving student' knowledge on Programming Introductory by doing pre-test and post-test evaluation. We hope to provide more understanding and information on how games can helps students to learn especially in Malaysia context.

ACKNOWLEDGMENT

This study is funded by research grant under Universiti Teknologi Malaysia (UTM) Research University Grant (RUG) with Vote No. 18H63. Thank you also to the Ministry of Higher Education Malaysia and UTM Research Management Team for the continuous supports.

REFERENCES

[1] Nor Azan, M. and S. Wong, Game Based Learning (GBL) Model for History Courseware: A Preliminary Analysis, in International Symposium on Information Technology (ITSIM). H.e.a. (Eds.), Editor. 2008, UKM: Kuala Lumpur, Malaysia. p. 253-260.

[2] Wong, S.Y., Reka bentuk dan penilaian permainan pendidikan multimedia interaktif Sejarah (PPMIS), in Fakulti Teknologi dan Sains Maklumat. 2012, Universiti Kebangsaan Malaysia: Bangi.

[3] Roslina, I., C.Y. Rasimah, O. Hasiah, and J. Azizah, Students Perceptions of Using Educational Games to Learn Introductory Programming. Computer and Information Science, 2011. 4(1): p. 205 - 216.

[4] Becker, K., Video Games pedagogy: Good Games = Good Pedagogy in Lecture Notes in Computer Science S. Link, Editor. 2008, Springer Verlag Heidelberg. p. 73-125.

[5] Fedwa Laamarti, Mohamad Eid, and A.E. Saddik, An Overview of Serious Games. International Journal of Computer Games Technology, 2014. 2014: p. 1-15.

[6] Rozana, I. and I. Roslina. PDEduGame: Towards Participatory Design Process for Educational Game Design in Primary School. in 5th International Conference on Research and Innovation in Information Systems. 2017. Langkawi: IEEE.

[7] Roslina Ibrahim, Suraya Masrom, Rasimah C.M Yusoff, N.M.M. Zainuddin, and Z. Rizman, Student Acceptance of Educational Games in Higher Education. Journal of Fundamental and Applied Sciences, 2017. 9(3a): p. 809-829.

[8] Fedwa, L., E. Mohamad, and E.S. Abdulmotaleb, An Overview of Serious Games. International Journal of Computer Games Technology, 2015. 2014: p. 1-15.

[9] Trevi, G.N. and C. Pomales-García. How can a serious game impact student motivation and learning? in In Industrial and systems engineering research conference. 2014. Montreal. Norcross: IIE.

[10] Prensky, M., Digital Game-Based Learning. 2001, New York: Mc Graw Hill.

[11] Nacke, L., Facilitating the education of game development, in Department of Computer Science. 2004, Otto-von-Guericke University Magdeburg: Magdeburg.

[12] Kirriemuir, J. and A. McFarlane, Literature review in games and learning, in Futurelab Series, Futurelab, Editor. 2004, University of Bristol: Bristol.

[13] Gee, J.P., What video games have to teach us about learning and literacy. 2003, New York: Palgrave MacMillan.

[14] Gee, J.P. Good Video Games and Good Learning. 2006 10 September 2010].

[15] Oblinger, D.G., Games and learning :Digital games have the potential to bring play back to the learning experience. Educause quarterly 2006(3): p. 5-7.

[16] Olga, S., V. Pavel, K. Alexander, and T. Alexey, Game based approach in IT education. International Book Series "Information Science and Computing", 2009. 12: p. 63-70.

[17] Martinovic, D., Ezeife, C. I., Whent, R., Reed, J., Burgess, G. H., Pomerleau, C. M.. Critic-proofing of the cognitive aspects of simple games. Computers & Education, 2014. 72(2014): p. 132-144.

[18] Baniqued, P.L., Lee, H., Voss, M. W., Basak, C., Cosman, J. D., DeSouza, S, Selling points: what cognitive abilities are tapped by casual video games? Acta Psychologica, 2013. 142(1): p. 74-86.

[19] Yolanda, A.R., M. McKenzie, W.S. Marcus, and G. Bruce, User centered game design: evaluating massive multiplayer online role playing games for second language acquisition, in Proceedings of the 2008 ACM SIGGRAPH symposium on Video games. 2008, ACM: Los Angeles, California.

[20] David, P., W. Nelson, and S. Tadeusz, Heuristic evaluation for games: usability principles for video game design, in Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems. 2008, ACM: Florence, Italy.

[21] F.A.S. Harnessing the power of video games for learning. 2006 8 May 2010].

[22] Song, M. and S. Zhang, EFM: A Model for Educational Game Design, in Lecture Notes in Computer Science, S. Link, Editor. 2008, Springer US. p. 509-517.

[23] Hsu, C.-L. and H.-P. Lu, Consumer behavior in online game communities: A motivational factor perspective. Computers in Human Behavior, 2007. 23(3): p. 1642-1659.

[24] Shin, D.H. and Y.J. Shin, Why do people play social network games? Computers in Human Behavior, 2011. 27(2).

[25] Chuang, T.Y., Chen, W.F, Effect of Computer-Based Video Games on Children: An Experimental Study. Educational Technology & Society, 2009. 12(2): p. 1-10.

[26] Kuang Chao Yu, H.S.H., Fu Hsing Tsai, The Implementation and Evaluation of Educational online gaming system. IEEE, 2005.

[27] Hye, S.K.S., Baeg Kim. An Integrated Course Based on Educational Games. in Proceedings on the International Conference in Information Technology: Coding and Computing (ITCC'05). 2005. IEEE Explore.

[28] Teh, C.L., M.F.W.I. Wan, and S.C. Toh, Why use computer games for learning?, in 1st International Malaysian Educational Technology Convention (IMETC). 2007: Johor, Malaysia. p. 835-843.

[29] Norizan, M.D., Pendekatan Bercerita dan Permainan dalam Pembangunan Perisian Kursus Akhlak Islamiah. 2003, UKM: Bangi.

[30] Vansiri Nagalingam and R. Ibrahim. Finding the Right Elements User Experience Elements for Educational Games. in ICEEG 2017. 2017. Turku, Finland: ACM.

[31] Alzahrani, A.I., I. Mahmudb, T. Ramayah, O. Alfarraj, and N. Alalwan, Extending the theory of planned behavior (TPB) to explain online game playing among Malaysian undergraduate students. Telematics and Informatics, 2017. 34(2017): p. 239-251.

[32] Eow, Y.L., W.Z.W. Ali, R. Mahmud, and R. Baki, Form one students' engagement with computer games and its effect on their academic achievement in a Malaysian secondary school. Computers & Education, 2009. 53(2009): p. 1082-1091.

[33] Eow, Y.L., W.Z.b.W. Ali, R.b. Mahmud, and R. Baki, Computer games development and appreciative learning approach in enhancing students' creative perception. Computers & Education, 2010. 54(1): p. 146-161.

[34] Izam Shah, B., Perisian Pengembaraan Multimedia :Edutainment Dalam Pendidikan Agama Islam Sekolah Rendah, in Fakulti Teknologi dan Sains Maklumat. 2007, Universiti Kebangsaan Malaysia: Bangi, Malaysia.

[35] Rubijesmin, A.L. Understanding Malaysian students as gamers: Experience. in Proceedings of the 2nd International Conference on Digital interactive Media in Entertainment and Arts 2007. Perth, Australia: ACM.

[36] Wong, S.Y. and S. Ghavifekr, User experience design of history game: An analysis review and evaluation study for Malaysia context. International Journal of Distance Education Technologies, 2018. 16(3): p. 46-63.

[37] Zuhaira, M.Z., Pembangunan Perisian Permainan Multimedia untuk Sains Prasekolah: Dam Cuaca. 2007, UKM: Bangi.

[38] Jaspaljeet, S., L.W. Ling, S. Mohana, S.G. Saraswathy, and K.D. Siva, Designing Computer Games to Introduce Programming to Children, in Information Technology and Multimedia at UNITEN (ICIMU' 2008), Uniten, Editor. 2008: Selangor, Malaysia. p. 643-647.

[39] Noor Azli, M., M. Nor Azan, and C. Shamsul Bahri. Digital Games Based Learning. in International Symposium on Information Technology (ITSIM). 2008. Kuala Lumpur, Malaysia.

[40] Rosas, R., M. Nussbaum, P. Cumsille, V. Marianov, M. Correa, et al., Beyond Nintendo, design and assessment of educational video games for first and second grade students. Elsevier Computers and Education, 2003. 40: p. 71-94.

[41] Squire, K., M. Barnett, J.M. Grant, and T. Higginbotham. Electromagnetism Supercharged! in International Conference of the Learning Sciences 2004. 2004. Los Angeles.

[42] Natvig, L.L., Steiner. Age of Computers − Game-Based Teaching of Computer Fundamentals. in ITiCSE. 2004. Leeds, United Kingdom: ACM.

[43] Coles, C.D., Dorothy C. Strickland, Lynne Padgett, and L. Bellmoff, Games that ''work'': Using computer games to teach alcohol-affected children about fire and street safety. Science Direct Research in Developmental Disabilities, 2007. 28: p. 518-530.

[44] Barnes, T., E. Powell, A. Chaffin, A. Godwin, and H. Richter. Game2Learn: Building CS1 Learning Games for Retention. in ITiCSE. 2007. Dundee, Scotland: ACM.

[45] Wong, W.L., S. Cuihua, N. Luciano, C. Eduardo, T. Fei, et al. Serious Video Game Effectiveness. in International Conference on Advances in Computer Entertainment Technology. 2007. Salzburg, Austria.: ACM.

[46] Garzotto, F. Investigating the Educational Effectiveness of Multiplayer Online Games for Children. in Interaction Design and Children (IDC 2007). 2007. Aalborg, Denmark: ACM Press.

[47] Roslina, I. and J. Azizah. Using educational games for learning introductory programming: initial study on student perceptions. in IADIS International Conference Game and Entertainment Technologies 2010. 2010. Freiburg Im Breisgau, Germany.

[48] Wong, S., Reka bentuk dan penilaian permainan pendidikan mutimedia interaktif Sejarah (PPMIS) 2012, Universiti Kebangsaan Malaysia: Bangi, Selangor.

[49] Hirumi, A. and C. Stapleton, Applying pedagogy during game development to enhance game based learning, in Book Technologies for E-Learning and Digital Entertainment 2008, Springer Verlag Berlin Heidelberg. p. 509-517.

[50] Overmars, M. and J. Habgood, The Game Maker's Apprentice: game development for beginners. 2006, Berkeley, California: Apress. 86-91.

[51] Leemkuil, H., Is it all in the game? Learner support in an educational knowledge management simulation game. 2006, University of Twente.

[52] Fisch, S.M., Making Educational Games "Educational", in Conference on Interaction Design and Children IDC 2005, ACM, Editor. 2005, ACM: Colorado, USA. p. 56 - 61.

[53] Fu, F.L., R.C. Su, and S.C. Yu, EGameFlow: A scale to measure learners' enjoyment of e-learning games. Computers and Education, 2009. 52: p. 101-112.

[54] Novak, J., Game Development Essentials: Second Edition. 2008, New York: Thomson Delmar Learning.

[55] Grassioulet, Y., A Cognitive Ergonomics Approach to the Process of Game Design and Development. 2002, University of Geneva.

[56] Davis, F.D., Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Quarterly, 1989. 13(Sep 1989): p. 319-340.

[57] Venkatesh, V., M.G. Morris, G.B. Davis, and F.D. Davis, User acceptance of information technology: Toward a unified view. MIS Quarterly, 2003. 27(3): p. 423 - 478.

[58] Bourgonjon, J., M. Valcke, R. Soetaert, and T. Schellens, Students' perceptions about the use of video games in the classroom. Computers and Education, 2010. 54(2010): p. 1145-1156.

[59] Davis, F.D., User acceptance of information technology: system charateristics, user perceptions and behaviroal impacts. International Journal of Man Machine Studies, 1993. 38(1993): p. 475-487.

[60] Ha, I., Y. Yoon, and M. Choi, Determinants of adoption of mobile games under mobile broadband wireless access environment. Information & Management, 2007. 44(3): p. 276-286.

Accident Detection and Smart Rescue System using Android Smartphone with Real-Time Location Tracking

Arsalan Khan, Farzana Bibi, Muhammad Dilshad,
Salman Ahmed, Zia Ullah
Department of Computer Science and Software Engineering
Al-hamd Islamic University, Islamabad, Pakistan

Haider Ali
Department of Pharmacy (Pharm-D)
Sarhad University of Science and Technology, Peshawar, Pakistan

Abstract—**A large number of deaths are caused by Traffic accidents worldwide. The global crisis of road safety can be seen by observing the significant number of deaths and injuries that are caused by road traffic accidents. In many situations the family members or emergency services are not informed in time. This results in delayed emergency service response time, which can lead to an individual's death or cause severe injury. The purpose of this work is to reduce the response time of emergency services in situations like traffic accidents or other emergencies such as fire, theft/robberies and medical emergencies. By utilizing onboard sensors of a smartphone to detect vehicular accidents and report it to the nearest emergency responder available and provide real time location tracking for responders and emergency victims, will drastically increase the chances of survival for emergency victims, and also help save emergency services time and resources.**

Keywords—Traffic accidents; accident detection; on-board sensor; accelerometer; android smartphones; real-time tracking; emergency services; emergency responder; emergency victim; SOSafe; SOSafe Go; firebase

I. Introduction

The number of deaths due to traffic accidents is very high. Looking at the number of deaths and injuries due to road traffic accidents shows the global crisis of road safety. Nearly 1.3 million people are killed every year and about 50 million injured worldwide due to road accidents, which averages to 3,287 lives lost every day. More than 50 percent of road traffic deaths affect young adults between the age of 15-44. Around 400,000 individuals under the age of 25 dies in road traffic accidents every year. Even in countries with very good road safety measures, the number of road accident deaths is getting higher every year [1]. More than 90% of road traffic deaths occur in middle-income countries. In low-income countries the figure is even higher.

In Pakistan the last 10 year of statistics shows that an average of 15 individuals lost their lives due to traffic accidents daily. According to data from Pakistan Bureau of Statistics on traffic accidents in Pakistan from 2004 to 2013 [2], the overall deaths in road accidents are about 55 percent, which according to the specialists is very high. According to the data, total 51,416 individuals died in 97,739 road accidents across the country. Furthermore, the data shows that deaths per accident are 55 per cent around the country [3].

The most likely reason for an individual's death in an accident is lack of the first aid provision that is because of emergency services not receiving information about accident in time. Emergency response time is extremely vital when it involves incidents involving vehicle accidents. Analysis shows that if we decrease just 1-minute in accident response time that can increase chances of saving an individual's life up to six percent [4]. In order to reduce response time, implementation of enhanced traffic technologies would be necessary, which will help scale back response time and therefore reduce fatalities.

The purpose of this research is to design and implement such an automated system that uses smartphone to detect vehicle accidents and report it to the nearest available responders to help counter these emerging problems and reduce casualties as much as possible. The detection system would help reduce fatalities due to vehicle accidents by decreasing the response time of emergency services. The system will also provide other emergency services like Fire Brigade, Police Department and Medical emergency services.

In this work we are utilizing android smartphone to detect accidents and report it to the nearest available emergency responders with the exact location of victims in emergency. On an emergency responder side, the system will inform responders about the incidents that occur near to them and provide them with real time tracking of emergency victims on a Google map. This will help emergency responders keep track of victim's location and rescue them as soon as possible.

This paper is organized as follows: Section II describes the related works that has been done in the field, Section III presents various technologies that are utilized in our system, Section IV describes questions related to this work, Section V presents the architecture and implementation of the proposed system, Section VI presents front end design of the proposed system, Section VII presents backend design of the system, Section VIII describes performance results and tests performed of the proposed system, Section IX presents contributions of this work and finally the last Section X is the conclusion and future work for the proposed system.

II. Related Work

Using smartphones to identify road traffic accidents is not a new subject. There are completed algorithms for systems which utilizes accelerometer as well as GPS to detect vehicle accidents using smartphones to detect accidents dates back to 2011. Because there is already a lot done on this subject, what we decided to do was to develop a complete system that is more reliable and have much more functionality than the existing ones, designed for the ongoing project in mind.

In [5] the authors developed a system which used Android smart-phones and ODB-II connection in a vehicle. When the system detects an accident, will sends an SMS to emergency contacts specified by the user, SMS will contain information about the accident and also a call to the emergency services is made automatically. All modern vehicles have ODB-II connection installed which transmits data about the vehicle in real-time such as acceleration, oil pressure, speed, etc. For the system to work a vehicle must support OBD-II standard. In US and this standard is necessary since 2001, European countries have also implemented a version of this standard, so vehicles in the US and in Europe can use this system and is not available to all vehicle in other countries. Other than that, upgrading and maintenance of this system is very expensive process [6].

In [7], the authors at the University of Baghdad Iraq developed a system which made use of the accelerometer, GPS and microphone to detect accidents. Upon detection of an accident sends an emergency notification to the web server and also sends an SMS to the emergency contacts, emergency responders have to access the web server to find out about an accident. Their system made use of the same sensors and hardware that the algorithm presented in this research work makes use of, except for a few features. The main issue with this system is that the notifications are sent to a web server and responders needs to check the web server for accident notification, there is no system for individual responder that responds to the emergency to track victim's location and also the system lacks the functionality to send emergency notification to the nearest emergency center in case there is more than one emergency center in the area.

In [8], the authors developed a system called WreckWatch which involves reading data from the accelerometer and acoustic data from the microphone to detect accidents. If an incident has occurred, the application contacts nearby emergency services and provides GPS-coordinates of accident location.

In [9], the authors have developed an android application that uses accelerometer sensor to detect accidents. After sensing the accident, application automatically sends a voice message to 108 ambulance emergency response service that is running in India. The issue is that this system is for specific emergency response services, only applicable in India. Also, the system is prone to increased false positives because there is no filter in place to verify if an accident detected by the smartphone is a real accident or just false alarm due to dropping smartphone, etc.

In this study we looked at various technologies and existing systems providing us with broad analysis and helped us in developing our system. From analysis we found that these systems can play a very important role in saving human lives. A new system is to be developed based on unique features that will help counter emergencies.

III. Technologies Utilized

A. Smartphones

Smartphones are mobile phones that have considerably a lot of functionality than a regular mobile phone. they're mobile computers. Smartphones are powerful and versatile as a result of built in sensors, powerful processors, multiple network interfaces and a high amount of memory for such small devices.

B. Android

Android is a Linux kernel based open-source mobile operating system which was developed by Google for phones, tablets, watches, TVs, cars and other electronic devices [10]. Being open-source, everyone has full access to the Android source code, with one restriction, it cannot be used for personal profit or any financial gain. It is the most popular mobile operating system.

Android alternatives include iOS by Apple, Windows Phone, BlackBerry, Symbian and a few others. We chose Android as it is the operating system that have the most programming experience with. Android's market dominance and cheap application release costs were also deciding factors.

C. Android Studio

Android Studio is the primary Android IDE (Integrated Development Environment). It provides an Android developer all the necessary tools to develop an Android application. More specifically, it allows writing code with auto-completion tools, debugging, testing [11], running the code on a physical or a virtual device and setting programming related or visual preferences. Java and XML are the only languages required to create Android applications with Android Studio [12]. Android Studio does not have any alternatives worth considering. It is possible to develop Android applications with Eclipse by using the Android Developer Tools plugin, but it is no longer supported by Google [13].

D. Java

Java is a class based general-purpose, object-oriented programming language [14]. It is a high-level, strongly typed language with garbage collection that incorporates concepts from several languages including C and C++, but it is not entirely the same. For example, Java does not allow writing unsafe code that might cause vulnerabilities and unexpected behavior. The main building blocks of a Java application are classes, interfaces and packages.

E. Accelerometer

An accelerometer works by detecting proper acceleration affecting the accelerometer to determine the G-forces affecting the accelerometer [15]. Proper acceleration means acceleration that is relative to free-fall [16]. An object in free-fall would as such have no acceleration affecting it while an object at rest on the surface of the earth would experience an acceleration of 9,81 m/s2 upwards due to the surface pushing the object upwards to negate gravity. Accelerometers in smartphones bases their functionality on micro-electromechanical-systems (MEMS), which measure electric currents based on compression of a seismic mass, often silicon, caused by acceleration [17].

F. Google Play Services

Google Play services provide application developers a comprehensive set of useful features, for example, Maps and Google+ sign-in. The services include the Google Play services client library and the Google Play services Android Package Kit. The client library makes it possible to access any feature with a user's account and deals with different issues that may occur when using the services. The Android Package Kit communicates with the client library and provides access to a specific service when necessary. The use of Google Play services is a must when using Firebase. Important functionalities in Smart Rescue System, for example, viewing on a map and obtaining a user's location also rely on the services.

G. Google Location API

The Google Location Services API [18] is part of Google Play Services, provides a more robust, high-level framework that automatically chooses a suitable location provider and power management. Location Services also provides new features like activity detection which is not provided by framework API. Developers should consider using Location services API if they are using framework API and also if they are making their apps location –aware.

H. Android Google Map API

The Google Maps Android API is a service which is part of the Google Play services library. Allows access to Google Maps server automatically, displaying map, downloading data, and map gesture response. It also allows to add markers, polygons, and basic map overlays, and to transition the user's context of a specific map area [19].

I. Google Places API Web Service

The Google Places API Web Service is a service which returns information about places like locations, geographic, establishments and prominent points of interest using HTTP requests [20]. The main alternative of Google Places API is Foursquare Venues. In free version Google Places allows 150000 queries per day [21] and Foursquare Venues allows 120 000 queries per day [22]. These two services are mostly similar, we chose Google Places API because we were more familiar and experienced using it.

J. Google Directions API

The Google Directions API [23] is a service that uses HTTP request to calculate distance between locations. When calculating directions, the API returns the most economic routes. The API decides which route is most efficient on the basis of travel time, number of turns, distance, etc.

K. Retrofit

Retrofit [24] is a type-safe HTTP client for Android and Java, provided by Square. Retrofit makes it easy to communicate with a web server and get back data, as java objects, it automatically supports a large collection of response types, including converting JSON objects into Plain Old Java Objects. Alternative of Retrofit is Googles Volley [25], which is also a HTTP library, for our system we chose Retrofit because it is light weight and has more documentation.

L. Backend as a Service

Backend as a Service (BaaS) is a hosted backend that has been premade for developing a web or mobile application. Developers do not have to write any or much backend specific code. It has all the necessary features of a backend and even more features such as Facebook and Google sign-in integration and cloud messaging are common. The features can be accessed by documented APIs that simplify the application development process.

M. Firebase

Firebase is one of many implementations of the BaaS model. Like other BaaS implementations, Firebase provides storage, push notifications, user authentication and a database. Other than the basic BaaS features, Firebase also give a test lab that permits testing a Firebase connected application with different configurations and devices. A feature that makes Firebase different from other BaaS implementations is the real-time database. When new data is added to the database, it becomes accessible instantly to all the users of the application.

N. Firebase Cloud Messaging

FCM (Firebase Cloud Messaging) is another adaptation of GCM (Google Cloud Messaging). It is a cross-platform messaging solution that allows us to reliably deliver and receive messages without any cost. Using FCM, we can send notification messages in order to re-engage users. [26].

O. GeoFire

GeoFire is an open-source library for Android/Java that allows us to save and query a set of keys according to their geographic location. At its core, GeoFire simply saves locations with string keys. Its main advantage however, is that it allows querying key in a specific geographic area in real time [27].

IV. Research Questions

Sr.no	Research Questions	Motivation
1	What are the benefits of such system?	This question will elaborate the pros and cons of accident detection system.
2	What are the research contributions?	This research aims to provide a complete system for both emergency victims and emergency responders.

V. System Overview

The designed system consists of:

- **SOSafe:** An android application for emergency victims.
- **SOSafe Go:** An android application for emergency responders.
- **Firebase:** as Server, Database, File Storage, Cloud Messaging, Auth.

A. Used Services

a) Google Play services

b) FCM (Firebase Cloud Messaging) services.

c) Firebase Realtime Database

d) Firebase Authentication

e) Firebase Storage

f) Google API Client

g) Google Maps API

h) Google Location API

i) Google Direction API

j) Google Places API

k) Retrofit

B. Use Case Diagrams

1) SOSafe

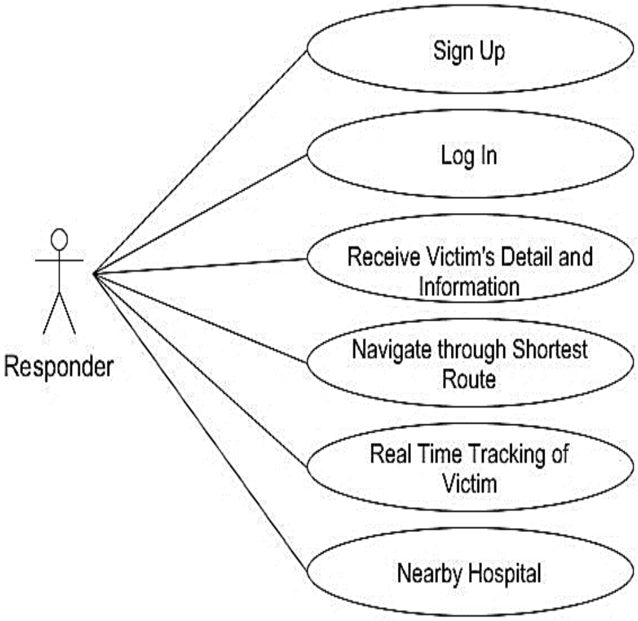

Sign Up
Log In
Receive Victim's Detail and Information
Navigate through Shortest Route
Real Time Tracking of Victim
Nearby Hospital
Responder

Fig. 1. SOSafe use case.

0 shows the Top-Level Use Case of SOSafe, which indicates the user's full interaction with the system. It shows the user is firstly registering himself, user can then log in to the system using firebase email and password authentication. He/she can view/update his profile at any time after authenticated. The user can turn on Automatic Monitoring which will register an accelerometer service running in the background, it can now detect all kinds of jerks produced by the user on his/her smartphone and correctly differentiate between accidents and normal routine jerks.

Upon detecting the right accidents, the system will generate an alert containing an alarm sound on the emergency victim's phone. users will be able to cancel sending alert to emergency responders in case of false alert (accident didn't occurred) under 15 seconds. SOSafe will get victim's location using Google Location API and save it to firebase real time database, then search for nearest emergency responders from victim's location and will send an alert notification to the nearest emergency responder (containing victim's location) using FCM. SOSafe will also send SMS to the emergency contacts containing location of the victim. When emergency responder accepts the request sent by victim, SOSafe will show real time location tracking of responder to the emergency victim on a Google map. SOSafe will also provide details about responders (name, vehicle number, phone, etc.).

In case of other emergencies users can select the type of emergency he/she is in (Fire, Ambulance, Police), then by pressing panic button the system will search and notify nearest responders available for the selected type of emergency.

2) SOSafe Go

The system also consists of an application for emergency responders. Responders will be able to select the type of emergency services they provide and other information related to it. This application will show emergency notifications that are sent by emergency victims and provide real time location tracking of their locations. In case of medical emergencies, the system will also guide responders to a nearest hospital from emergency location.

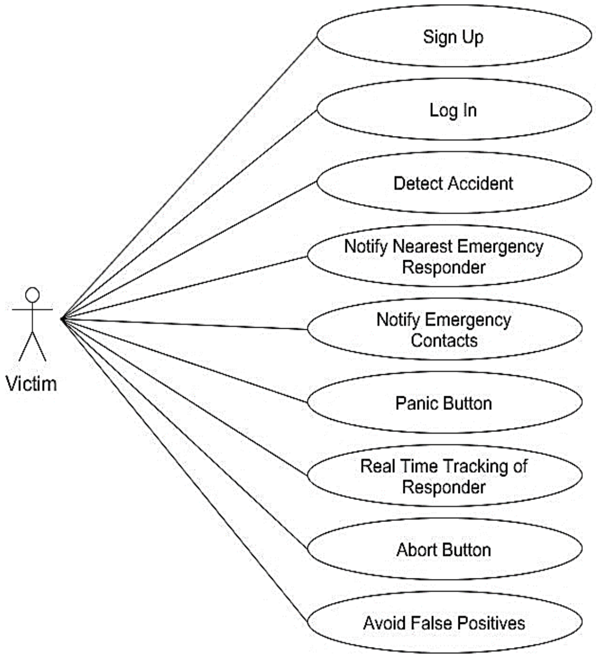

Sign Up
Log In
Detect Accident
Notify Nearest Emergency Responder
Notify Emergency Contacts
Panic Button
Real Time Tracking of Responder
Abort Button
Avoid False Positives
Victim

Fig. 2. SOSafe go use case.

Fig. 2 shows the Top-Level Use Case of SOSafe Go, which indicates the emergency responders' full interaction with the system. It shows the responders firstly registering themselves, responders can then log in to the system using firebase email and password authentication, responders can view/update thier profile at any time after authenticated. Emergency responders can then receive emergency requests from emergency victims using FCM.

When SOSafe Go receives emergency requests, an alert with sound and vibration triggered will be shown with location of the emergency. When requests are accepted by the responders, responders will be able track the location of emergency victims in real time on a Google map with shortest route to location of emergency victims using Google Directions API. SOSafe Go will also show details about victims (name, address, blood group, etc.) when emergency victim is rescued, if victims are in need of medical assistance SOSafe GO will guide responders to the nearest hospital from emergency locations by utilizing Google Places and Google Directions API.

C. *Activity Diagram*

0 is showing the sequence of activities held in the system.

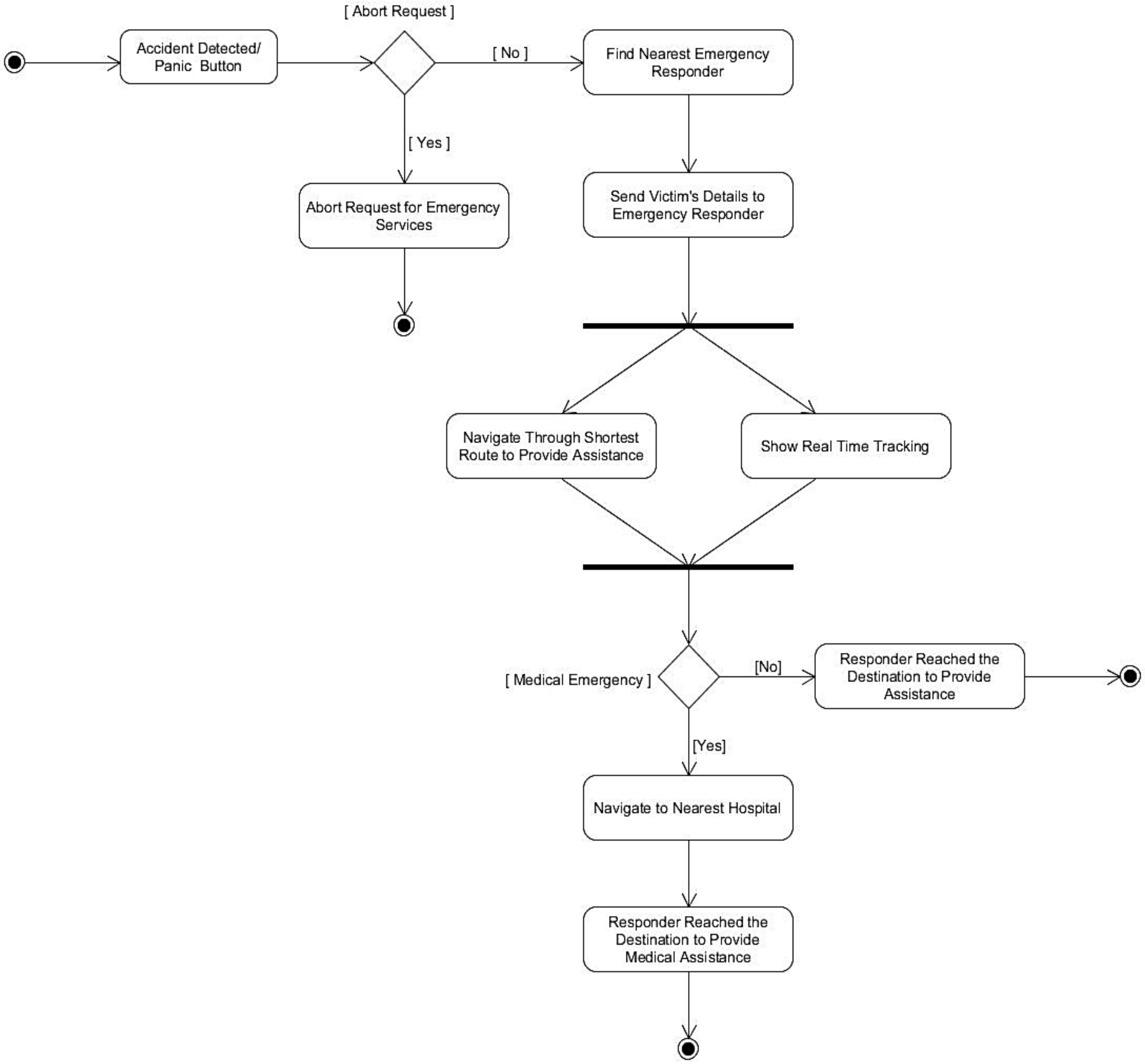

[Abort Request]
Accident Detected/ Panic Button
[No]
Find Nearest Emergency Responder
[Yes]
Abort Request for Emergency Services
Send Victim's Details to Emergency Responder
Navigate Through Shortest Route to Provide Assistance
Show Real Time Tracking
[Medical Emergency]
[No]
Responder Reached the Destination to Provide Assistance
[Yes]
Navigate to Nearest Hospital
Responder Reached the Destination to Provide Medical Assistance

Fig. 3. Overall system activity diagram.

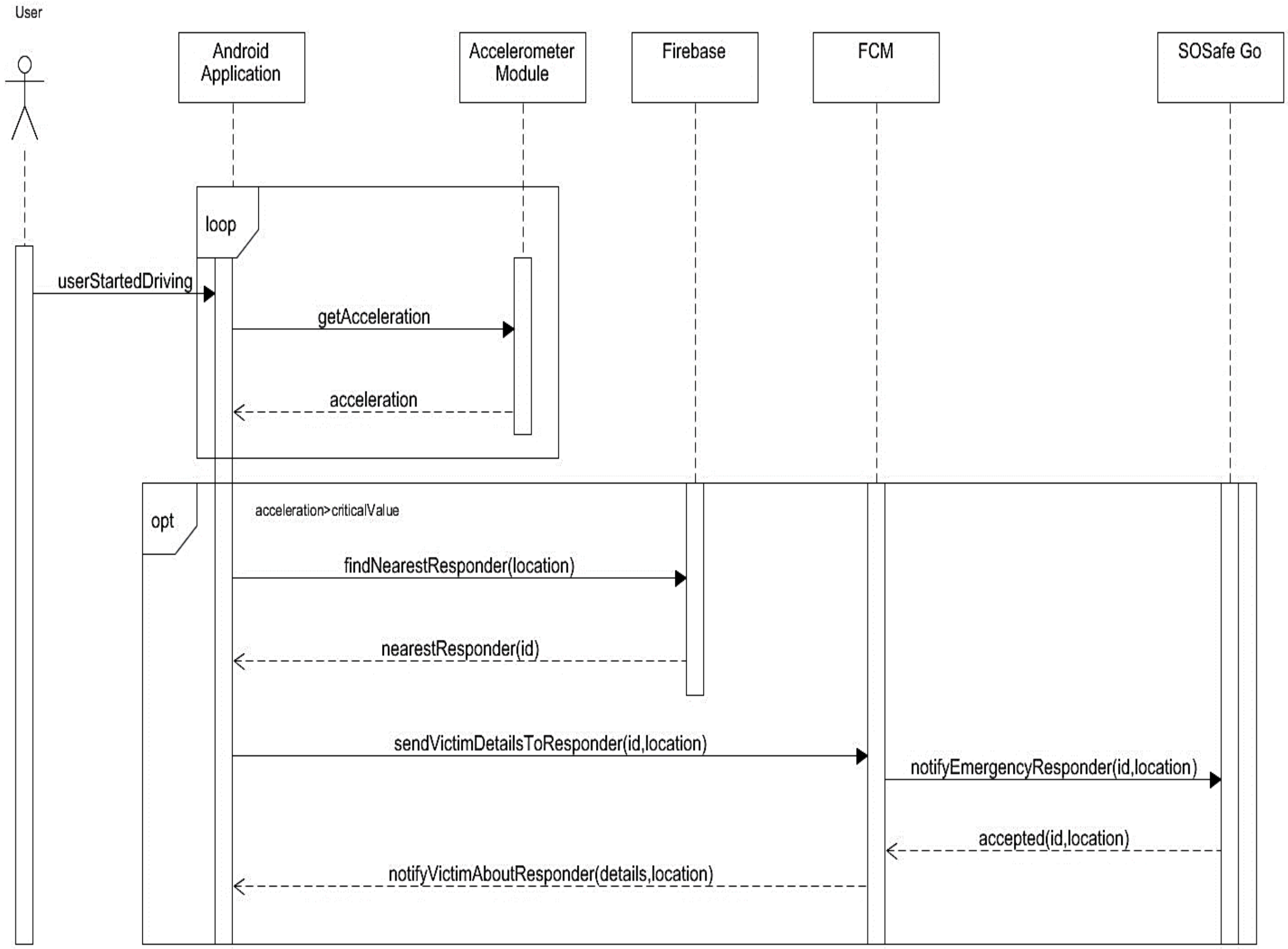

User
Android Application
Accelerometer Module
Firebase
FCM
SOSafe Go
loop
userStartedDriving
getAcceleration
acceleration
opt
acceleration>criticalValue
findNearestResponder(location)
nearestResponder(id)
sendVictimDetailsToResponder(id,location)
notifyEmergencyResponder(id,location)
accepted(id,location)
notifyVictimAboutResponder(details,location)

Fig. 4. System sequence diagram.

D. Sequence Diagram

In 0, sequence diagram shows the sequence in which emergency victims' application, Firebase and responders' application are performing their work.

E. Firebase as Backend

The brain of the whole system is Firebase. Firebase Authentication is used to authenticate users by Email and Password. Firebase Authentication also provides a number of Authentication API's like Google, Facebook, GitHub, etc. All information about emergency victims and emergency responders (availability, location. etc.) are stored at Firebase database. Information about emergency requests sent from emergency victims' side is also stored and is processed by matching attributes of the request to a nearest emergency responder available. A notification message about emergency request is then sent to the available responder through FCM. Firebase is used to store real time location data, using GeoFire library for Firebase. Firebase Cloud Messaging is used to exchange data and send notifications between emergency victim's app and emergency responder's app. The system will also deal with unexpected scenarios like when there is no responder available and avoiding false positive.

F. Algorithm for Accident Detection

Algorithm that uses on-board accelerometer sensor of a smartphone to detect accidents was developed for our system in this research work. The question is that what to do with the values that are being generated by the accelerometer sensor and setting a threshold value that will trigger the accident alert. Accelerometer delivers acceleration values for each of the three axes. Accelerometer values are based on (1).

$$A_D = -g - \left(\frac{\sum F_s}{m}\right)$$

Equation (1) shows that the acceleration values that an accelerometer sensor (A_D) generates is force (F_s) divided by mass (m) which is affected by gravitational acceleration (-g). Acceleration for each axes (ax, ay, az) is derived based on (1).

Now we will use Pythagorean Theorem to derive values from accelerometer as shown in (2).

$$a = \sqrt{a_x^2 + a_y^2 + a_z^2}$$

By using the value of (a) we can calculate the value of g-force (gravitational force). Equation (3) shows how g-force is calculated.

$$G = \frac{a}{g}$$

Equation (3) shows that g-force value is calculated from acceleration (a) divided by gravitational acceleration (G). G-force value will be approximately 1 if smartphone is resting on a table and will exceed 1 if the device is moving. Accidents can be detected by checking if g-force value exceeds a specified threshold, which in our system will be 4g, which then will generate an emergency alert.

G. *Avoiding False Positives*

Since accident detection system that uses smartphone can dispatch emergency services, it is necessary to avoid false alerts. Differentiating between accidents that occurred versus dropping your phone or sudden stop is relatively hard due to smartphone mobility. If the system cannot analyze and avoid false positives precisely, but can make it meaningless by wasting emergency services resources on false incident reports. We have added some features to improve our systems reliability, accuracy and avoid false positives.

1) Acceleration filter prevents false positive:

SOSafe will not trigger emergency event if the G-force value is below 4g. This value can detect accident but will avoid triggering emergency event on dropping a smartphone or sudden stop.

2) Count down timer alert to prevent false positive:

In case the system detects an accident, the system will generate a countdown alert dialog with sound and vibration for 15 sec. in case of false alert (accident didn't occur) the user will be able to cancel sending emergency alert to emergency responder under 15 secs. This will help in reducing false positive, as the user will be able to cancel sending an emergency notification in case of false event.

VI. Front End Design

A. *SOSafe*

Emergency victims' side application SOSafe is developed in Java programming language using Android Studio as IDE. This prototype application is developed for android operating system having a minimum API level 17, and target API level 26. The application is fully working and implemented on the Android smartphone.

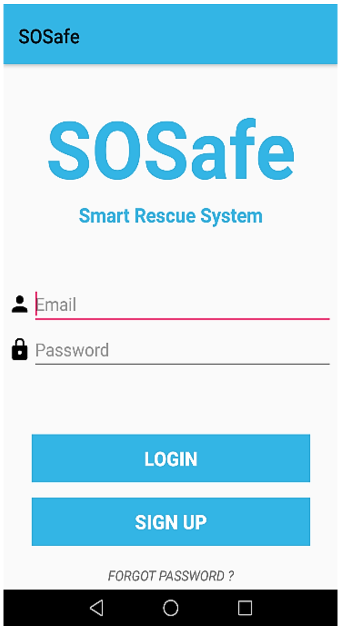
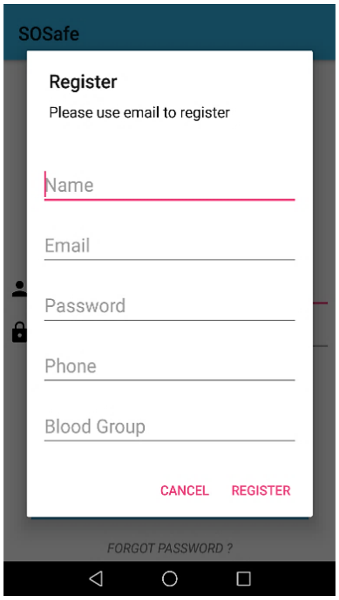


Fig. 5. Log in screen (SOSafe).

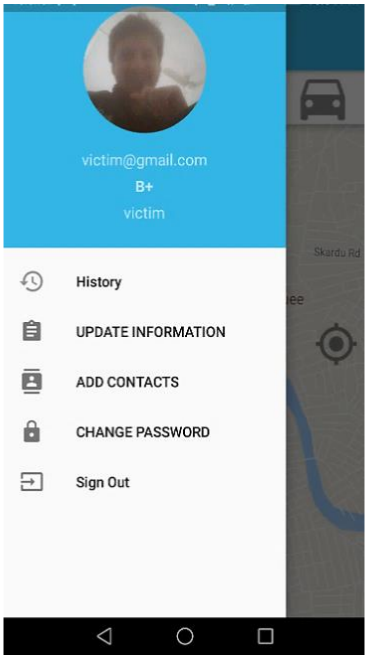

Register

Please use email to register

Fig. 6. Sign up screen (SOSafe).

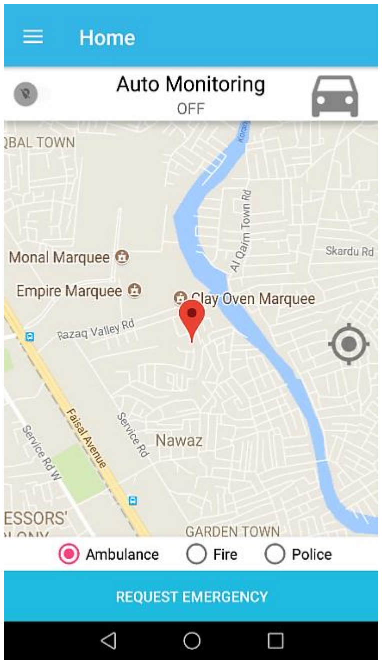
Fig. 5 shows log in screen of SOSafe, users can use email and password that they used to register, to log in to the system. After users logs in to the system, they will be able to use all system features.

Fig. 6 shows sign up screen of SOSafe, users provides email, password and other details (name, phone and blood group) in order to register. After registration all information will be saved at Firebase database.



victim@gmail.com
B+
victim

Fig. 7. Navigation drawer (SOSafe).

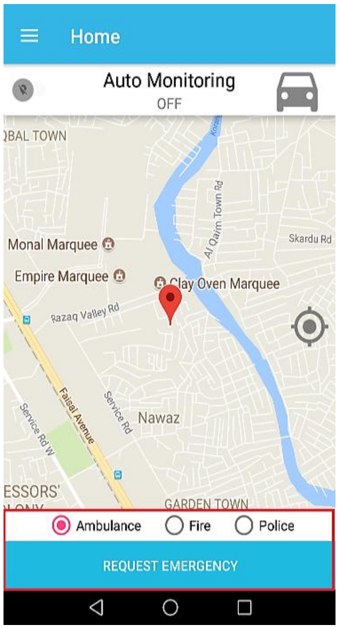


Auto Monitoring
OFF

Fig. 8. Home screen (SOSafe).

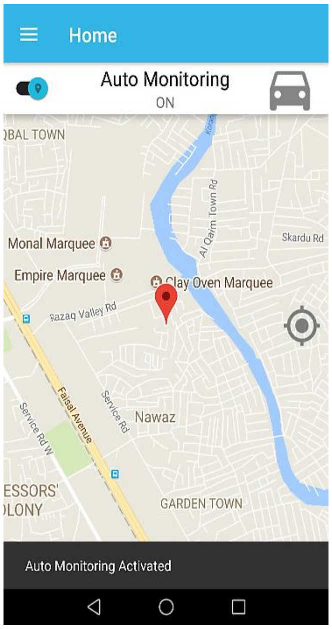
Fig. 7 shows navigation drawer of SOSafe, users can view history of previous emergencies, update their information, add emergency contacts numbers, change password and sign out.

Fig. 8 shows home screen of SOSafe, user can turn on Auto Monitoring for automatic accident detection or send emergency request manually.



Auto Monitoring
OFF

Fig. 9. Panic button (SOSafe).



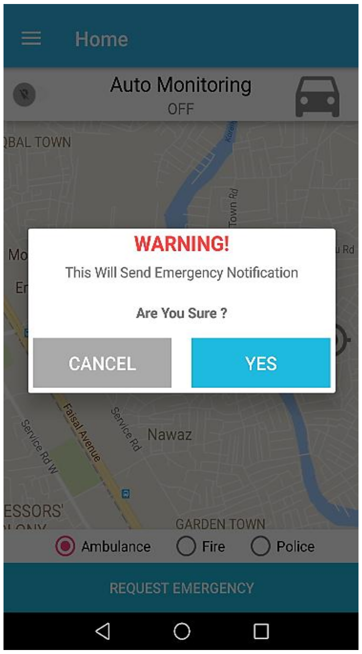
Auto Monitoring
ON

Auto Monitoring Activated

Fig. 10. Auto monitoring/auto accident detection (SOSafe).

Fig. 9 shows Panic Button feature of SOSafe, users can select the type of emergency service and press "Request Emergency" button to manually send an emergency request to selected emergency service responder near to them.

Fig. 10 shows, users of SOSafe can turn on "Automatic Monitoring", the system will register an accelerometer service running in background; it can now detect accidents.

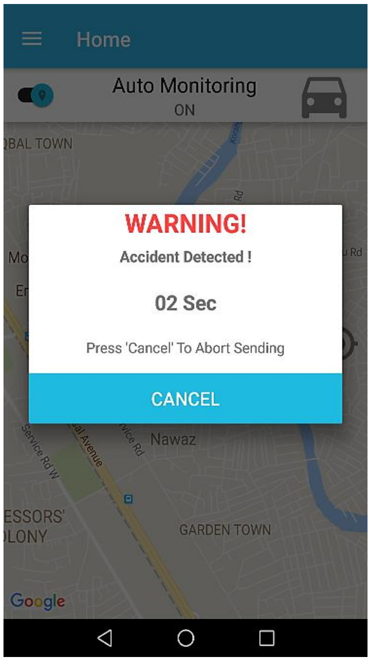


WARNING!

This Will Send Emergency Notification

Are You Sure ?

Fig. 11. Panic button alert dialog (SOSafe).



WARNING!

Accident Detected !

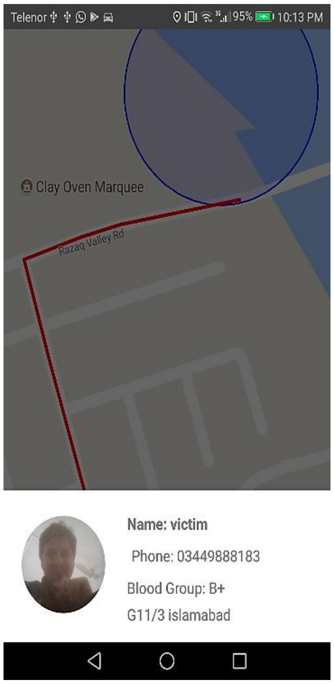
02 Sec

Press 'Cancel' To Abort Sending

Fig. 12. Accident detected alert dialog (SOSafe).

In Fig. 11, when user of SOSafe presses the panic button, the system will present an alert dialog to confirm the action before sending an emergency request to responder, this will help in situations when panic button is pressed accidently.

In Fig. 12 when SOSafe detects an accident, will present an alert dialog with a 15 sec count down timer, in case of false alert the user can abort sending request by pressing “Cancel” button. If there is no response from the user in 15 sec, it will be considered as an actual accident and the system will send an emergency alert to the nearest emergency responder and also to the emergency contacts.



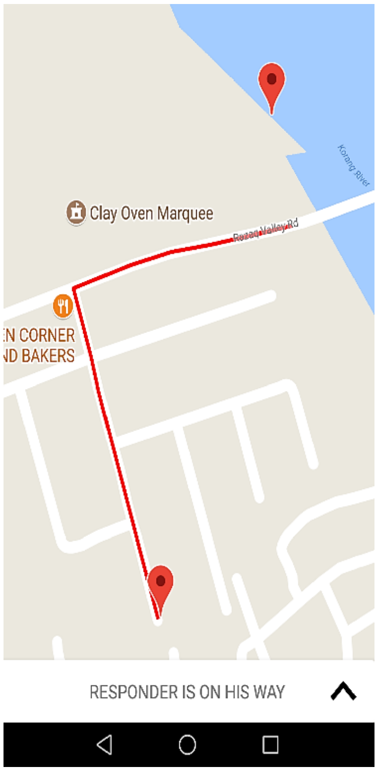
Name: victim

Phone: 03449888183

Blood Group: B+

G11/3 islamabad

Fig. 13. Responder details (SOSafe).



RESPONDER IS ON HIS WAY

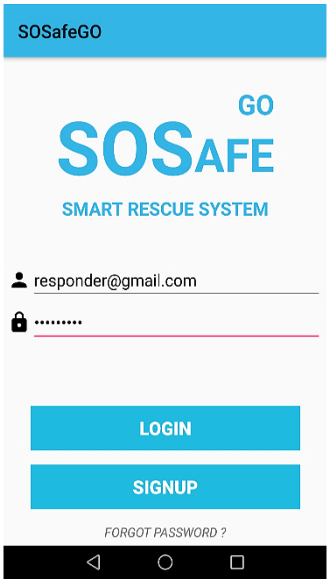
Fig. 14. Real time tracking of responder (SOSafe).

Fig. 13 shows, when an emergency request sent by emergency victim is accepted by emergency responder, SOSafe will show details about that responder to the victim.

In Fig. 14, SOSafe will show real time location of emergency responder on a Google map to the emergency victim.

B. *SOSafe Go*

Emergency responders' side application SOSafe Go is developed in Java programming language using Android Studio as IDE. This prototype application is developed for android operating system having a minimum API level 17, and target API level 26. The application is fully working and implemented on the Android smartphone.



responder@gmail.com
•••••••••

Fig. 15. Log in screen (SOSafe Go).

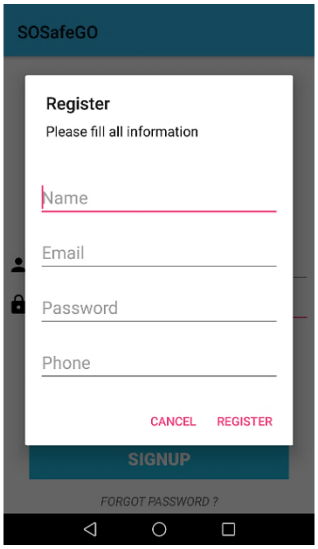


Fig. 16. Sign up screen (SOSafe Go).

Fig. 15 shows log in screen of SOSafe Go, emergency responder can use email and password that they used to register, to log in to the system. After emergency responders logs in to the system, they will be able to use all system features.

Fig. 16 shows sign up screen of SOSafe Go, emergency responder needs to provide email and password and name in order to register. After registration all information will be saved at Firebase database.

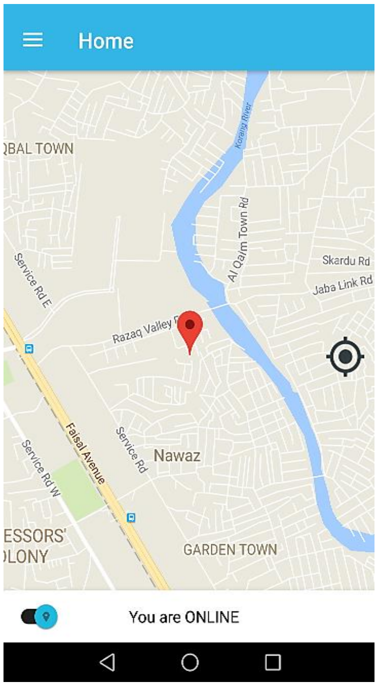
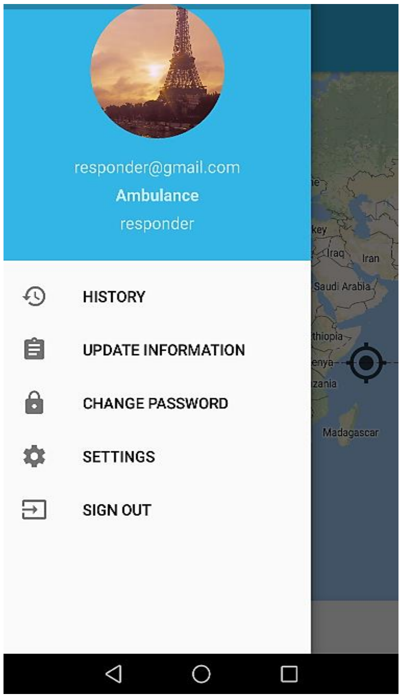


Fig. 17. Home screen (SOSafe Go).

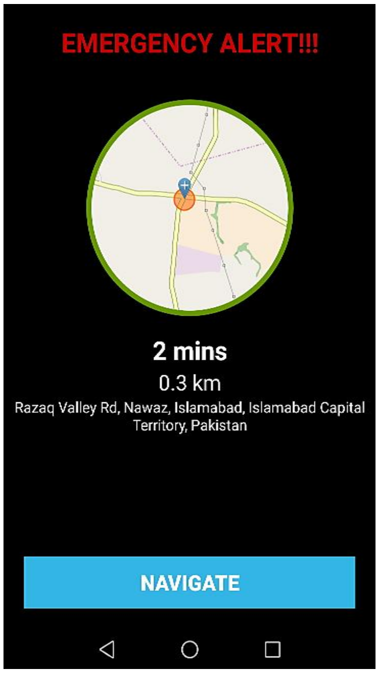


responder@gmail.com
Ambulance
responder

Fig. 18. Navigation drawer (SOSafe Go).

Fig. 17 shows home screen of SOSafe Go, responders can turn on the switch to go Online, responders can now receive emergency requests sent by emergency victim.

Fig. 18 shows the navigation drawer of SOSafe Go, emergency responders can view history of previous emergencies, update their information, change password, and select the type of emergency services they provide, e.g. Fire Brigade, Police Department or Ambulance.

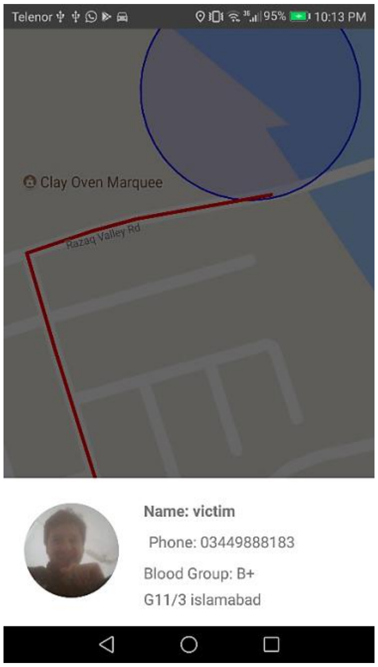

EMERGENCY ALERT!!!

2 mins

0.3 km

Razaq Valley Rd, Nawaz, Islamabad, Islamabad Capital Territory, Pakistan

Fig. 19. Emergency alert screen (SOSafe Go).



Name: victim

Phone: 03449888183

Blood Group: B+

G11/3 islamabad

Fig. 20. Victim's details (SOSafe Go).

In Fig. 19, when SOSafe Go receives an emergency request, the system will show an emergency alert screen to the responder with sound and vibration turned on. Emergency responder can press "Navigate" button to get direction to the emergency location.

Fig. 20 shows, when an emergency request by emergency victim is accepted by emergency responder, SOSafe Go will show details (name, phone, blood group, and address) about that emergency victim.

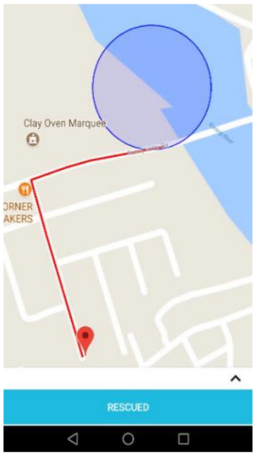


Fig. 21. Real time tracking of victim (SOSafe Go).

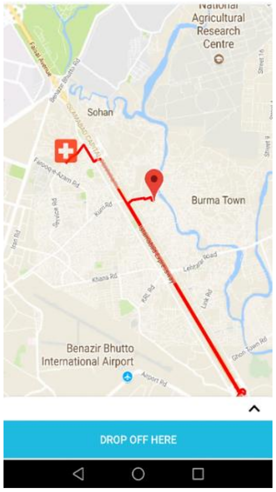

National Agricultural Research Centre
Sohan
Burma Town
Benazir Bhutto International Airport
DROP OFF HERE

Fig. 22. Nearby hospital route (SOSafe Go).

In Fig. 21, SOSafe Go will show real time location of emergency victim on a Google map to the emergency responder.

In Fig. 22 shows, upon rescuing emergency victim, if the victim is in need of medical assistance, SOSafe Go will guide emergency responders to the nearest hospital from emergency location.

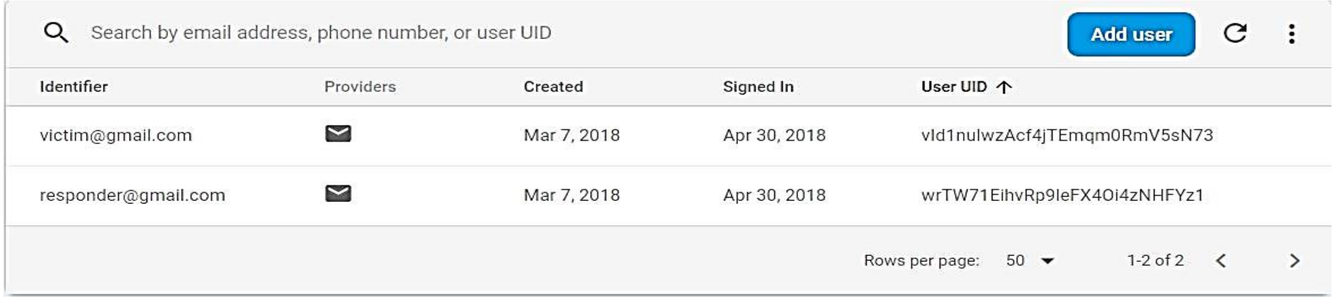
VII. Back End

In our proposed system we used only cloud-based server Firebase for data storage, user authentication, file sharing, location sharing and push messaging. Here we will discuss how firebase is utilized in our system.

A. *User Authentication*

1) Registration

In case of registering to the system user has to provide name, email address, password, phone, etc. Once user is registered into the system a passive user id will be generated and this id will always be used to identify user and access backend.



Search by email address, phone number, or user UID — Add user

Identifier	Providers	Created	Signed In	User UID ↑
victim@gmail.com	✉	Mar 7, 2018	Apr 30, 2018	vld1nulwzAcf4jTEmqm0RmV5sN73
responder@gmail.com	✉	Mar 7, 2018	Apr 30, 2018	wrTW71EihvRp9leFX4Oi4zNHFYz1

Rows per page: 50 1-2 of 2

Fig. 23. User authentication database snapshot.

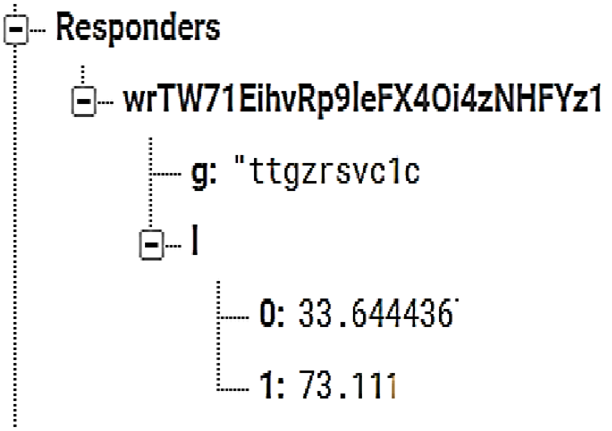
2) Log In

User has to provide email and password to login. Once the user is logged in, it is not necessary to login every time unless user is logged out. The firebase authentication system provides the user id which is synced with a device token that matches the user authenticity. Fig. 23 shows registered users, these users are authenticated and can log in to the system to use all system features.

B. Real Time Database

1) Responders

Fig. 24 shows the node in Firebase database where real time location data of responders currently online will be saved.

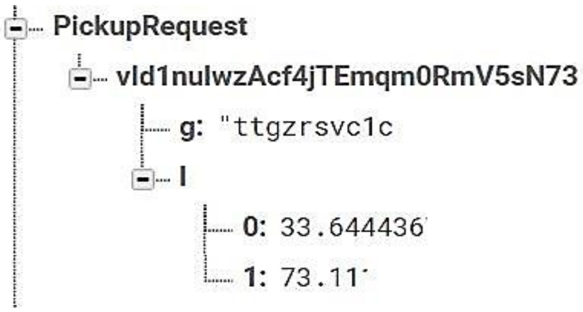


```
Responders
  wrTW71EihvRp9leFX4Oi4zNHFYz1
    g: "ttgzrsvc1c
    l
      0: 33.644436
      1: 73.111
```

Fig. 24. Online responders database snapshot.

2) PickupRequest

In Fig. 25, this node in Firebase database will contain emergency requests that are sent by emergency victims, with their location. Each request will be created with a separate child node using the id of victim that sent the request.

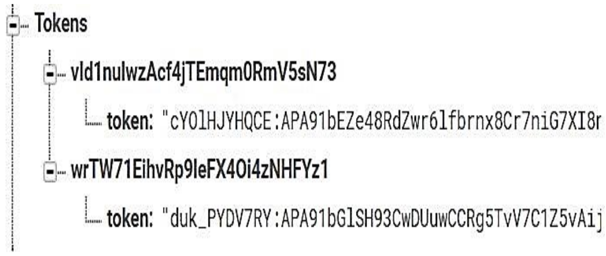


```
PickupRequest
  vld1nulwzAcf4jTEmqm0RmV5sN73
    g: "ttgzrsvc1c
    l
      0: 33.644436
      1: 73.11
```

Fig. 25. Emergency requests database snapshot.

3) Tokens

Is an id issued by Firebase Cloud Messaging connection servers to client applications allowing it to receive messages. Tokens will be used to send notification messages to both responder and victim. Fig. 26 shows the node in Firebase database where tokens for each user will be saved.

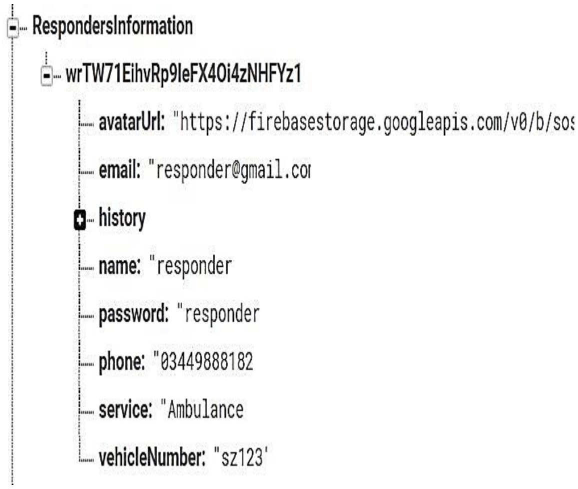


```
Tokens
  vld1nulwzAcf4jTEmqm0RmV5sN73
    token: "cYOlHJYHQCE:APA91bEZe48RdZwr6lfbrnx8Cr7niG7XI8r
  wrTW71EihvRp9leFX4Oi4zNHFYz1
    token: "duk_PYDV7RY:APA91bGlSH93CwDUuwCCRg5TvV7C1Z5vAij
```

Fig. 26. Tokens database snapshot.

4) RespondersInformation:

In Fig. 27, when SOSafe Go user signs up, all information about that responder will saved under this node in Firebase database. It will contain details about responders like, name, email, phone, emergency service type, vehicle number, history, etc.

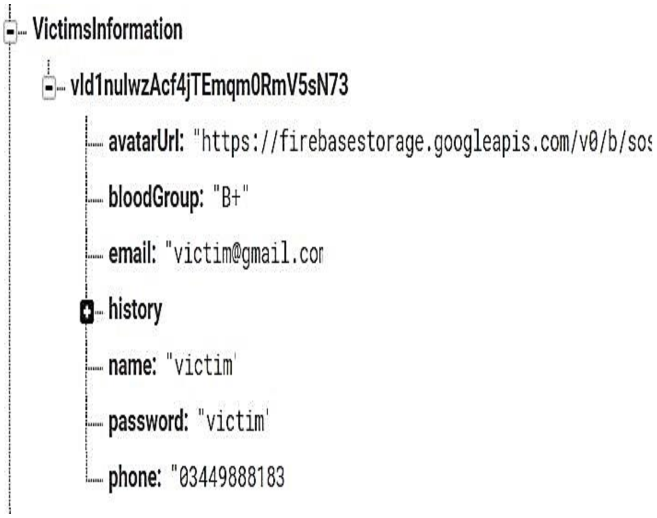


```
RespondersInformation
  wrTW71EihvRp9leFX4Oi4zNHFYz1
    avatarUrl: "https://firebasestorage.googleapis.com/v0/b/sos
    email: "responder@gmail.cor
    history
    name: "responder
    password: "responder
    phone: "03449888182
    service: "Ambulance
    vehicleNumber: "sz123'
```

Fig. 27. Responders information database snapshot.

5) VictimsInformation

In Fig. 28, when a SOSafe user signs up, all information about that user will saved under this node in Firebase database. It will contain details about victims like, name, email, phone, blood group, history, etc.

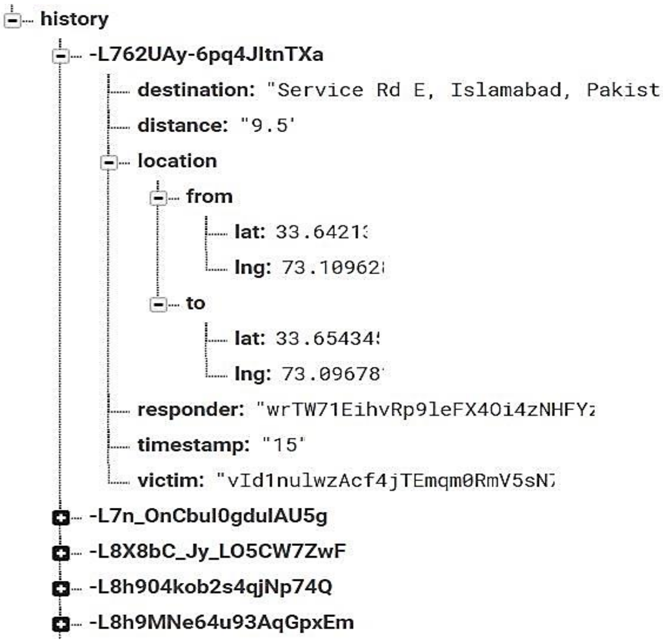


```
VictimsInformation
    vld1nulwzAcf4jTEmqm0RmV5sN73
        avatarUrl: "https://firebasestorage.googleapis.com/v0/b/sos
        bloodGroup: "B+"
        email: "victim@gmail.cor
        history
        name: "victim'
        password: "victim'
        phone: "03449888183
```

Fig. 28. Victims information database snapshot.

6) History

In Fig. 29, this node in Firebase Database will contain information about previous emergencies that each responder has successfully responded to. It will contain victim id, responder id, location, destination, time, etc. for each emergency.



```
history
    -L762UAy-6pq4JltnTXa
        destination: "Service Rd E, Islamabad, Pakist
        distance: "9.5'
        location
            from
                lat: 33.6421
                lng: 73.10962
            to
                lat: 33.65434
                lng: 73.09678
        responder: "wrTW71EihvRp9leFX4Oi4zNHFY
        timestamp: "15'
        victim: "vId1nulwzAcf4jTEmqm0RmV5sN
    -L7n_OnCbul0gdulAU5g
    -L8X8bC_Jy_LO5CW7ZwF
    -L8h904kob2s4qjNp74Q
    -L8h9MNe64u93AqGpxEm
```

Fig. 29. History database snapshot.

VIII. Results and Discussion (RQ1)

From emergency victim's point of view, during fatal accidents, emergency victims usually are not able to call an ambulance by themselves, in these situations the designed system will automatically detect the accident and will send an emergency notification to the nearest emergency responder available, to hopefully save victim's life. Sending an emergency alert is a lot easier and convenient because all essential functionalities reside together. In case of other emergencies, the system also provides functionality to send request to the desired emergency service.

From emergency responder's point of view, the application will show the location of the emergency that occurs near to them, this will help in reducing response time, so that they will be able to track victims in real time and rescue them as soon as possible, resulting in a more efficient usage of emergency services resources.

We conducted some tests by dropping a smartphone from height of 10, 15, 30 and 40cm and recorded the g-force values which can be seen in Table I.

We also mounted the smartphone in a car and recorded g-force value during sudden brakes while driving. Due to limited resources and lack of vehicle crash test labs in Pakistan we were not able to test our system by conducting a real vehicle crash test, but the tests we have done by applying sudden brakes in a vehicle, are somewhat close. The results can be seen in Table II.

TABLE I. G-Force During Free Fall

Height	Max	Min	Average
10 cm	2.484621402	1.126285167	1.805453284
15 cm	2.568356721	1.201817380	1.885087050
30 cm	2.981591310	1.766139158	2.373865234
40 cm	3.315491403	2.041593813	2.628542608

Table I shows the results from our drop testing we drop tested a smartphone as many as 10 times and recorded the results, we recorded that accidently dropping smartphone will not cause the system to generate an alert. The maximum g-force value we got in our testing by dropping smartphone from the height of 40cm was 3. 315491403. Our system will generate alert if the g-force value exceeds 4g.

TABLE II. G-Force While Applying Brakes in Vehicle.

Brakes	Max	Min	Average
Hard	3.025675430	2.220679102	2.623177266
Normal	2.357143091	1.301464250	1.829303670

Table II shows the results from testing by mounting a smartphone in a vehicle and applying brakes while driving, even in situation like applying sudden brakes in a vehicle the g-force values were lower than g-force during dropping a smartphone. The maximum value we got by applying sudden brakes in vehicle while driving was 3.025675430, which is lower than our critical threshold value. The threshold value at which our system will generate emergency alert is set to 4g, because during fatal traffic accidents the g-force value exceeds 4g. It can detect when an accident occurs but will avoid false positives in case of dropping a smartphone or applying sudden brakes in a vehicle.

IX. Research Contributions (RQ2)

Many systems exist that uses conventional in vehicle sensors and also some work done in the field of Smartphone based accident detection systems, we aim to develop a

Smartphone based accident detection system with much more features and better user experience. We developed a complete system for both emergency victims and emergency responders. The system uses on-board sensors of a typical Smartphone to detect accidents and report it to nearest emergency responders. Emergency responders will be able to track the exact location of the victims on a Google map. The system also provides help during other types of emergencies like Fire, Police etc. With innovative and creative ideas, we intend to take the system to another level, staying a step ahead from system which pre-exists.

The proposed system will provide much more functionalities than the existing systems like.

- Directly sends emergency notifications to the nearest available responder.
- Real time location tracking for both responders and victims on a Google map.
- An android application for emergency responders that directly receives notifications about the emergency that occurs near to them and is provided with real time location of the victim and is also provided with the details about victim such as name, blood group, address.
- Provide directions to the nearest hospital during medical emergencies.
- Provide other emergency services such as Fire Department, Ambulance and Police.

X. Conclusion and Future Work

In this research, we developed the accident detection and smart rescue system, which uses on board accelerometer sensor to detect accident and generate emergency alert and send it to the nearest emergency responder and will also send an SMS to emergency contact containing location coordinates of the accident. With real time location tracking for both victim and responder the system will drastically increase the survival rate of an accident victim by providing emergency aid in time. The system will also provide help during other emergencies such as during fire, robberies/theft and other medical emergencies. Emergency responder will be able pin point victim’s location on a Google map in real time.

The probability of false positives in a smartphone-based accident detection and rescue system is inevitable. We have added some features to reduce these issues. Here are some features we added to reduce false positives.

- **Acceleration filter:** The system will ignore g-force values lower than 4g.
- **Count down timer alert:** On detection of an accident the system will present an alert dialog with 15 sec count down, which the user will be able to cancel in case accident didn’t occur.

For future work, more research is needed in order to make the accident detection part more reliable and accurate which will help in reducing false positives. Adding additional sensors in combination with accelerometer for accident detection like gyroscope, microphone, camera (to automatically take pictures of the accident) and a voice recognition module to detect noises during a vehicle crash like noise when air bags are deployed, will drastically increase the reliability and accuracy of the system.

References

[1] Asirt.org. (n.d.). Road Crash Statistics. [online] Available at: http://asirt.org/Initiatives/Informing-Road-Users/Road-Safety-Facts/Road-Crash-Statistics [Accessed 10 Dec. 2017].

[2] Pbs.gov.pk. (n.d.). Traffic Accidents (Annual) | Pakistan Bureau of Statistics. [online] Available at: http://www.pbs.gov.pk/content/traffic-accidents-annual [Accessed 11 Dec. 2017].

[3] Traffic accidents kill an average 15 people in Pakistan daily. (2015). [Blog] Available at: https://www.thenews.com.pk/print/58036-traffic-accidents-kill-an-average-15-people-in-pakistan-daily [Accessed 11 Dec. 2017].

[4] Evanco and William M., “*The Impact of Rapid Incident Detection on Freeway Accident Fatalities*", technical report available from Mitretek , McLean, Virginia, USA, report No .WN 96W0000071, June 1996.

[5] J. Zaldivar, C. T. Calafate, J. C. Cano and P. Manzoni, "Providing accident detection in vehicular networks through OBD-II devices and Android-based smartphones," 2011 IEEE 36th Conference on Local Computer Networks, Bonn, 2011, pp. 813-819.

[6] Shahbaz Ahmed Khan Ghayyur, Salman Ahmed, Mukhtar Ali, Adnan Naseem, Abdul Razzaq and Naveed Ahmed, “A Systematic Literature Review of Success Factors and Barriers of Agile Software Development” International Journal of Advanced Computer Science and Applications(IJACSA), 9(3), 2018.

[7] Zainab S. Alwan Hamid M. Ali. “Car Accident Detection and Notification System Using Smartphone”. In: International Journal of Computer Science and Mobile Computing 4.4 (Apr. 2015), pp. 620–635.

[8] J. & Dougherty B. & Albright A. & Schmidt DC Chris T. & White. “WreckWatch:Automatic Traffic Accident Detection and Notification with Smartphones”. In: Journal of Mobile Networks and Applications manuscript(2011).

[9] Patel K.H., "Utilizing the Emergence of Android Smartphones for Public Welfare by Providing Advance Accident Detection and Remedy by 108 Ambulances", International Journal of Engineering Research & Technology (IJERT), Vol.2, Issue 9, PP 1340-1342, September – 2013.

[10] Chao Wang, Wei Duan, Jianzhang Ma and Chenhui Wang, "The research of Android System architecture and application programming " *Proceedings of 2011 International Conference on Computer Science and Network Technology*, Harbin, 2011, pp. 785-790.

[11] Shahbaz Ahmed Khan Ghayyur, Salman Ahmed, Adnan Naseem and Abdul Razzaq, “Motivators and Demotivators of Agile Software Development: Elicitation and Analysis” International Journal of Advanced Computer Science and Applications(ijacsa), 8(12), 2017.

[12] Android Developers. (n.d.). Meet Android Studio | Android Developers. [online] Available at: https://developer.android.com/studio/intro/ [Accessed 17 Dec. 2017].

[13] Android Developers. (n.d.). ADT Plugin (UNSUPPORTED) | Android Developers. [online] Available at: https://developer.android.com/studio/tools/sdk/eclipse-adt [Accessed 17 Dec. 2017].

[14] Gosling J., Joy B., Steele G., Bracha G., Buckley A. The Java Language Specification. Java SE 8 Edition. New Jersey: Pearson Education. 2014.

[15] Dimension Engineering LLC. A beginner’s guide to accelerometers. 2017. [online] Available at: https://www.dimensionengineering.com/info/accelerometers [Accessed 1 Jan. 2018].

[16] Edwin F. Taylor & John Archibald Wheeler. “Spacetime Physics”. In: 1st ed. Vol. 1. 1. San Francisco: W.H. Freeman, July 1966. Chap. 1, pp. 97–98. ISBN: 071670336X.

[17] James B. Angell & Stephen C. Terry & Phillip W. Barth. “Silicon Micromechanical Devices”. In: Scientific American 4.4 (Apr. 1983), pp. 44–55.

[18] Android Developers. (n.d.). Location and context overview | Android Developers. [online] Available at: https://developer.android.com/training/location/ [Accessed 4 Jan. 2018].

[19] Google Developers. Overview | Maps SDK for Android | Google Developers. [online] Available at: https://developers.google.com/maps/documentation/android-sdk/intro [Accessed 4 Jan. 2018].

[20] Google Developers. (n.d.). Overview | Places SDK for Android | Google Developers. [online] Available at: https://developers.google.com/places/android-sdk/intro [Accessed 15 Jan. 2018].

[21] Google Developers. (n.d.). Places API Usage and Billing | Places API | Google Developers. [online] Available at: https://developers.google.com/places/web-service/usage-and-billing [Accessed 15 Jan 2018].

[22] Developer.foursquare.com. (n.d.). Rate Limits - Foursquare Developer. [online] Available at: https://developer.foursquare.com/docs/api/troubleshooting/rate-limits [Accessed 15 Jan. 2018].

[23] Google Developers. (n.d.). Developer Guide | Directions API | Google Developers. [online] Available at: https://developers.google.com/maps/documentation/directions/intro [Accessed 4 Feb. 2018].

[24] Square.github.io. (n.d.). Retrofit. [online] Available at: http://square.github.io/retrofit/ [Accessed 4 Jan. 2018].

[25] Android Developers. (n.d.). Volley overview | Android Developers. [online] Available at: https://developer.android.com/training/volley/ [Accessed 4 Jan. 2018].

[26] Firebase. (n.d.). Firebase Cloud Messaging | Firebase. [online] Available at: https://firebase.google.com/docs/cloud-messaging/ [Accessed 8 Feb. 2018].

[27] GitHub. (n.d.). firebase/geofire-java. [online] Available at: https://github.com/firebase/geofire-java [Accessed 26 Jan. 2018]

Dist-Coop: Distributed Cooperative Transmission in UWSNs using Optimization Congestion Control and Opportunistic Routing

Malik Taimur Ali
Department of Computer Science
Iqra National University
Peshawar, Pakistan

Mian Ahmed Jan
Department of Computer Science
Abdul Wali Khan University
Mardan, Pakistan

Sheeraz Ahmed
Department of Computer Science
Iqra National University
Peshawar, Pakistan

Saqib Shahid Rahim
Department of Computer Science
Abasyn University
Peshawar, Pakistan

Atif Ishtiaq
Department of Computer Science
Iqra National University
Peshawar, Pakistan

Mukhtar Ahmad
Department of Electronics
Islamia College Peshawar
Peshawar, Pakistan

Mukhtaj Khan
Department of Computer Science
Abdul Wali Khan University
Mardan, Pakistan

M. Ayub Khan
Department of Computer Science
Iqra National University
Peshawar, Pakistan

***Abstract*—One of the real issues in UWSN is congestion control. The need is to plan an optimized congestion control scheme which enhances the network life time and in addition limits the usage of energy in data transmission from source to destination. In this paper, we propose a routing protocol called Dist-Coop in UWSN. Dist-Coop is a distributed cooperation based routing scheme which uses mechanism for optimized congestion control in noisy links of underwater environment. It is compact, energy proficient and high throughput opportunistic routing scheme for UWSN. In this proposed protocol architecture, we present congestion control with cooperative transmission of data packets utilizing relay sensors. The final objective is to enhance the network life time and forward information utilizing cooperation procedure, limiting energy consumption amid transmission of information. At destination node, combining strategy utilized is based on Signal-to-Noise Ratio (SNRC). Simulation results of Dist-Coop scheme indicate better outcomes in terms of energy consumption, throughput and network lifetime in contrast with Co-UWSN and EH-UWSN routing protocols. Dist-Coop has expended substantially less energy and better throughput when contrasted with these protocols.**

***Keywords*—Opportunistic routing; cooperation; congestion control; signal-to-noise ratio**

I. Introduction

Wireless networks are such type of networks which use Infrared or Radio Frequencies signals to share data or information between the devices connected with each other. There are number of wireless devices available now a days; for example, laptops, mobile sets, small size PCs, wireless sensors and satellite receivers among others. The new fourth era of cellular communication has greatly increased the data transmission speed, which provides the variety of high speed mobile data rates. At the same time, different new standards like Bluetooth, Infrared, 802.11 for short range radio helps to produce a variety of new application for enterprises and home networking, which enables wireless multimedia and data communication in offices and homes. Examples of wireless networks are cellular networks, Terrestrial Networks, Satellite Communication Networks, Ad hoc Networks and Wireless Sensor Networks [1].

A. Types of Wireless Sensor Networks

Sensor nodes are most commonly deployed in Remote areas and underwater. In these areas, the sensor network faces different challenges according to the environment. Mainly there are five types of WSNs:

1) Terrestrial Wireless Sensor Networks
2) Underground Wireless Sensor Networks
3) Underwater Wireless Sensor Networks
4) Multi-media Wireless Sensor Networks
5) Mobile Wireless Sensor Networks

1) Terrestrial Wireless Sensor Networks: These types of networks mainly consists of hundreds to thousands small inexpensive sensor nodes, which are deployed in a specific region, either in ad hoc (unstructured) manner or in pre-defined (structured) manner. In ad hoc manner, the sensor nodes are deployed in a specific region randomly. In terrestrial WSNs, the sensor nodes must be able to communicate successfully with the base station, while limited battery power is available [2].

2) Underground Wireless Sensor Network: In Underground WSNs, the sensor nodes are placed underground in a specific targeted area. Basically it is used to observe the underground situation and communicate this information to the surface area. These sensor nodes transmit the information to the sink nodes, which are further transmitted to the base station. The underground WSNs are more costly as compared to terrestrial WSNs. Due to signal losses and attenuation, the wireless communication becomes a challenge in these different environments. To increase the life time of the network, careful planning and cost consideration is required in underground WSNs [3].

3) Underwater Wireless Sensor Networks: In these types of networks, numbers of nodes are deployed into the water. Underwater WSNs are more costly and very difficult to deploy it in underwater. Sensor nodes in underwater WSNs communicate with each other and with the sink nodes using acoustic signals. Acoustic communication is again big challenge in underwater due to limited bandwidth, large propagation delay, signal fading problems and sensor node failure. These sensor nodes have a limited battery power and very difficult to recharge or replace a battery in harsh environment of water. For efficient use of energy, underwater WSNs are required to develop efficient underwater communication and networking techniques as shown in Fig. 1 [4].

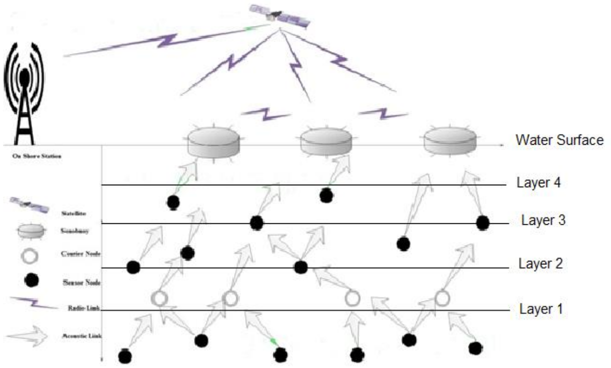


Fig. 1. Underwater wireless sensor network architecture with layers.

4) Multi-media Wireless Sensor Networks: To monitor events or tracking any object, wireless sensor networks are a suitable platform. It consists of low price nodes while having large in number to cover a specific region. These nodes have different properties and connected with each other via different sources. Nodes are arranged in the atmosphere in a pre-defined manner for a specific task [5].

In multi-media applications such as video streaming requires high bandwidth to transmit the whole contents. Therefore, high energy is consumed for high data rates. Multi-media WSNs needs to develop such transmission techniques which have high bandwidth and low energy consumption. Due to variable delay and channel capacity, Quality of Service (QoS) is difficult to preserve in multi-media WSNs. It is necessary to obtain certain level of QoS for reliable delivery of the contents [5].

5) Mobile Wireless Sensor Networks: In such networks, nodes can freely move within the environment on their own capability. These nodes performing different operation like sensing and communicating as a fixed nodes. These nodes have the ability to divert their location and manage themselves accordingly. In a network, mobile node communicates with another mobile node within a range of each other and transfers gathered information [5].

B. Major Challenges in UWSNs

Quantities of difficulties are looked in brutal submerged conditions. These issues are looked in two angles; Technical and Research challenges, for successful utilization of submerged sensor sensors. A portion of the significant issues are given below [1]:

- Hard to energize batteries and have very limited battery power.
- Accessible bandwidth is limited.
- Channel is influenced by multipath fading.
- Higher size of end to end defer when contrasted with earthly WSN.
- Mistake rates of bits are high.
- Solar energy cannot be used in submerged.
- Corrosion and Fouling may cause to come up short submerged sensors.
- Localizations.
- Data extraction.

One of the significant issues seen in UWSN is battery usage constraint, it's energizing and substitution. Because of unforgiving submerged circumstance, it is exceptionally troublesome and costly to revive or supplant sensor sensors batteries. Based on specified difficulties, the administration and planning of UWSNs routing techniques design is a major test for scientists and researchers [6].

II. Literature Review

According to authors in [2], it is exceptionally hard to give effective routing administrations in UWSNs. The radio signals does not work appropriately in underwater condition that's why acoustic signs are utilized. Acoustic signals have lesser transfer speeds and longer spread deferral when contrasted with radio signs. Besides, because of water ebb and flow, Network topology in UWSN in nature is dynamic and sensors move latently. Here the proposed protocol is known as Depth Based Routing convention (DBR) to tackle the issue of restriction up to some degree. DBR does not require data of sensor sensors of full-dimensional area. It requires just neighborhood profundity data that can be effectively accomplished by putting a low value profundity sensor with each submerged node.

Researchers in [7], use that a standout amongst the most vital issue in UWSN is energy impediment. Because of unforgiving submerged condition, the charging and substitution of sensor sensors batteries is exceptionally troublesome and costly. The arrangement proposed is to outline an energy effective directing convention named Energy Efficient Depth Based Routing (EEDBR) to take care of the said issue up to some degree in UWSNs. The proposed convention ascertains the profundity of the sensor sensors alongside its lingering energy, to build the life time of the system. Researchers in [8], assess diverse routing conventions like VBF, DBR, H2-DAB, and QELAR and so on based on limitation systems, minimization of energy and ascertain the holding time in UWSNs. Each directing convention has diverse objectives like decrease in energy utilization, accomplishment of strength and so on. This paper for the most part centers to feature the difficulties looked in the plan of directing convention for UWSNs.

Specialist in [9], use that in thick submerged condition, the significant issues are the variety in arrange topology, high blunder rate, and more noteworthy energy utilization for the transmission of information. Expansion with these quantities of issues, there are some real focal points of UWSNs, for example, submerged administration, oil investigations and a few calamities administration issues. Versatile portability of Courier sensors in Threshold-Optimized Depth-based (AMCTD) directing plan has planned to accomplish more noteworthy system lifetime based on sensors profundity, diminishing the utilization of energy particularly amid the steadiness time frame. The versatile development of messenger node maintains the system throughput in the extra state of system. In [10], proposed routing convention in light of sending capacity (FF) for UWSNs named enhanced Adaptive Mobility of Courier sensors in Threshold-Optimized profundity based directing (iAMCTD). Contrasting and the current profundity based routing conventions; developed convention misuses the thickness of the system for time basic data. To beat way misfortune, proliferation dormancy and flooding holding time is ascertained and utilize directing measurements. Which contained confinement free signals to-clamor proportion (LSNR), signals quality list (SQI), Energy cost work (ECF) and Depth Dependent Function (DDF). Reckoning of FF procedure increments the life time of a system and decreases the transmission adversity.

Researchers in [11], used that to diminish the issue of more prominent postponement, the deferral touchy routing convention is required. The proposed plans known as Delay-Sensitive Depth-Based Routing (DSEEDBR) and affectability based levels having limit esteems to enable the routing in view of its profundity to diminish the three plans are consolidated to configuration defer effective Delay-Sensitive Holding time (DSHT) and Priority Factors (PF). In request to figure the transmission misfortune and got signals speed, ideal weight work (WF) are presented in these plans. Besides, to tackle the issue of postponement, it needs better forward to limit the transmission delay in such locale where sensors are put in low profundity. In DSDBR, WF and Fi are utilized for better forwarder choice and to evaluate DSHT, they presented dth. The significant reason for high engendering delays is low profundity Transmission. Researchers in [12] proposed Dual Sink Efficient and Balanced Energy utilization Technique (DSEBET) for Underwater Acoustic Sensor Networks (UASNs). In UASNs organize the system lifetime crumples because of restricted energy asset. The real issue is to adjust energy in low system lifetime. In this way, to expand the system lifetime, the utilization of energy must be adjusted. In DSEBET, interfaces between the sensors are made based on their little separation "Nj" hand-off node is chosen for information transmission. In transmission of information, each node has parallel Energy Level Numbers (ELNs).

Researchers in [13], proposed impedance mindful and solid directing convention for UWSNs. Based on built up way from end-to-end, the following forwarder has been chosen for transmission of information parcels. Based on this approach, the void gaps can be disposed of utilizing the proposed routing convention. For dependable correspondence, interference with in the divert is additionally considered in directing metric amid the choice of sending node. The proposed directing convention chooses neighbor node as a next forwarder from source to goal, so crash likelihood can be decreased at arrange layer. Later, the momentous progression in the field of UWSNs has been accomplished in [14]. Number of routing conventions is configured to fathom the issues looked in UWSNs. In said plot, the goal is chiefly center to comprehend the working standards of those routing plans. Three investigation strategies are proposed for this reason; Clustering based, limitation based and collaboration based routing. Numerous routing conventions center around energy productivity, security of system and time proficiency diminish the postponement in time and increment the life time of the system.

Researchers in [15], suggested that in agreeable WSN condition, energy proficient plan is utilized as a part of quantities of sensors and Data Gathering Node (DGN). Based on DGN, diverse setup is utilized like the Number of Input to Number of yield and Number of Input to Single Output. In proposed convention, helpful energy plot for MIMO (C-MIMO) is utilized, where mistake is rectified by utilizing low thickness equality check (LDPC). The length of LDPC relies upon the measure of the message and mistake checking bits; where the rate of LDPC code changed with the extent of message and equality bits. Analyst in [16], used that in energy constrained systems, Cooperation by utilizing single transfer is more basic and successful correspondence. For the most part center to research and select a hand-off having least energy with control transmission. For agreeable transmission, Relays ascertain their base required power with a specific end goal to oversee collaboration. Just the best one hand-off is chosen among all to diminish the power use. The decision of transfer depends on appropriated path with little overhead.

Authors in [17] offered Energy Harvesting (EH-WSN) technique having the upside of productive collaboration and transfer choice in view of little power use. The convention is assessed in two stages, in initial step, helpful correspondence is assessed and afterward the energy of every node is figured. This plan characterizes unassuming and valuable system for EH-WSN with participation and straightforwardly connected in genuine cases. To accomplish energy productive collaboration based correspondence, it can likewise coordinate with different conventions. In [18], proposed to get exceptional capacity by utilizing new form of codes to redesign programming, called shrewd Internet of things (IOT). Because of cruel channel qualities, information gathering is still issues in WSN because of long deferral, energy utilization and retransmission. Information is transmitted agreeably by refreshing programming to accomplish dependable correspondence. Dependability enhanced agreeable correspondence RICC scheme is proposed to enhance the unwavering quality of the system in multi-hop helpful correspondence without diminishing system lifetime. In WSNs, sensors are put at various positions, so every node handle diverse bundles, so uneven energy utilization happened in the system. To proficiently utilize the lingering energy, in RICC, energy utilization is embraced by the prerequisites. Low power utilization is embraced to keep up long system lifetime.

Analysts in [19], argued that the information transmitted by WSNs are utilized for various purposes in various parts like security, arrange topology and correspondence conventions are happens in the sending of WSNs. Distinctive conventions are utilized by the prerequisites of the application, for example, verification, remove, number of parcels transmitted amid particular timeframe. In [20], Author proposed a Wireless Network (WPCN) and the proposed plot comprises of one source to goal combining and having various forward (DF) and translate transfers. This plan proposed energy limit based multi-hand-off choice (ETMRS) for WPCN. The transfers are sent arbitrarily and are known by just its profundity data for switching between Energy and Information Forwarding gathering modes. The limited battery asset of every sensor node, the charging/releasing of node happens that required ETMRS plot over blended Nakagami-m and Rayleigh blurring channels.

In [21], authors proposed a technique expanded the proficiency and unwavering quality of the system by utilizing the sink portability and helpful routing. Numerous conventions are created to perform participation with a specific end goal to enhance connect productivity by examining physical and MAC layer perspectives. The proposed protocols basically center to investigate arrange layer with sink versatility. The goal and hand-off node is chosen based on remaining energy and its profundity data. Sink versatility by social occasion information from various sensors specifically increment the effectiveness of the system. The proposed plot works in various stages; information procurement stage, arrange usage stage, limit based information detecting and directing stage. Every node ascertains its alive neighbor sensors to refresh profundity edge level in organize instatement stage. Researchers in [22], use that to take care of various issues happened in UWSNs, Improved routing arrangement is required for productive information sending. This paper proposed enhanced agreeable plan to build the life time and unwavering quality of the UWSNs. The proposed protocol receives participation on arrange layer inside existing non-agreeable routing plan, Depth Based Routing (DBR), to enhance the dependability and throughput. The determination of hand-off node depends on its profundity and information is sent from source to goal helpfully by utilizing transfer sensors.

In [23], researchers broke down that the cruel submerged condition because of blurring and unavoidable clamors makes it hard to perform mistake free transmission of information. The choice of transfer depends on node profundity and its leftover energy. Analyst in [24], use that to enhance the correspondence quality, agreeable correspondence is received in UWSNs by utilizing transfer sensor sensors. To enhance the execution of the system, proposed a helpful plan known as Analytical approach towards Reliability with Cooperation for UWSNs (ARCUN). The proposed convention has high throughput and energy proficient routing plan for UWSNs. Transfer chose from a gathering of hand-off sensor sensors, ascertain the separation and SNR proportion of the submerged channel. The proposed protocol uses collaboration and SNR for delay-delicate application to expand the parcel conveyance proportion and soundness time of the system. In transmission of information without collaboration implies information is transmitted by utilizing direct connection from source to goal.

In [25], authors use that the two principle challenges are; radio waves can't function admirably in submerged condition and the second is that the acoustic correspondence is moderate. The proposed plot centers around area base collaboration. The region is separated in various clusters and after that collaboration between sensors take place. The proposed protocol named Energy Efficient Adaptive Cooperation Routing convention for UWSNs (EACE) accomplished longer system lifetime and less utilization of energy with participation between sensors. Specialist in [26] use that dependability is one of the imperative factors to enhance the general execution of UWSNs. Uproarious condition and poor channel quality lessening the dependability and influence the system execution by influencing the trustworthiness of information. Helpful directing in UWSNs enhances the unwavering quality and honesty of the information. The proposed protocol called Improved Adaptive Cooperative Routing in UWSNs (IACR), comprises of two transfer node and one ace node is chosen among the accessible sensors from source to sink node for transmission of information.

In [27], researchers recommended that Cooperative condition increment the life time of the system in testing condition. Neighbor sensors are utilized for transmission of information helpfully as transfer sensors. The planned protocol named Cooperative UWSN (Co-UWSN), which builds the system life time by expanding unwavering quality, energy effectiveness and expanding throughput in UWSNs. To beat blurring, agreeable assorted variety is presented. Choice of hand-off node in light of channel quality and separation among neighbor node for quality and separation among neighbor node for fruitful information bundle transmission. Loss of information parcel is diminished by happening variety inside and out limit level. The system soundness period and load adjusting is accomplished by utilizing ideal weight calculation and helpful condition.

III. Motivation

Co-UWSN and SPARCO are proficient plans yet it might have a few issues in adjusting of load in sensors having low profundity. Vast energy is devoured by these low profundity sensors when transmitting information, which create scope gaps in a system. These gaps are essentially created because of load administration among alive sensors. In the event that the profundity expands, the utilization of energy amid transmission is additionally expanded.

All sensor nodes have ability to transfer theirs data as well as data that came from other sensor nodes by the route that ending at sink. Nodes utilize also relay sensor nodes when transferring the data to BS. Before collecting data packets, a routine process performed to decide an optimal node amongst neighbors of each sensor node to minimize cost function. Sensed and incoming data packets traffic is transferred to this sensor node. The establishing of these paths guarantees that collected data are forwarded to BS. Issues of the multi hop routing are attempted here by two stage technique. According to 1st stage, optimum cooperation based transmission approach is imitative for every sensor node for sending data to their neighbor sensor nodes in case of selected. Each sensor node recognizes link cost for its neighbors. In 2nd stage, a routing algorithm is used called Bellman-Ford distributed algorithm for obtaining minimum cost route from any sensor node to BS. Cost values of link attained in 1st stage that is utilized in 2nd stage. Data is labeled by creating sensor nodes and organized as bit data packets.

According to broadcasting stage, source sensor node broadcast data packets with index of destination sensor node. The subset of neighboring relays get signals and in 2nd stage use the cooperation based approach for transmission such as decode and forward strategy to send data packets to destination. It is also possible that source take part too in 2nd stage in case if these outcomes in lesser total consumption of power. Power of transmission in cooperation based transmissions also requires determining. Therefore with fix destination, the source requires to determine broadcast power that is represented by Pb. In addition, SNR at destination is mandatory to be high by set verge that is stated formerly. Generally relays are also used by other sensor nodes in network. Therefore for cooperation based transmission, destination requires to recognize to relays. Hence index of destination is also required to transfer to relays. The main purpose of any cooperation based transmission is delivering data packets from a source to a destination when satisfying SNR, with least likely total power of transmission.

It is also important to note that resolution of this issue may be direct which non-cooperative transmission from source to destination is. It is also be noted that by selecting a greater values of Pb, more sensor nodes are involved in cooperation based transmission that is leads to minimized total power that is obligatory for cooperation based transmission. In contrast the lesser values of Pb leads to high power obligatory for the transmission of data in 2nd stage. This trade-offs stated in certain research papers and in-depth exploration is planned as an approach to find optimum value of broadcast power [13].

According to current paper, an efficient structure is presented for methodically resolve the optimization issue which is addressed above.

IV. Flow-chart and Methodology of Dist-Coop

The proposed look into system is quickly clarified through the accompanying flow chart in Fig. 2. Irregular sending of 250 sensors submerged with 5 sinks on the water surface in a domain of 500x500x500. Each sensor will begin to find its neighbouring sensors. This will make the sensors to recognize the best option of neighbours which can be chosen for data transfer based on remaining energy and estimated cost function. It is imperative on the grounds that each node which is encompassed by different sensors which is diverse with various sensors because of node thickness. In the focused on territory sensors are scattered arbitrarily.

After estimation of neighbour sensors, in subsequent stage we ascertain the residual energy (RE) of the considerable number of sensors conveyed. Here we separate the sensors on edge energy level 1 and 2. In the event that the R.E of the node is more prominent or equivalent to edge energy level 1, at that point this node can be chosen as a transfer node. In the event that the RE is not as much as limit energy level 1, at that point hand-off node again process its RE, if the RE is more noteworthy than edge level 2 and not as much as edge level 1, so it can have the capacity to use as a typical sensor node. On the off chance that, if R.E is not as much as edge level 2 than it begins its Energy Harvesting (EH) to increase its energy level.

V. Dist-Coop Protocol Mathematical Modelling

In this segment, we present our estimated optimized congestion control and link aware with Cooperation based scheme Dist-Coop which guarantees to enhance the Network Lifetime, Packets Delivery Ratio of the system and lessened Energy use by sensors.

A. Network Topology

System limit, energy utilization and the unwavering quality of a system relies upon arranged topology. The greatest scope of a sensor node is not sufficient to cover the whole system, so multi-hop correspondence is utilized. Information got from the source node at sink node is accumulated. It might consider that sink node has no energy requirement that may speak with any of the sensors without collaboration. System is isolated into various layers based on profundity and is made out of heterogeneous sensors. The lower profundity sensors send information to higher profundity sensors and the procedure goes ahead till the information comes to at the surface of the water. The transfer sensors are propelled node, since they have double duties of transmitting its own information and also transferring of neighbor sensors. In Fig. 6, if there should be an occurrence of ordinary information transmission the source node transmit information toss two hand-off sensors in a collaboration mode, on the grounds that if hand-off sensors isn't accessible or may dead, so there will be another hand-off and an immediate connection accessible for the information exchange.

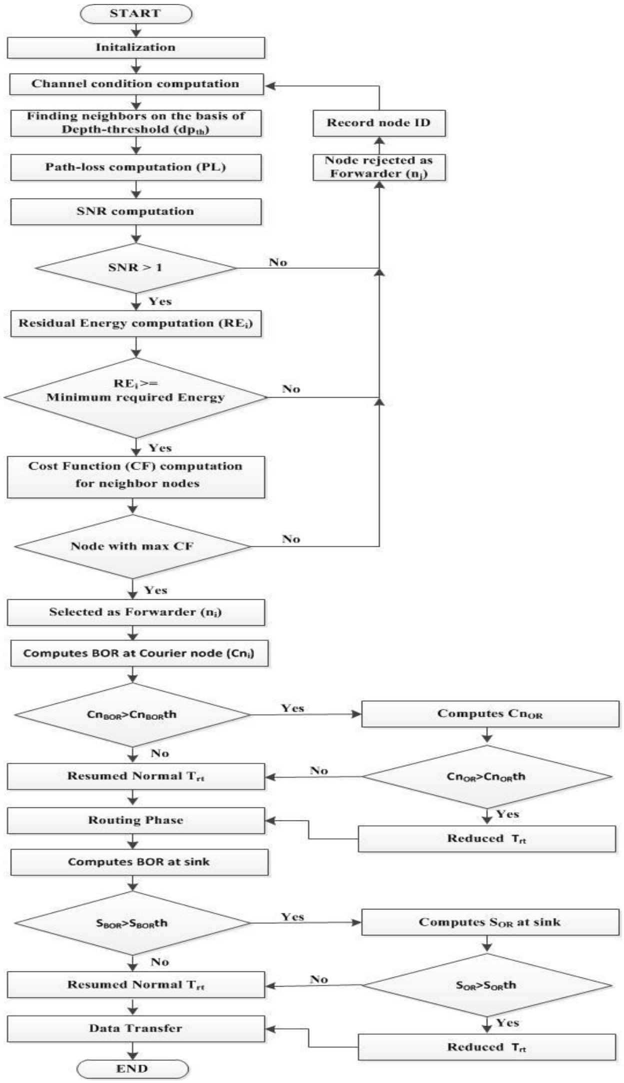

START
Initalization
Channel condition computation
Finding neighbors on the basis of Depth-threshold (dp_th)
Path-loss computation (PL)
SNR computation
SNR > 1
No
Yes
Record node ID
Node rejected as Forwarder (n_j)
Residual Energy computation (RE_i)
RE_i >= Minimum required Energy
No
Yes
Cost Function (CF) computation for neighbor nodes
Node with max CF
No
Yes
Selected as Forwarder (n_i)
Computes BOR at Courier node (Cn_i)
Cn_BOR>Cn_BORth
Yes
Computes Cn_OR
No
Resumed Normal T_rt
No
Cn_OR>Cn_ORth
Yes
Routing Phase
Reduced T_rt
Computes BOR at sink
S_BOR>S_BORth
Yes
Computes S_OR at sink
No
Resumed Normal T_rt
No
S_OR>S_ORth
Yes
Data Transfer
Reduced T_rt
END

Fig. 2. Flow chart of the Scheme EH-UWSN.

B. Absorption and Scattering Models

Two noteworthy reasons of constriction in submerged condition are scrambling and ingestion misfortunes. In disseminating, the electromagnetic signals is diverted far from its unique bearing, and in assimilation, the electromagnetic signals energy is changed over starting with one shape then onto the next frame like warmth or substance. Consequently:

$$c(\lambda) = a(\lambda) + b(\lambda) \quad (1)$$

Where a and b refer to ingestion and dissipating separately, estimated in m-1, and λ is the wavelength of signals in nm. Utilizing the co-productive of weakening Beer, s Law decide the construction of an acoustic signals for a separation d is given by [25]:

$$I = Io\, e^{-c(\lambda)}\, d \quad (2)$$

Where Io is a normalizing steady. The standardization esteem is a non-negative esteem, relies upon the circumstance to make the change precisely equivalent to 1. c(λ) is co-proficient of lessening, d is a separation.

C. Ambient Noise

Surrounding commotion is one of the vital factor in submerged acoustic channel condition. Encompassing commotion is fundamentally a connection between the measure of data worried about air state of the sea, ocean condition of the sea, wind speed and sea life natural impacts. Four fundamental sources show the distinctive overwhelming levels of surrounding commotion. They are: Turbulence, waves, delivery and warm clamor. The aggregate power unearthly thickness of clamor is communicated in db and is given by [27]:

$$NL = NLtb + NLsh + NLwv + NLth \quad (3)$$

Where

$$NLtb = 27\text{-}30 \log f ,$$

$$NLsh = 40+20(s-0.5) + 26 \log f - 60 \log (f+0.03)$$

$$NLwv = 50+7.5 (w-0.5) +20 \log f -40 \log (f+0.4)$$

$$NLth= -25+20 \log f$$

f in KHz , w is wind speed (m/s) and s- shipping activity factor.

D. Signal-to-Noise Ratio (SNR)

The SNR of an underwater acoustic signal at a receiver side can be calculated in dB by sonar equation [27] as follows:

$$SNR = SL\text{-}TL\text{-}NL\text{-}DI+c(\lambda) \quad (4)$$

Where

NL is ambient noise level in ocean (dB),

TL is transmission loss (dB),

DI is the directivity index and is set to zero and SL is the source level of transmission (dB) is given by

$$SL = 10 \log \left(\frac{\text{Pit}}{0.67\text{x}10\,(-18)}\right)(\text{dB}) \quad (5)$$

Where Pit is the transmission power intensity.

In shallow water, the Intensity, Pit is given in watt/m2 as follows.

$$\text{Pit} = \frac{\text{Pt}}{2*\pi*z} \quad (6)$$

In deep water, Pt, is given in watt/m2 as follows

$$\text{Pit} = \frac{\text{Pt}}{4*\pi*d} \quad (7)$$

Where, Pt is the transmitted power (watt) and d is the depth (m)

E. Initialization Phase

Three unique assignments are performed in this phase. Every sensor has information about its neighbor, sink node on the water surface is recognized, and every conceivable course toward various sink is assessed. Sensors communicate a packet, which contains data of the node like its profundity, node ID, and energy status. This packet has been gotten by the neighbor node and utilizes these for advance transmissions. Sink present on the water surface sends a Hello packet to each of its associated sensors. Utilizing high bundle transmission, with in the given transmission extend, every node distinguish its neighbor and independently kept up a line of neighbor under profundity limit to recognize the best sent node for its information transmission; Each node figure its weight utilizing the recipe given beneath [27].

$$\mathrm{Wi} = \frac{\max(\mathrm{SNR}(d_{\mathrm{SiR1}},f),\mathrm{SNR}(d_{\mathrm{SiR2}},f)\mathrm{SNR}(d_{\mathrm{SiDi}},f)+\max(\mathrm{R.E}_{\mathrm{R1}},\mathrm{R.E}_{\mathrm{R2}},\mathrm{R.Edi})}{\min(|d_{\mathrm{SiR1}}|^2, |d_{\mathrm{SiR2}}|^2, |d_{\mathrm{SiDi}}|^2)} \quad (8)$$

Where SNR(dSiR1,f), SNR(dSiR2,f), SNR(dSiD1,f) are the SNR of the corresponding links from Si to Di respectively, R.E is the residual energy of the corresponding sensors; dSiR1, dSiR2,and dSiDi are the distances from the corresponding source to its relays and immediate destination respectively.

F. Underwater Acoustic Cooperative Transmission

Let N fixed agents that are sensed and they represented by Si where the i ∈ N:= {1, . . . , N-1, N}. Relay sensor nodes are also arbitrarily set up in 2dim area. Suppose index of neighbor relays and other nodes for the ith node belongs to sets Li and Ni. Assume relaying and sensing sensor nodes are capable for adjusting power of transmissions while they all have same maximum power of transmission that is denoted by Pmax. Transmission range of agents that are sensing, directly related to maximum power of transmission. The sensor nodes are known as neighbors if theirs Euclidean distances are minimal than transmission range. In this process the major aim of sensor nodes deployments is observing the geographical area and transfers the data to a static BS. BS is likely as the sink and other nodes are the sources that sporadically measure data.

G. Cooperation based Model of Transmission

Assortment system stated as a system which takes more than two same replicas of transferred signals from the transmitter. According to this research paper, we deliberate a cooperation based protocol with two relay nodes as Fig. 3. It contains one source node, two relay nodes with a single destination node.

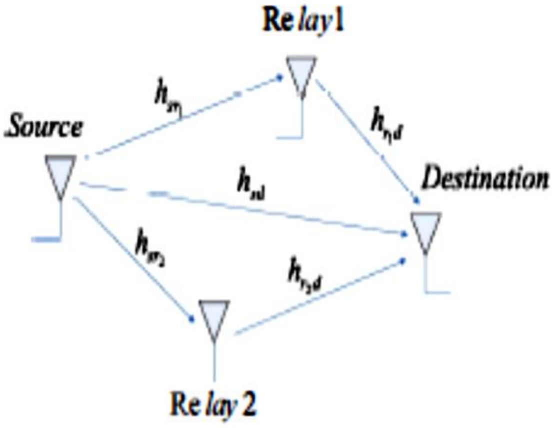


Fig. 3. Cooperative Transmission with Two Relays.

In Fig. 3, relay nodes depend on reformative cooperation while relay nodes sensed and transfers original message that comes from source node prior to forward decrypted bits to terminus [7]. In TWSNs transmission time is lesser while processed using phase synchronization procedures. Due to long transmission delays in UWSNs, addition scheme depends on analogue domain which flops as three signals reach to terminus at different time. Assortment merging methods are auspicious resolutions for processing the receive signals at terminus for UW communications with maximal ratio diversity, equal gain diversity and assortment diversity system [13]. It is observed that assortment methods diminish impacts of declining and increased channel excellence. These methods are mostly model for analogue waves in radio transmission that’s propagation delays greatly low as compare to processing delays at all sensor nodes. Though, the data signals travel alongside more than one channel (relays and terminus sensor nodes) with multi extents and arrived to terminus at diverse timings in UW environ. It is not possible to apply addition model to received signals at terminus in term of analogue data signals. Thus, our technique adopts diversity concepts by merging procedures which are applied for analogue field at physical layer to receive data signals at packet level relying on channel state information (CSI).

H. Selection of Relay and Destination

According to literature [8], [14], a source sensor node tries to search optimal relays to make possible the cooperative communications for TWNs with supposition the destination is pre-defined. This methodology depends on terminated timers at the possible relay nodes, is unsuitable for UW links as it unusually maximizes delays of UW channels. Moreover, link states possibly modify by high delays due to terminated timers or collision evasion amongst devices. This research examines cooperative base communications by distributed mode for UWSNs. In this paper it is assumed that a source sensor node, that requires transferring their message to sink node by group of hops, have n neighboring sensor nodes in its range of transmission that is shown in Fig. 4. Source sensor nodes rely on instant link state for determining that which ones amongst neighbors have maximum consistent links to transfer the information to sink node containing the relay and destination nodes. Channel properties are taken in account by electing processes. It contains SNR from all links to source sensor node and distance node from all neighbors to sink node. It is important to note that GPS systems that are fortified for terrestrial wireless based devices not able to work fine in case of the environments of UW by the restrictions of link properties as well as frequencies. Transmission processes consist of a sequence of six phases that are:

- Request to Send (RTS)
- Clear to Send (CTS)
- Source based computations
- Source transmissions
- Relays transmissions
- Acknowledgement.

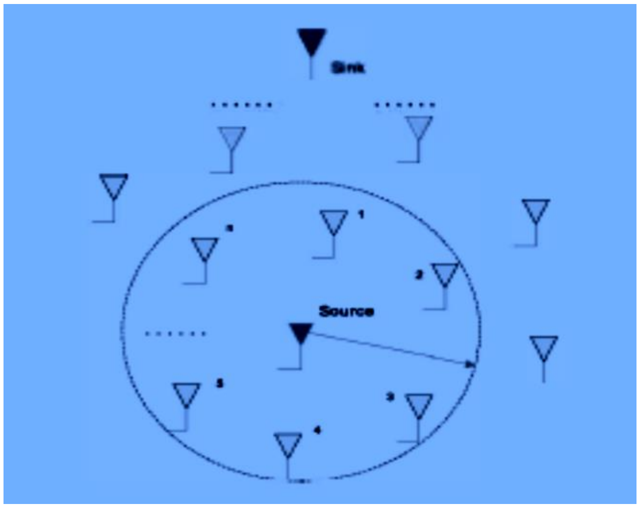


Fig. 4. Network model with source and neighbors at each hop.

I. Creating Neighbors List

A list of neighbor sensor nodes are created by every sensor node with particular hop-count that is known as distance interchangeably to sink node by advertising (ADV) data packet that sporadically transmits from sink node to each node afterward every pre-defined time period. Depending on hop-count to sink node, source node just transmits the data packet to sink node by a set of relative neighbors. After initiative time, the network state is constant; and ADV data packet shows part of altering link state among local sensor nodes and neighboring nodes.

J. Collecting Condition of Link\Channel

Initially in case that a source sensor node has data then this node broadcast RTS packet towards neighboring nodes that potential relay nodes and destination to relays packets. The size of a RTS and CTS is lesser then data packet's for reducing the consumption of energy for neighboring sensor nodes. A RTS packet consists of hop count of source sensor node to sink node. As a neighboring sensor node gets RTS packet, it must compare their local hop-count to sink node with hop-count of source sensor node. Just neighboring nodes having shorter distance than source sensor node, possibly become next relay. It restricts number of sensor nodes to join in to rivalry to become destination/ relays to decrease the consumptions of energy and packet loops. After inspection of distance, potential relay replies a CTS packet to source node. Note that according to this technique, a sensor node impossible measure channel status info between itself and destination node such as [8], [14]. It is by ambivalent to the destination node. Though, link state from relay node to destination node is a significant parameter that effects by the selection of relay node. Hence all sensor nodes sensed neighboring link state in terms SNR by eavesdropping data packets from their neighbors and inset average value in CTS.

In case of distributed scenario, more than one source nodes needed to forward data packet, CTS from all neighbors recognizes to strike on links. However, data sensed rate from environ measures for all sensor nodes supposed to minor to decrease data packets collisions. Source sensor nodes estimate Time of Arrival (ToA) and SNR on the reception of CTS packet conforming to particular neighbors. ToA is a packet that travels from a sensor node to other. That parameter is for the measuring physical distance from a source node to a potential relay and terminus. SNR kept conforming to a neighbor that is be around between sensed SNR from CTS at source node while be an average of SNR which is approximate by eavesdropping at every neighbor.

K. Destination/Relays Selection

After the RTS/CTS exchange procedure, the source node achieves a list of candidates with the distance, ToA, and channel conditions, respectively. The source runs Algorithm 1 to select the appropriate and reliable relay nodes to forward the data message.

```
Algorithm 1: Destination/Relays Selection.

Input   :  SNR, Hop_count, ToA by parents Childs with siblings

Output: Destination with two relay nodes

max := total neighbours (n)   with  d(x) ≥  d(n);

for x = 1 to max

    If C (f , l) ≥ R then

            Requirement in Eq10

      If d(x)> d(n) then

          Add with sort using ToA in parents list;

      else

          Add with sort using ToA in siblings list;

    // where d(x) = d(n)

Select (top - down) three members from parents list with siblings for destination and two relay nodes;
```

The siblings and parents are neighbors (n) whom hop-count are lesser and equal to source (x) on link to sink node. Neighbors with hop count lower or equal to the source sensor nodes combined into relays computation. The three parameters used for evaluating a contender with SNR, ToA and hop-count. Initially source node examines computed link capacity correspond to neighbors and requisite rate of data included in candidates list. Afterward the algorithm examines that sensor node is parent or sibling, hence then it includes to particular lists as well as sort by ToA. The timer is also set for getting CTS packet; information of neighbors updates and stops if the timeout happens. Till each candidate is tested or timeout, source sensor node selects three members to be a terminus as well as two relay nodes from two that lists. Source node can select destination or relay node soon on ending of every list for flexibility.

L. Techniques for Diversity Combining

The UW links faces a quite great propagation delays that becomes the reasons of great difference for arriving data signals. Links with source nodes, relay nodes and the destination nodes are greatly lengthier as compared to direct source node to destination route. Therefore the techniques of diversity combining are appropriate choice for processing the signals at destination node, while incoming signals from the source node and relay nodes are operated at packet level, instead of at physical layer. In 1st phase, intermediary relay nodes translates message that comes from a source node and transfer it to next. Destination node then takes one or more copies from the source node and relay nodes. The technique called maximal ratio combing (MRC) is then applied for recovering the messages from source node as well as intermediary relay nodes. We used BER estimation model to compute BER of data packets that's parameters assessed by the technique named maximum likely-hood estimation (MLE).

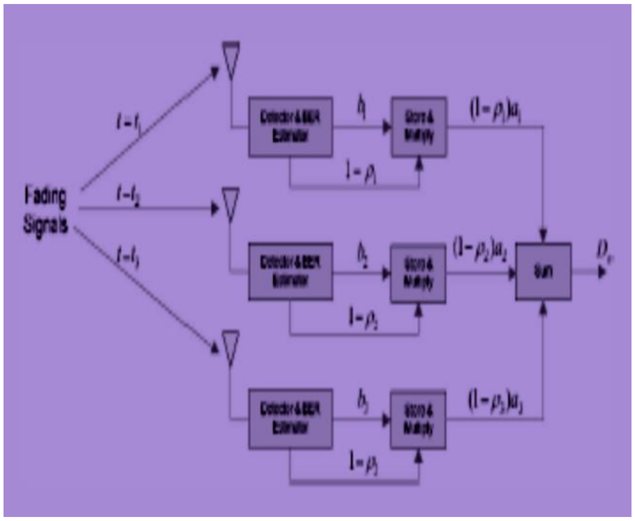


Fig. 5. The maximal ratio combining.

MRC is shown in Fig. 5 by a diagram for Multiple Input and Single Output (MISO) with three inputs. We also used three incoming routes at t1; t2 and t3 times that are from source node, 1st relay node and 2nd relay node. At destination node, arrived signals from link are de-modulated and translated to get a binary result that is bi where i = 1, 2 and 3. Sensed binary sequence is retained and multiplied by the respective weight factors that are 1- Pi for the recovery of original message from various copies. Moreover the bits are transformed to biased value that is a_i prior to multiply, where

$$a_i = \begin{cases} 1 & \text{for binary 1} \\ -1 & \text{for binary 0} \end{cases} \tag{9}$$

The output binary value b is determined by the sum Do based on the threshold which is set to zero.

$$D_0 = \sum_{i=1}^{3}(1 - p_i).a_i\ , \tag{10}$$

$$b = \begin{cases} 1 & D_0 \geq 0 \\ 0 & D_0 \;\; i \;\; 0 \end{cases} \tag{11}$$

M. Routing Model

A K-hop cooperative route ` is a sequence of K links $\{l_1 \dots\dots\dots.. l_k\}$, where each link $l_k = \langle t_k\,,\ r_k \rangle$ is formed between a transmitting node t_k and a receiving node r_k, using the two-stage cooperative transmission at the physical layer. The sequence of links l_k connects a source 's' to a destination 'd' in a loop-free path. Our objective is to find a path that minimizes end-to-end transmission power to reach the destination.

Cost of Link: The link cost $l_k = \langle t_k\,,\ r_k \rangle$ that is represented by following

$$C = (t_k\,, r_k)$$

The above equation can be stated as least expected power of transmission for delivering a message from t_k to r_k by a cooperative transmission that consist of two-stage system that is subject to rate λ with outage probability which is denoted by p_e . Then, the problem of energy efficient routing can be formulated as follows:

$$\min_{l_k \in L} \sum_{l_{k \in l}} C(t_k\,, r_k) \tag{12}$$

In above equation the L represents set of entire possible links in network (any free of loop sequence of sensor nodes from source node to destination node is a possible path of this model).

N. Optimal Cooperation-based Routing

1) Link Cost Formulation: Let we have a link $\langle t_k\,, r_k \rangle$ that is designed between two sensor nodes that are t_k and r_k by two-stage cooperation based transmission. The t_k is the set of cooperative sensor nodes in 2nd stage transmission approach. The power allocation vector to form the cooperative link $\langle t_k\,, r_k \rangle$ is represented by p. Expected cost of cooperative link $\langle t_k\,, r_k \rangle$ represented by C $(t_k\,, r_k)$ that is given by following 6.5 optimization issue,

$$C\ (t_k\,, r_k) = \min_{p \in P} \frac{\sum_{t_i \in T_k} p_i}{S(T_k, p\,, r_k)} \tag{13}$$

The P in above equation represents set of all possible power allocation vectors denoted by p, where $P_{max} \geq\ p_i$ is power that allocated to transmitter $t_i \in T_k$.

The major advantage of cooperation based transmissions is in fading environs in which diversity is used for conflict fading. Applications which take advantage from cooperation based transmissions are usually have a severe obligation in terms of reliability of link. Link cost construction is addressed in (5). Though, it not give any particular target outage probability and no check on the number of re-transmissions with consequently link delay. To control this problem, we alter optimization 5 to comprise a restriction on target outage probability as following.

Suppose p_ϵ represent target per link outage probability which can be accepted. Then cooperative link cost C $(T_k\,, r_k)$ is the solution to the following constrained optimization problem:

$$C\ (t_k\,, r_k) = \min_{p \in P} \sum_{t_i \in T_k} p_i \tag{14}$$

$$\text{Subject to } S(T_k,\ p\,, r_k) \geq 1 - p_\epsilon \tag{15}$$

Till this point, entire cost of transmission to create link $\langle t_k\,, r_k \rangle$ is sum of transmissions power in 1st and 2nd stage. It is expressed by following

Total Power to create $\langle t_k\,, r_k \rangle$ = C $(T_k\,, r_k)$ + P_b

The T_k is cooperation based transferring set that is created in 1st stage.

2) Least Cost Route Selection: Now we model the network as weighted graph that is denoted as following:

G = (N, E, C),

Where N is set of sensors of the network, E is set of entire conceivable links in sensor nodes that are given below:

$E = \{(t_k\ ,\ r_k)|t_k\ , r_k \in N\}$ and

$C = \{C(t_k\ ,\ r_k)|(t_k\ ,\ r_k) \in E\}$

It is set of costs of link that is well-defined over ends. The issue of efficient energy routing now expressed as shortest path problem on G where G is the graph. By Use of Dijkstra's algorithm, least route in source node and destination node calculated by O (N logN), in case if costs of link C is known. Though C is calculated one time and it is possible to calculate offline. The calculating link costs comprise enumerating the exponential amount of cooperative set called T. To ease it, a strategy is reducing search space for T that is addressed in next section.

O. Opportunistic Route

Due of anycast technique, the messages arrive at destination by possibly different paths. A devious path is amalgamation of multiple conceivable paths within the source node and the destination node that is created by a choice of candidate relay nodes at every intermediary sensor node.

1) Anycasted Link Cost: Deliberate a transmitter denoted by t_k with the corresponding candidate relay node set R_k. In phase 1, t_k broadcast a message with power Pb. The sensor nodes which positively took message join t_k to form a cooperation based communicating set T_k . In phase 2, T_k cooperatively anycasted message to entrant relays set R_k.

2) Anycast Link Cost: Anycasted link cost $l_k = \langle t_k, R_k \rangle$ is represented by $C(t_k, R_k)$ which can be stated as least expected power of transmission to send message from t_k to any sensor node in R_k by using two stage cooperation based transmission subject to rate λ with outage probability which is denoted by p_ϵ. Suppose $C(T_k, R_k)$ represent least power that is requisite for cooperation based anycasting from T_k to R_k. Hence $C(T_k, R_k)$ is expressed by following optimization issue:

$$C\ (T_k\ , R_k) = \min_{p \in P} \sum\nolimits_{t_i \in T_k} p_i \qquad (16)$$

$$\text{Subject to}\ \ A(T_k,\ p\ , R_k)\ \geq 1 - p_\epsilon \qquad (17)$$

Where $A(T_k,\ p\ , R_k)$ represents probability which at least one sensor node in set of R_k positively received messages that is uttered as

$$A(T_k,\ p\ , R_k) = 1 - \prod\nolimits_{r_j\ \in R_k} \left(1 - S\left(T_k,\ p\ , r_j\right)\right) \qquad (18)$$

Using (18), anycasted cost of link $C(t_k\ , R_k)$ is expressed by following optimization issue on broadcast power that is denoted by P_b:

$$C(t_k\ , R_k) = \min_{P_b \leq P_{max}} \left\{ \begin{matrix} P_b + \left(1 -\ A(t_k,\ P_b\ , R_k)\right) x \\ \sum_{T \in N-\{t_k\}} S(t_k,\ p_b\ , T)\ .\ C\ (T\ U\{t_k\}, R_k) \end{matrix} \right\} \qquad (19)$$

Where $S(t_k,\ p_b\ , T)$ is given by (15), and $A(t_k,\ P_b\ , R_k)$ can be computed from (18) by substituting $T_k = \{t_k\}$

P. Optimization Parameters

The variables of decision for the issue of global optimization are broadcast power at source, cooperation based transmission power for relay node and sources. For different single hop transmission, certain relay nodes might be engaged by different sources at different timing. Hence different the values of cooperation based transmission power are determined. Assume that ith source contains ni =dim(Ni) neighbor relays, and represent the values of cooperation based transmission power by Pk,i , for any k ∈ {0,1, . . . ,ni}. Hence entire transmission power for ith sensor node is entire transmission power of sensor node in broadcast and cooperative transmissions stages, i.e.:

$$P_{coop,i}\ :=\ P_{b,i} + \sum\nolimits_{k=0}^{n_i} P_{k,i} \qquad (20)$$

Note that $P_{b,i}$ is assigned power for source in broadcast stage and $P_{0,i}$ is cooperation based transmission power for this sensor node in 2nd stage. The values of transmission power must accomplish power constraint at sources, relays and SNR restrictions at relays and at corresponding destination for every source. Set of restrictions for cooperation based transmission from ith source to their destination is represented by

$$C_i\left(P_{b,i}\ , \{P_{k,i}\}_{k\ =\ 0}^{n_i}\right) \leq 0$$

Following stated lemma demonstrate the underlying global optimization issue which is possibly decomposed in to local issues.

Lemma : Let take issue of minimize the entire cooperation based transmission power for sensor nodes as below,

$$\min_{.} \sum\nolimits_{k=0}^{n_i} \left[P_{b,i}\ +\ \sum\nolimits_{k=0}^{n_i} P_{k,i}\right] \qquad (21\ a)$$

$$\text{subject to}\ C_i\left(P_{b,i}\ , \{P_{k,i}\}_{k\ =\ 0}^{n_i}\right) \leq 0\ , i \in \{1, \dots \dots , N\} \qquad (21\ b)$$

The resolution of above can be acquired by resolving given below N individual constrain optimization issues for i = 1, . . . N-1, N,

$$\min\left[P_{b,i}\ +\ \sum\nolimits_{k=0}^{n_i} P_{k,i}\right] \qquad (22\ a)$$

$$\text{subject to}\ C_i\left(P_{b,i}\ , \{P_{k,i}\}_{k\ =\ 0}^{n_i}\right) \leq 0 \qquad (22\ b)$$

Proof: Let the set of values $P^*_{coop,i}\ \{P^*_{coop,i}\ , \{\ P^*_{k,i}\}_{k\ =\ 0}^{n_i}\ \}$

for i = 1, . . . N-1, N, is optimum solution to set of optimization issues (3). The values create a viable solution to optimization issue (2). Variables that are related to ith source not in constraint set for other sources. For proving suitability of solution $\{P^*_{b,i}\ , \{\ P^*_{k,i}\}_{k\ =\ 0}^{n_i}\ \}$ by illogicality and to be supposed that following is a different possible solution

$P'_{coop,i}$ for i = 1, . . . N-1, N with corresponding $\{P'_{b,i}, \{P'_{k,i}\}_{k=0}^{n_i}\}$ values that results in a improved performance index for (2).

It means that there is an index $\bar{\imath}$ for which $P'_{coop,\ \bar{\imath}} < P^*_{coop,\ \bar{\imath}}$ or equally $P'_{b,\bar{\imath}} < P^*_{b,\bar{\imath}}$ or $P'_{k,\bar{\imath}} < P^*_{k,\bar{\imath}}$ for $k \in n_{\bar{\imath}}$. As values $\{P'_{b,\bar{\imath}}, \{P'_{k,\neg\imath}\}_{k=0}^{n_i}\}$ create a possible solution to $\bar{\imath}$-th optimization problem in (3). This produce a improved value for individual optimization issue and this denies optimality of $\{P^*_{b,\bar{\imath}}, \{P^*_{k,i}\}_{k=0}^{n_i}\}$.

In consequence, the issue of minimum power cooperation based transmission is expressed for a source. Subscript i is released henceforth for ease. The effective tackles of mathematical modelling systems are then engaged to convert underlying optimization issue in to mixed integer linear programming issue.

VI. Simulations, Results and Discussion

The proposed protocol Dist-Coop UWSN's performance and effectiveness is evaluated by comparing it with EH-UWSN and SPARCO protocols. 225 nodes having node energy of 0.7 J were randomly placed in underwater and network dimensions used were 500m x 500m x 500m. The protocol was executed for a total of 100 rounds, having average radius value of 100m. Network parameters used in simulation are given below in Table I.

TABLE I. Network Parameters

Parameters	**Values**
Network Volume	500m x 500m x 500m
Total Nodes	225
Relay Nodes	10
Initial Node Energy	0.7 J
Number of Rounds	100
Average Radius value of sensing	100m
Number of Sinks	5
Sensor Node Activation Energy	6.2500e-06
Amplifying Energy (amp)	5.0000e-09
Transmitting Energy (Et)	3.0000e-06

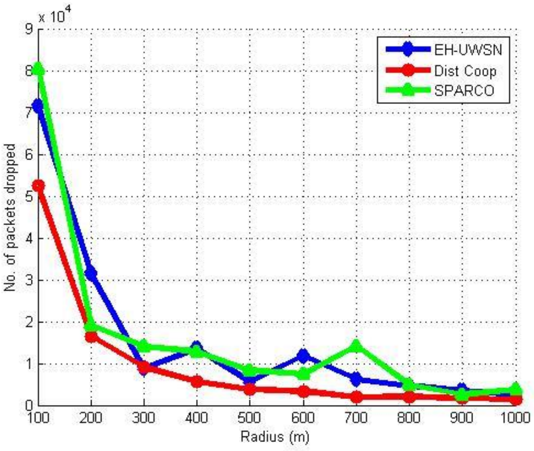


Fig. 6. Number of dropped packets v/s radius (m).

A. Numbers of Dropped Packet

Fig. 6 shows a comparative analysis of packets dropped at the sink while transmitting sufficient data towards the destination in three schemes. Our presented scheme Dist-Coop is compared with two existing schemes SPARCO and EH-UWSN. As our scheme is mainly concentrating on the link reliability and congestion control through the use of cooperation and anycasting strategies, hence Fig. 9 proved that drop of packets in Dist-Coop scheme is much lower as compared to SPARCO and EH-UWSN techniques. Dist-Coop scheme assures that one of the sinks is selected at one time so that data is transmitted in queues and no congestion occurs at the destination sinks. Also opportunist routing enables data to select those paths where path loss is less and link is stable and reliable. Average loss in packets in case of Dist-Coop is 110 while SPARCO and EH-UWSN have 121 and 124 respectively.

B. Energy Tax

The energy tax level of Dist Coop protocol in Fig. 7 is better than the compared schemes. The average energy consumption of Dist Coop in the figure is near to the energy consumption of SPARCO and EH-UWSN as these schemes mainly focused on the decreasing of energy consumption. Our focused in Dist Coop was on congestion control, increasing stability period and throughput as well as load balancing of motes. Therefore the goal is achieved by showing effective outcomes in these parameters. The Dist Coop is also showing improvement in energy consumption as compared to SPARCO and EH-UWSN.

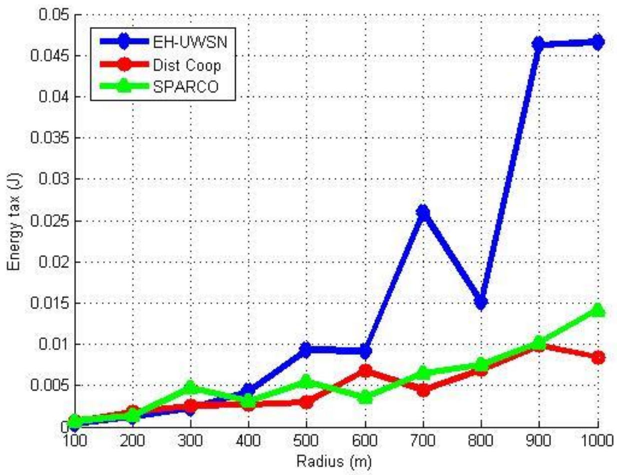


Fig. 7. Energy tax v/s radius (m).

C. Number of Packets Received

The concept here is to deliver data effectively from nodes to surface sink. It is also called Packet Delivery Ratio (PDR). Two types of links used in Dist Coop are from sensor node to forwarder and from forwarder to surface sink. Latter one transfers maximum data than the first one. Therefore, in Dist Coop mostly sensors come close to the surface sink and transfer data packets directly which improves the PDR. Fig. 8 shows major efficiency difference between Dist Coop and SPARCO, EH-UWSN schemes.

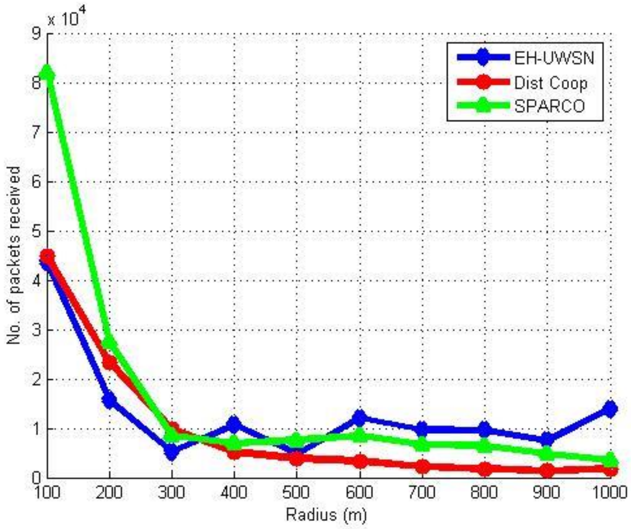


Fig. 8. Received packets v/s radius (m).

D. Network Stability Period

Fig. 9 depicts the performance of Dist Coop protocol which is optimal as compared to SPARCO and EH-UWSN techniques. Stability Period is an operational time of the network until 1st mote expires. At seconds 10000 after network starts, Dist Coop loses 102 nodes and 123 nodes are alive while SPARCO and EH-UWSN lose 140 and 178 nodes respectively. This shows that due to balanced usage of energy by the nodes the performance of Dist Coop is improved.

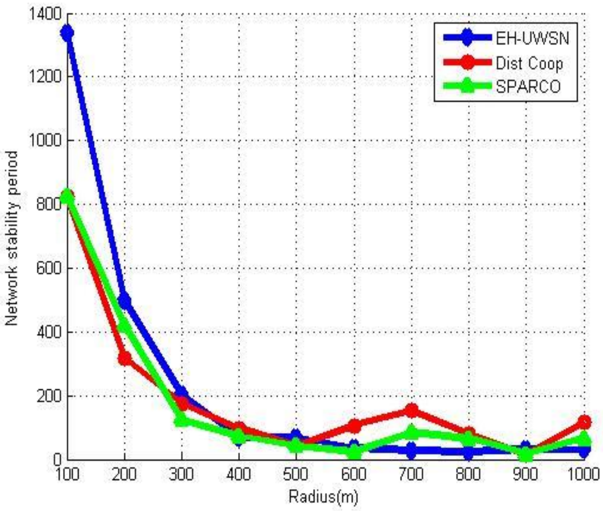


Fig. 9. Network stability period v/s radius (m).

E. Network Lifetime

Stability Period is an operational time of the network until 1st mote expires. Fig. 10 depicts the optimal performance of Dist Coop protocol as compared to SPARCO and EH-UWSN techniques. After 12000 seconds Dist-Coop lost 93 nodes and 132 nodes are alive while SPARCO and EH-UWSN loses 104 and 115 nodes respectively. This better performance of Dist-Coop is due to balanced usage of energy by the nodes.

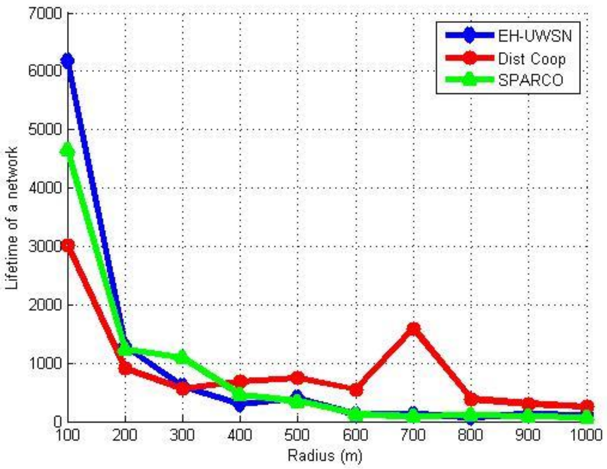


Fig. 10. Network lifetime v/s radius (m).

VII. Conclusion and Future Work

In this research, we have presented a congestion control scheme whose focus is on the network stability period extension and link awareness called Dist-Coop to tackle the aforementioned issues of UWSNs upto some extent. After the analysis of simulation results, we conclude that Dist-Coop showed extraordinarily better execution in connection than the existing schemes i.e. SPARCO and EH-UWSN as far as system Lifetime, Packets Delivery Ratio, Energy Consumption, Packets Received Ratio and Stability Period of the system are concerned.

Dist-Coop has devoured substantially less energy up to 3 times when contrasted with SPARCO and EH-UWSN amid the transmission of information. In Fig. 7, at first from sweep 100(m) to 500(m) the two plans have expended roughly level with energy, however after 500(m) the SPARCO consumed much energy when contrasted with EH-UWSN protocol. Huge contrast is appeared in the normal estimations of Dist-Coop and EH-UWSN plans. In Dist-Coop plot, sooner or later where energy is collected, than little energy is devoured in that segment of the plot like 500(m) to 600(m).

Furthermore, Dist-Coop improved the system lifetime by gathering energy from the earth to make a sensor alive for longer time. In Fig. 8, the underlying life time organize estimation of SPARCO from radius 100(m) to 400(m) is substantially more prominent when contrasted with EH-UWSN protocol. At sweep 400(m), both of the plans have risen to values around. On the off chance that we analyze the plot of EH-UWSN protocol, it plainly demonstrates that its system life time esteems rot fastly when range increments from 100 (m) to 1000 (m). The EH-UWSN esteem bit by bit diminishes when the sweep increments, while plot of Dist-Coop protocol demonstrates energy collecting at various focuses in Fig. 9. At the point when sweep is equivalent to 500(m), Than Dist-Coop protocol has little system life time esteem, so it begins collecting energy up to some level from span 500(m) to 600(m) with a specific end goal to expand the life time of a system. Energy gathering is rehashed as appeared in Fig. 10, when the energy level reductions from the characterize limit esteem. The reproduction comes about shows Improvement in Stability time of the Network in Dist-Coop amid the examination with existing plan SPARCO in Fig. 10. The normal estimation of Dist-Coop protocol is 106% enhanced when contrasted with UWSN existing schemes. As the energy productivity and dependability of a system relies upon the parameters under thought, henceforth clearly Dist-Coop is more energy proficient, link aware and congestion controlled than SPARCO and EH-UWSN schemes.

In our scheme, we utilized SNR consolidating methodology (SNRC) at destination to join the signals received from various directions. In future, we can utilize distinctive different methodologies like, Fixed Ratio Combining (FRC), Maximal Ratio Combining (MRC) and Selection Combining (SC) at destination to inspect and contrast the results and the introduced strategies so as to outline more productive routing schemes for various UWSNs and WSNs environments. By utilizing this method, researchers can additionally enhance system life time by growing further developed protocols.

References

[1] Kaur, and S. Monga, "Comparisons of wired and wireless networks: A review." International Journal of Advanced Engineering Technology 5, no. 2 (2014): 34-35.

[2] Frodigh, P. Johansson, and P. Larsson. "Wireless ad hoc networking: the art of networking without a network." Ericsson review 4, no. 4 (2000): 249.

[3] Penttinen, "Research on ad hoc networking: Current activity and future directions." Networking Laboratory, Helsinki University of Technology, Finland. See also http://citeseer. nj. necm. com/533517. html (2002).

[4] Wahid and K. Dongkyun, "Analyzing routing protocols for underwater wireless sensor networks." International Journal of Communication Networks and Information Security 2, no. 3 (2010): 253.

[5] Heidemann, W. Ye, J. Wills, A. Syed and Y. Li, "Research challenges and applications for underwater sensor networking." In Wireless Communications and Networking Conference, 2006. WCNC 2006. IEEE, vol. 1, pp. 228-235. IEEE, 2006.

[6] Heidemann, M. Stojanovic and M. Zorzi. "Underwater sensor networks: applications, advances and challenges." Phil. Trans. R. Soc. A 370, no. 1958 (2012): 158-175

[7] Wahid and D. Kim, "An energy efficient localization-free routing protocol for underwater wireless sensor networks." International journal of distributed sensor networks 8, no. 4 (2012): 307246.

[8] S. Ahmed, I. U. Khan, M. B. Rasheed, M. Ilahi, R. D. Khan, S. H. Bouk and N. Javaid,"Comparative analysis of routing protocols for under water wireless sensor networks." arXiv preprint arXiv:1306.1148(2013).

[9] M.R. Jafri, S. Ahmed, N. Javaid, Z. Ahmad and R. J. Qureshi, "Amctd: Adaptive mobility of courier sensors in threshold-optimized dbr protocol for underwater wireless sensor networks." In Broadband and Wireless Computing, Communication and Applications (BWCCA), 2013 Eighth International Conference on, pp. 93-99. IEEE, 2013.

[10] Javaid, M. R. Jafri, Z. A. Khan, U. Qasim, T. A. Alghamdi and M. Ali, "Iamctd: Improved adaptive mobility of courier sensors in threshold-optimized dbr protocol for underwater wireless sensor networks." International Journal of Distributed Sensor Networks 10, no. 11 (2014): 213012.

[11] Javaid, M. R. Jafri, S. Ahmed, M. Jamil, Z. A. Khan, U. Qasim and S. S. Al-Saleh, "Delay-sensitive routing schemes for underwater acoustic sensor networks." International Journal of Distributed Sensor Networks 11, no. 3 (2015): 532676.

[12] A. Khan, N. Javaid, A. Majid, M. Imran and M. Alnuem, "Dual sink efficient balanced energy technique for underwater acoustic sensor networks." In Advanced Information Networking and Applications Workshops (WAINA), 2016 30th International Conference on, pp. 551-556. IEEE, 2016.

[13] Majid, I. Azam, T. Khan, Z. Ali Khan, U. Qasim and N. Javaid, "A reliable and interference-aware routing protocol for underwater wireless sensor networks." In Complex, Intelligent, and Software Intensive Systems (CISIS), 2016 10th International Conference on, pp. 246-255. IEEE, 2016.

[14] Ahsan, S. Ahmed, F. Hadi, F. Wahab and I. Ahmed, "A Recent Study on Routing Protocols in UWSNs."

[15] R. Islam and Y. S. Han, "Cooperative MIMO communication at wireless sensor network: An error correcting code approach." Sensors 11, no. 10 (2011): 9887-9903.

[16] Zhou, S. Zhou, J.H Cui and S. Cui, "Energy-efficient cooperative communication based on power control and selective single-relay in wireless sensor networks." IEEE transactions on wireless communications 7, no. 8 (2008).

[17] Wu, W. Liu and K. Li, "Power allocation and relay selection for energy efficient cooperation in wireless sensor networks with energy harvesting." EURASIP Journal on Wireless Communications and Networking 2017, no. 1 (2017): 26.

[18] Chen, M. Ma, X. Liu, A. Liu and M. Zhao, "Reliability Improved Cooperative Communication over Wireless Sensor Networks." Symmetry 9, no. 10 (2017): 209.

[19] Bravo, E. Palomar, A. Gardel and J. L. Lázaro, "Trusted and Secure Wireless Sensor Network Designs and Deployments." (2017): 1787.

[20] Gu, H. Chen, Y. Li and B. Vucetic, "Distributed multi-relay selection in accumulate-then-forward energy harvesting relay networks." arXiv preprint arXiv:1602.00339 (2016).

[21] Umar, M. Akbar, S. Ahmed, N. Javaid, Z. A. Khan and U. Qasim, "Underwater wireless sensor network's performance enhancement with cooperative routing and sink mobility." In Broadband and Wireless Computing, Communication and Applications (BWCCA), 2014 Ninth International Conference on, pp. 26-33. IEEE, 2014.

[22] Nasir, N. Javaid, H. Ashraf, S. Manzoor, Z. A. Khan, U. Qasim and M. Sher, "CoDBR: cooperative depth based routing for underwater wireless sensor networks." In Broadband and Wireless Computing, Communication and Applications (BWCCA), 2014 Ninth International Conference on, pp. 52-57. IEEE, 2014.

[23] Nasir, N. Javaid, M. Murtaza, S. Manzoor, Z. A. Khan, U. Qasim and M. Sher, "ACE: Adaptive cooperation in EEDBR for underwater wireless sensor networks." In Broadband and Wireless Computing, Communication and Applications (BWCCA), 2014 Ninth International Conference on, pp. 8-14. IEEE, 2014.

[24] Ahmed, M. Akbar, R. Ullah, S. Ahmed, M. Raza, Z. A. Khan, U. Qasim and N. Javaid, "ARCUN: Analytical approach towards reliability with cooperation for underwater WSNs." Procedia Computer Science 52 (2015): 576-583.

[25] Hafeez, N. Javaid, U. Shakeel, S. Hussain and H. Maqsood. "An Energy Efficient Adaptive Cooperative Routing Protocol for Underwater WSNs." In Broadband and Wireless Computing, Communication and Applications (BWCCA), 2015 10th International Conference on, pp. 304-310. IEEE, 2015.

[26] Javaid, H. Maqsood, A. Wadood, I. A. Niaz, A. Almogren, A. Alamri and M. Ilahi, "A localization based cooperative routing protocol for underwater wireless sensor networks." Mobile Information Systems 2017 (2017).

[27] S. Ahmed, N. Javaid, F. A. Khan, M. Y. Durrani, A. Ali, A. Shaukat, M. M. Sandhu, Z. A. Khan and U. Qasim. "Co-UWSN: cooperative energy efficient protocol for underwater WSNs." International Journal of Distributed Sensor Networks (2015).

Traffic Predicting Model for Dynamic Spectrum Sharing Over 5G Networks

Ahmed Alshaflut**,** Vijey Thayananthan
Faculty of Computing and IT
King Abdulaziz University,
Jeddah, Saudi Arabia

***Abstract*—Recently, wireless networks and traffic requirements have been rapidly aggregated in diverse applications in 5G environments. For this reason, researchers have investigated the influences of this growth based on a user's requirements inside these networks. However, the stream of traffic has been considered a crucial role for the user's needs over 5G network. In this paper, gigantic data traffic is considered for enabling dynamic spectrum sharing over 5G networks. Thus, various accessing plans are covered to manage the overall network traffic. Additionally, it proposes a traffic predicting model for a technique of managing traffic when multiple requests are received to decrease delays. It has considered different significances related to a large size of traffic practices. Additionally, this work will guide us to enhance traffic solutions within massive requests over outsized networks. Systematically, it has focused on the traffic flow, starting from the accessing steps until passing on requests to suitable spectrum carriers.**

Keywords—Component; traffic predictions; software defined multiple access; dynamic spectrum sharing; 5G networks

I. Introduction

Wireless technologies are increasingly being improved over a variety of scopes, with some of these improvements going far beyond the current needs of users. Reasons pushing this rapid growth are mostly due to either enhancing capacity or reducing latency. There is an obvious need to provide more services for different users without consuming more resources. Enormous applications need to be guaranteed in order to satisfy more users and this has simultaneously increased the cost of the providing the systems necessary for these applications. Thus, traffic issues are considered one of the biggest challenges facing the next generations. Accordingly, systems must achieve the delivery of different services by considering both the system's capacities and quality of service.

The importance of traffic predicting is to balance between a system's capacities and a user's requirements. It can manage to lessen the effects of the system's limitations. It is important to plan using a system's resources as well as to provide services for different users equally. Moreover, predicting traffic assures the correct resource allocations for enormous systems.

Different 5G networks have promised to provide different services in the best possible way for a variety of circumstances. Hence, it is an important to keep the delay at the minimum rate. As a result, this work has focused on enhancing the performance of 5G networks as well as the quality of service. It balances between both growing number of services and the large amount of users simultaneously. Primarily, choosing efficient traffic techniques improves user's experience and the quality of the provided services.

Principally, this work investigates predicting traffic by employing the recent technologies of different traffic concerns. Furthermore, it adopts fairness practices for dynamic spectrum sharing. This approach would assure the best delivery of provided services to end users by managing the available resources. It deals with each service as an individual request, and then current and previous requests are measured for forthcoming reactions. This solution will consider different requirements for providing fair spectrum sharing to enable easier service delivery. Thus, Section II presents the state of the art of data traffic management. After that Section III presents the traffic prediction model. Moving forward to operational scenarios as well as indicative results is presented in Section IV. Then, the analytical section is presented in Section V. Finally, Section VI presents the conclusion and further directions.

II. State of the Art

This section discusses the data traffic management in 5G networks. Also, it presents the concept of SoDeMa as well as operational behaviors of SoDeMa. Furthermore, it would cover the concept of spectrum sharing for different scenarios.

A. Data Traffic in 5G Networks

Recently, traffic management has been raised with different fragments of large systems while traffic issues are being taken into consideration when designing large applications. Also, chosen data transmission mechanism has been gathering greater attention in recent developments, and challenging considerations have been presented by several researchers in the 5G era. Network traffic is expected to increase every nine months. As a reason, traffic loads will be increased over the next few years [1]. This massive growth causes raising access control issues for large applications. Several schemas have been proposed for managing traffic. For example, authors of [2] have taken on the traffic issues using different mechanism by Software-Defined Networks (SDN). They have shown the significance proposing new traffic management mechanism to enhance the quality of services. Once more, the growth of connected devices and their consequential needs is studied by authors of [3]. Therefore, an SDN controller is proposed for controlling dense networking requirements.

B. *Software-Defined Multiple Access*

In the massive applications's growth, decentralized management was a precise and significant solution for network management. Recently, several types of research have offered different mechanisms, including Non-orthogonal Multiple Access NOMA instead of orthogonal access. Mainly, NOMA was proposed to increase the spectral efficiency [4]. However, for flexibility purposes, SoDeMa is recommended in terms of access management schemas within expandable networks [5]. Though both SoDeMa and NOMA achieve retrieving tasks, but SoDeMa can enable an easier configurability. Obviously, this technique can be implemented by choosing an appropriate access technique. Fig. 1 demonstrates SoDeMa with different access mechanisms. However, both availability and user's requirements will indicate the most suitable schema.

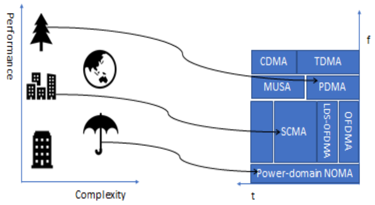


Fig. 1. SoDeMa design [5].

C. *Different Uses of SoDeMa*

SoDeMa is mainly to enhance numerous features of flexible programmability solutions. In fact, it was mainly proposed for providing higher resource allocations in large applications. It has mostly used several features from both NOMA and Software Defined Radio SDR. Thus, it efficiently allows requests to use appropriate NOMA schemas. Furthermore, SoDeMa enhances the quality of service alongside with supporting many services through 5G standards thus providing more options for applications to use the suitable schema.

D. *Spectrum Sharing in 5G*

In fact, 5G networks have undertaken serving enormous amounts of users in different applications types. Though, this can increase data traffic issues in large environments. Correspondingly, it requires efficient techniques of different resources to be effectively allocated. One of these resources is a spectrum that is defined as "precious limited resources required to build, maintain and expand of the Information and Communications Technology (ICT) infrastructure of any nations" [6]. Thus, efficient strategy and procedures for managing spectrum are significant points of deployment and coexistence of wireless technologies, services, and applications. Furthermore, spectrum management is a sensitive task that deals with incredible numbers of users and devices. Techniques of sharing spectrum are key facts for satisfying large applications users. Spectrum sharing has been deployed in several approaches based on the topology used and the user's preferences. Commonly, spectrum sharing is explored more fully in the following scenarios:

1) Licensed Approach: Spectrum in this scenario is allocated as primary use to wireless systems. Homogeneous horizontal sharing is the only relevant scenario of spectrum sharing. It is mainly considered as the least space of sharing among other modes since it has two methods to achieve sharing, namely limited spectrum pool as well as mutual renting. Mutual renting acts with bands of spectrum resources as subdivided blocks. Each block is licensed to a particular operator. The actual spectrum sharing occurs at this point since it operates mutually and can allow for renting limited parts of their licensed resources. Simultaneously, operators can rent resources from several other operators. However, the genuine owner of a resource has the highest level of priority among the operators accessing its resources. Limited spectrum pool allows operators to obtain a license to facilitate the band on shared basis but with limited numbers of authorized users. It could give an indication for current users of others' needs and agreements to provide more transparency.

2) Licensed Shared Access Approach: Initially, spectrum in this mode is licensed. Thus, the unused spectrum of an incumbent user can be accessed at particular times and locations. In fact, well-defined conditions determine the decision of sharing spectrum. Licenses tended to be active for long-term agreements. However, because of updates, the level of agreement varies. Thus, agreements need to be reissued. Several developments have evolved to manage the time and location preferences, and several reports have been published to enhance this mode. In fact, this approach has been leveraged from the full licensed approach but with more regulations for future systems.

3) Unlicensed Approach: In the unlicensed mode, the spectrum is wirelessly shared with other unlicensed systems. Thus, it enables more flexibility but with more required sharing regulations since this mode involves the most known scenarios of spectrums sharing cases. In this approach, both heterogeneous and homogeneous horizontal sharing modes are combined. Additionally, vertical sharing is done in the case of existing primary users. Therefore, this mode has multiple deployment scenarios and then needs more regulations to manage the perfect sharing among different systems. Different schemas have been proposed for dealing with spectrum sharing issues. Thus, several works have been using NOMA approaches in order to enable for more connectivity and massive users requirements [7]. Furthermore, others are combining both NOMA with cognitive radio technologies to leverage from both technologies in heterogeneous networks [8], [9]. However, further works have been proposing other techniques to enable the best of shared spectrum among 5G networks based on reconfigurable SDN technologies in massive devices [10]-[12]. Also, other researchers have focused on D2D based spectrum sharing for vehicular purposes to enable massive connections such as [13]. The spectrum is envisioned that 5G will be 10 times improved to satisfy the 1000 times of throughput improvements [14].

III. Traffic Predicting Model

This model focuses on predicting traffic of different requests for enabling an easier spectrum sharing among large systems. Thus, we will cover the entire process of the traffic predicting model from the beginning to assigning to the suitable spectrum sharing provider. Predicting traffic in 5G networks involves examining different reasons for solving this issue. Therefore, our solution enhances traffic predicting in different ranges. This means we have considered suitable accessing schemas, for serving user's needs at the best of resource saving. Also, it has considered the Quality of service in terms of spectrum sharing. In addition, it has focused on each slot of time with it served for estimation purposes.

This work applies two factors, starting by calculating the exact duration for each request in order to use the suitable spectrum carrier. This step is significant to allocating network resources. Consequently, requests are considered based on different ranges of priority to calculate the needed time. Requests mainly fall under the offered bandwidth within this system. It is a key point that a serving request goes to the nearest slot of time. Thus, it also accounts for the number of requests for future traffic issues. At this time, requests are classified by the traffic controller, based on the needed time and their types. Furthermore, this controller assigns each request to an appropriate accessing mechanism after calculating the time needed. However, spectrum carrier is individually considered based on its capacity. Consequently, it is necessary to identify the capacities of carriers to deal with future requests. Afterward, the assigning step assures accounting for both times needed for each request. Then, it calculates the traffic at the particular slot of time for calculating future traffic. Accordingly, previous steps are considered as the first scenario of this model.

The traffic controller assigns different requests to the needed service. Then, it leverages from current services for future estimations. Users are served based on their needs as well as the available resources. Before assigning to the spectrum carrier, decisions are already made by the controller for enabling an easier resource allocation. The history of previous requests is saved to estimate future traffic approximations. The future estimations are based on both location and times of requests. Thus, frequent requests are classified into their locations and the slot of time to be served with suitable spectrum carrier in future experiences.

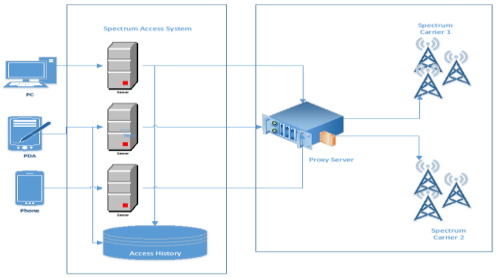


Fig. 2. An entire system model.

Our model is divided into two main scenarios. Scenario one internally measures traffic to provide requests with the time approximations as shown in Fig. 2. However, scenario two has calculated different factors of current requests for future traffic purposes. Therefore, it will consider the most frequent carrier for future purposes. It will not repeat the calculations of scenario one for the future calculated request. Thus, recording the most frequent decisions will be saved for specific requests. In fact, this step would help for more resource savings and better allocations. Generally, this includes implementing two steps in order to deal with future requests. Since scenario one will deal with the new requests, however scenario two analyzes the previous requests as well as their requirements based on the accessible resources.

In fact, replication process of scenario one is avoided by calculating frequently received requests. Moreover, it will assist in reducing the consumption of systems resources. Then, it would decrease the needed time when serving requests. This would enable for an easier decision making especially for a massive number of users to the suitable spectrum provider. Thus, a particular spectrum carrier will be assigned by the time of the systems. Therefore, it benefits both asynchronous requests and requests with real-time requirements. Different schemas will be classifies to serve frequent requests. Also, timestamps will help in determining the frequent request and the suitable used schemas.

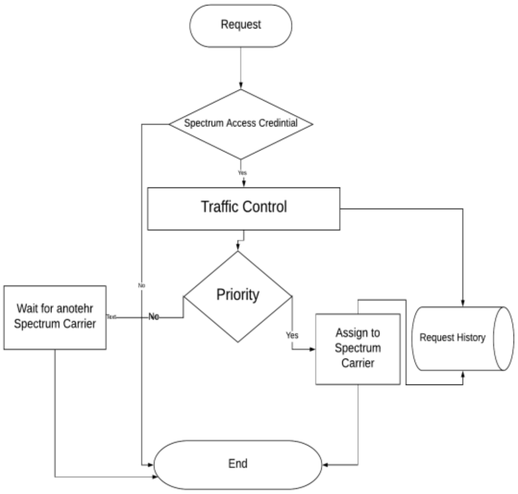

Request
Spectrum Access Credintial
Yes
Traffic Control
No
Priority
Wait for anotehr Spectrum Carrier
Text
No
Yes
Assign to Spectrum Carrier
Request History
End

Fig. 3. Simple request process of dynamic spectrum sharing.

As seen in Fig. 3, a simple process is used for incoming requests. In the beginning, it would be treated at the level of credentials. If this request is authorized, it would be processed to the next step. Then, after the request has been processed, the traffic control will be in charge of dealing with this request based on its previous experiences. During the first time, this request will be recorded for future purposes. However, if the coming request is regular, the control will check with its stored previous experiences. This will help determine the initial priority needs. Next, this request will be processed further in the priority stage. In the priority step, the request will be treated based on its previous and current needs. This step accounts for the accurate traffic levels for better decision making. Then, again, it will assign the request to most suitable spectrum sharing. For enabling dynamic sharing, each request is examined twice while processing for enabling more dynamicity of spectrum sharing. In fact, requests are served with an easier technique for analyzing previous experiences. The difference can be recognized through the benefit of using previous requests for future optimizations. Thus, it is focused on choosing the suitable spectrum sharing to manage the overall network traffic.

Moving to the main focus, estimating future traffic to enable better decision support, we have to consider the situations of past requests. Thus, the time taken for the previous request will be an important factor for our upcoming requests. This has benefits in the managing of resources and predicting the needs of current requests. However, the predicting phase depends on available bandwidth as well as the priority of the request. These predictions will be based on both the slot of time given as well as the network experience of previous requests. Henceforth, it helps in future traffic estimations to assign the incoming requests to particular carriers.

IV. Operational Scenarios and Indicative Results

This section describes the showcases of Estimating Data Traffic, followed by indicative results for managing the traffic among heterogeneous networks by using SoDeMa technology.

A. *Scenario 1: SoDeMa*

As presented in Section II, SoDeMa can play an important role in managing traffic within enormous networks, by enabling flexible configuration when choosing the suitable access schema. However, in this proposal, it is very important to manage the traffic to assign the incoming request to a suitable spectrum carrier. Thus, we will be discovering several schemas where SoDeMa can be adopted for enhancing the overall throughput as well as managing the traffic issues. We are assuming that several users with different requests wish to access an application with different privileges. Our scenario will focus on providing them with the best service while considering their bandwidth capacities. To be clear, overall latency requirements must be considered for service estimations. However, it is also obvious that showing current needs is significant for priority classifications. Users are informed by messages for spectrum capacities. Once the channel is free for incoming requests, users will be informed for assigning them to a suitable spectrum carrier.

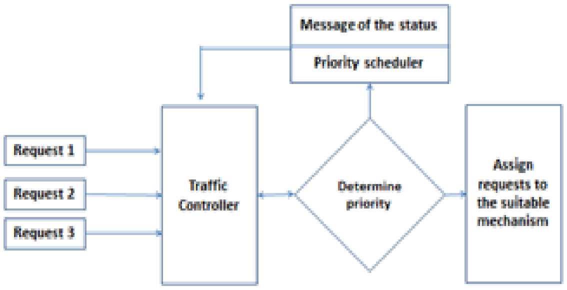


Fig. 4. SoDeMa architecture in 5G environments.

As seen in Fig. 4, incoming requests will be assigned to the traffic controller. This step can measure internal traffic to provide users with the service approximations. Also, this controller determines the priority of needed services in a dynamic method. The controller notifies users of the traffic status. However, to estimate the time of achieving different services, users will be notified by a message including both status and times needed. Therefore, a notifying message will be suitable for time management as well as for priority purposes.

Accordingly, at this step, computing the time needed for each request would be critical factor. All requests are individually considered based on the different priority levels. Thus, the available spectrum carriers results on serving different requests. Furthermore, it is an important to serve requests within the nearest slot of time. The controller will record the number of requests for future traffic purposes. This controller assigns the request to the suitable spectrum carrier after calculating the needed time. Accordingly, the capacities of available spectrums needs to determined to deal perfectly with current requests. In fact, this step assures calculating both the traffic within the slot of time and the time needed for each request.

B. *Scenario 2: Predicting Data Traffic*

At scenario 1, it is an important to calculate the time needed for serving each request, depending on both priority and the available spectrum. Hence, this scenario focuses on additional aspect of the traffic process. Thus, it calculates the most frequent spectrum carrier for flexible future assigning steps. Thus, it will overcome the calculations of scenario 1 for the future request. Therefore, recording repeated notifying message will be saved for particular requests. This scenario was achieved by implemented two steps as presented below:

1) Frequent requests: At this step, the time slot of the system's availability will be assigned to particular multiple accesing schemas. Thus, it helps supporting asynchronous requests or requests with real time needs. Thus, frequent requests will be kept as an array in the system's reccords. Requests will be identified by the time stamp to extract frequent requests. As a result, this step analyzes the current situation of the request and provides the history of decisions made in order to support future choices. Also, it will focus on the time of the request to for priorities practices.

2) Calculating data volume: It has considered the situations of past requests, in order to estimate future traffic. Thus, it enables for better decision making. Therefore, the previously time taken acts as a significant factor in all the future requests. Urgent requests will be directly assigned to a spectrum carrier. Thus, estimation steps firstly depend on both slot of time, then on the experience of spectrum sharing. Thus, the future traffic decisions are estimated by returning past made decisions of similar request.

C. *Scenario 3: Dynamic Spectrum Sharing*

Previous scenarios have mainly participated in achieving this scenario for dynamic sharing of spectrum. This supports an integrated solution in our proposal. For dynamicity support, SoDeMa enhances decision making for choosing the best available carrier. This decision is made according to different factors. These factors include time taken for the previous request, quality of experiences and so on. Individually, each request is considered in order to provide the best service as well as taking into consideration the quality of service. Furthermore, in scenario 2, predicting data traffic enhances the current status for the systems to estimate future traffic based on current traffic management as well as the current system's capacity. Thus, this scenario allows dynamic sharing of the spectrum based on past request experiences. However, it leverages the current requests for two reasons. The first is for predicting the traffic in the future. The second is for recording the current traffic of requests based on the time of the request as well the time taken for achieving this request.

As seen in Fig. 5, the scenario of dynamic spectrum needs to be shared perfectly. Thus, requests need an optimized assigning at the end of the process. The control unit classifies carriers into four main types. Green color means the carrier is ready for receiving the new request based on its capacity. A black path means the carrier is currently available, but another request just has been assigned to share this spectrum. Furthermore, a red color means this carrier is not available at this time and serving another request. The time taken for serving this request is very important for the next request to perfectly manage the time. The blue color means this carrier is prepared for another request and this color also provides the capacity limitations for dealing with new requests. Dynamically, colors change based on the current situations of the spectrum carrier. Thus, each status is understood by the traffic controller and the approximations of time. Colors are changed based on the notifying messages sent continuously to the controller. This approach will assure that all requests are treated equally, and at the same time, the status of the spectrum and providers are accurate.

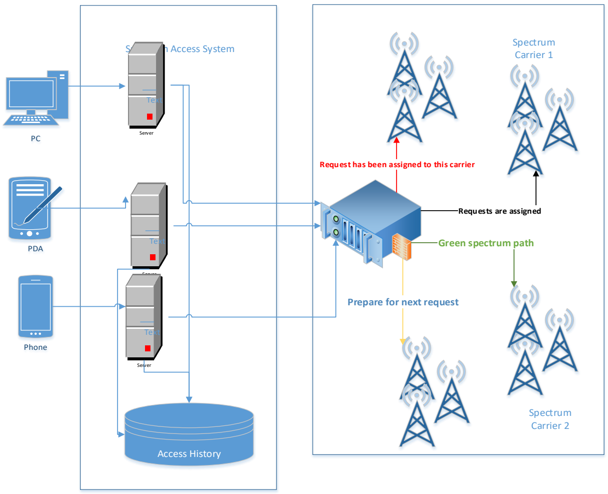


Fig. 5. Dynamic spectrum sharing scenarios.

V. Analysis

Traffic conditions are a significant factor in this work. This section analyzes the proposed model in order of predicting traffic. In fatc, this work is focused to be implemented on the future 5G networks. Therefore, we focused on several factors including spectrum sharing, time, packet loss, and frequent request reactions. Significantly, it is important to study the existing traffic solutions for investigating their usages within SDN and SoDeMa.

A. Time Management

This work invitistaiges traffic predicting by recent technologies for heterogeneous networks. Thus, time management is a crucial aspect and must be addressed for traffic enhancements. As a consequence, time management for enormous numbers of requests is an identical issue. Hence, it has focused on keeping the delay of requests at the minimum rates. Also, the time taken is considered in all steps of this work. Correspondingly, this solution ensures reducing time consumptions for future networks by predicting the traffic approximations, which will help decision making of frequent requests. Again, it has prevented the repetition of scenario 1 steps for frequent incoming requests. Also, classifying the user's services is implemented to assign them to the suitable spectrum provider.

B. Spectrum Sharing Schema

Another important issue has been dicovered in our model. This was regarding spectrum sharing which was solved by distributing the service colors. It was classified by based on the availability and frequently booked channels, as shown in Fig. 5. Thus, it is classified into four main colors, for representing different scenarios. This has helped classify the different users as well as to estimate the time of assigning the desired service to the suitable carrier. On the other hand, priority considerations were considered by implementing the quality of service of different users. In fact, several factors are affected in this issue including the large and various services. Also, the changeability of available spectrums is another factor at this step. For future considerations, each request is maintained by the needed time of delivering services.

C. Packet Loss

Huge data transmission could result on packet loss, which can be considered as a bottleneck in 5G systems. This is caused by the increasing number of users as well as the hug data traffic. This system has overcome this issue by estimating time for frequent users. Thus, the frequent user is given an estimated time before resending packets. Mainly, this solution is implemented by giving an approximate time before sending any packet. The time given is calculated by consdiering several factors as well as the available spectrum. Thus, requests will be given a range of time until it is serve, otherwise it will resend the lost packets.

D. Frequent Request Considerations

Mainly, the proposed model focuses on saving both time and network resources. Thus, requests at the first time will be served, but will not be considered as frequent ones. They, will be served and then enter the chain of previous request for better decision making in fture. After serving first time request, the next times will be easier for the systems to enable them with frequent user'sconsiderations. At the next steps, it will enable for more flexibility since scenario one procedures is avoided. However, they will be saved for the network records for future purposes. Though, it will not solve all the issues of coming requests but at least will consider frequent users request and then enable for a faster resources allocating.

VI. Conclusion and Further Directions

In conclusion, this model has investigated traffic predicting in 5G networks. Thus, It has proposed predicting data traffic by using a new technology, called SoDeMa. This dynamically enhances spectrum sharing process. Moreover, it has concluded that managing traffic, starting from the access schema would improve the throughput of large applications within 5G networks. It can be considered as an efficient approach to track the current traffic issues. The main difference in this model is to overcome the replications of dealing with different requests in future experiences. The outlined advantages of this paper are to solve traffic management through a new sharing mechanism. Thus, applying this technique would insure flexible services in heterogeneous networks. Also, it can flexibly achieve requests of large numbers of users.

Future directions are mainly based on the new requirements of the next generation. Coming generations will be mainly focusing on enhancing the service provided. Other attempts should investigate the load of traffic prediction on the overall throughput. Finally, the changeability status must be accurate for enabling more efficiency when applying dynamic sharing among spectrum carriers.

References

[1] B. Bangerter, S. Talwar, R. Arefi, K. Stewart, "Networks and devices for the 5G era", IEEE Comm. Mag., vol. 52, no. 2, pp. 90-96, Feb. 2014.

[2] K. Kosek-Szott et al., "Coexistence Issues in Future WiFi Networks," in IEEE Network, vol. 31, no. 4, pp. 86-95, July-August 2017.

[3] A. Muthanna, R. Gimadinov, R. Kirichek, A. Koucheryavy and M. S. A. Muthanna, "Software development for the centralized management of IoT-devices in the “smart home” systems," 2017 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus), St. Petersburg, 2017, pp. 190-194.

[4] J. Choi, "NOMA: Principles and recent results," 2017 International Symposium on Wireless Communication Systems (ISWCS), Bologna, Italy, 2017, pp. 349-354.

[5] Dai. L., Wang, B. Yuan, Y. Han, S., I, C. and Wang, Z., ” Non-orthogonal multiple access for 5G: solutions, challenges, opportunities, and future research trends”. IEEE Communications Magazine, 2015, 53(9), pp.74-81.M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.

[6] R. B. F. Da Silva and C. T. R. Da Silva, “Spectrum regulation in Brazil,” IEEE Wireless Communications, vol. 23, no. 3, pp. 2-3, 2016.

[7] L. Lv, J. Chen, Q. Ni, Z. Ding and H. Jiang, "Cognitive Non-Orthogonal Multiple Access with Cooperative Relaying: A New Wireless Frontier for 5G Spectrum Sharing," in *IEEE Communications Magazine*, vol. 56, no. 4, pp. 188-195, APR, 2018.

[8] Y. Liu et al., "Nonorthogonal Multiple Access In Large-Scale Underlay Cognitive Radio Networks", *IEEE Trans. Vehic. Technol.*, vol. 65, no. 12, pp. 10152-57, Dec. 2016.

[9] L. Lv et al., "Design of Cooperative Non-Orthogonal Multicast Cognitive Multiple Access for 5G Systems: User Scheduling and Performance Analysis", *IEEE Trans. Commun.*, vol. 65, no. 6, pp. 2641-56, June 2017.

[10] F. Mekuria and L. Mfupe, "Spectrum sharing & affordable broadband in 5G," *2017 Global Wireless Summit (GWS)*, Cape Town, 2017, pp. 114-118.

[11] X. Duan, X. Wang, L. Zhang, W. Li and Y. Wu, "Software defined orchestrated spectrum sharing enabled by 3D interference map," *2017 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*, Cagliari, 2017, pp. 1-6.

[12] X. Duan, X. Wang, "Authentication handover and privacy protection in 5g hetnets using software-defined networking", *IEEE Communications Magazine*, vol. 53, no. 4, pp. 28-35, April 2015.

[13] J. Liu et al., "Device-to-Device Communication in LTE-Advanced Networks: A Survey", *IEEE Commun. Surveys & Tutorials*, vol. 17, no. 4, pp. 1923-40, 2015.

[14] L. Zhang, M. Xiao, G. Wu, M. Alam, Y. C. Liang and S. Li, "A Survey of Advanced Techniques for Spectrum Sharing in 5G Networks," in *IEEE Wireless Communications*, vol. 24, no. 5, pp. 44-51, October 2017.

[15] A. Alshaflut and V. Thayananthan, "Estimating data traffic through software-defined multiple access for IoT applications over 5G networks," *2018 15th Learning and Technology Conference (L&T)*, Jeddah, 2018, pp. 59-66.

Implication of Genetic Algorithm in Cryptography to Enhance Security

Muhammad Irshad Nazeer, Ghulam Ali Mallah, Noor Ahmed Shaikh
Shah Abdul Latif University, Khairpur,
Sindh, Pakistan

Rakhi Bhatra, Raheel Ahmed Memon, Muhammad Ismail Mangrio
Department of Computer Science Sukkur IBA University,
Sukkur, Sindh Pakistan

***Abstract*—In today's age of information technology secure transmission of information is a big challenge. Symmetric and asymmetric cryptosystems are not appropriate for high level of security. Modern hash function based systems are better than traditional systems but the complex algorithms of generating invertible functions are very time consuming. In traditional systems data is being encrypted with the key but still there are possibilities of eavesdrop the key and altered text. Therefore, key must be strong and unpredictable, so a method has been proposed which take the advantage of theory of natural selection. Genetic Algorithms are used to solve many problems by modeling simplified genetic processes and are considered as a class of optimization algorithms. By using Genetic Algorithm the strength of the key is improved that ultimately make the whole algorithm good enough. In the proposed method, data is encrypted by a number of steps. First, a key is generated through random number generator and by applying genetic operations. Next, data is diffused by genetic operators and then logical operators are performed between the diffused data and the key to encrypt the data. Finally, a comparative study has been carried out between our proposed method and two other cryptographic algorithms. It has been observed that the proposed algorithm has better results in terms of the key strength but is less computational efficient than other two.**

Keywords*—Secure transmission; symmetric cryptosystems; invertible functions; genetic algorithms; efficient encryption

I. Introduction

Recently, secure data transmission over network has become a vital and critical issue due to increased demand of digital media transmission and unauthorized access of important data [1]. Cryptography uses mathematical techniques for information security, data integrity, confidentiality, non-repudiation and authentication. Cryptography is based on concepts of Encryption and Decryption [2]. When data is sent from sender to receiver, the data is converted to some unreadable form called encryption of data and at receiver side data is again converted to its original form called decryption of data. Both encryption and decryption process require the key. For protection of valuable information from unlawful imitation, eavesdropper's attack and modification, different types of cryptographic algorithms are designed. There are two major types of such algorithms: symmetric cryptography [3] and asymmetric cryptography [4]. In asymmetric key cryptography two different keys are used, one for encryption called public key and one for decryption called private key. Only one same key is used in symmetric scheme.

The applications of both schemes differ due to efficiency of scheme; symmetric scheme is mostly used for encryption of data due to its high performance while asymmetric is often used for digital signature and distribution of key. Moreover, no any symmetrical ciphering technique such as AES, DES, Advanced AES, and IDEA has taken any benefit from most recent advances in information processing technology. Various kinds of modern data encryption techniques [2], [5] are found in the literature. Genetic Algorithms (GAs) [6] are among such techniques.

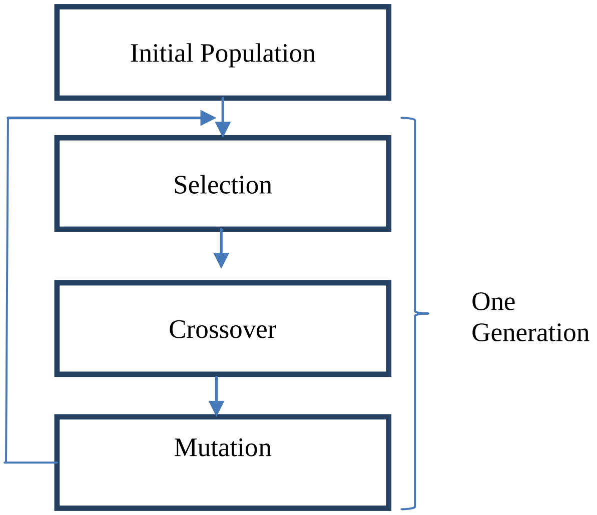


Fig. 1. Flow chart of genetic algorithm.

GA is kind of adaptive search algorithms which make use of the mechanics of natural selection and genetics. GA is part of Evolutionary Algorithms; which are used to solve optimization problems with the help of biological mechanism like selection, crossover and mutation [7]. Fig. 1 shows the process of solving optimization problems using Genetic Algorithms.

The key idea of GA is to imitate the randomness of the nature where natural selection process and behavior of natural system make population of individuals able to adapt the surrounding. We can say the survival and reproduction of the individuals is supported by exclusion of less fitted individuals. The population is generated in such a way that the individual with the highest fitness value is most likely to be replicated and

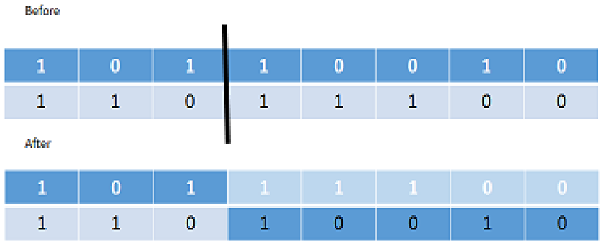
This work has been sponsored by the Higher Education Commission of Pakistan through Indigenous Ph D fellowships program.

unfitted individual is discarded based on threshold set by an iterative application of set of stochastic genetic operators [8].

Genetic Algorithm performs following operations to transform the population to new population based on fitness value.

A. Crossover

Crossover is a genetic operator which joins two chromosomes to form a new chromosome. The newly generated child chromosome is composed of chromosomes from each parent.



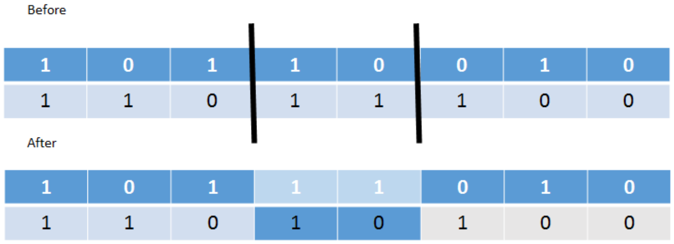
1	0	1	1	0	0	1	0
1	1	0	1	1	1	0	0

1	0	1	1	1	1	0	0
1	1	0	1	0	0	1	0

Fig. 2. Single point crossover.

Crossover is classified as single point, two point and uniform crossover. In Single Point only one crossover point is selected to generate new child (Fig. 2).

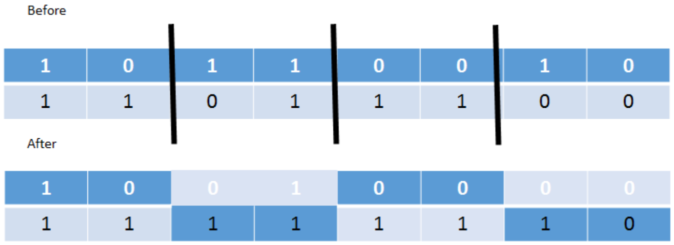
In Two Point crossover two crossover points are selected to generate new child (Fig. 3). In Uniform crossover bits are selected uniformly from each (Fig. 4) [8].



1	0	1	1	0	0	1	0
1	1	0	1	1	1	0	0

1	0	1	1	1	0	1	0
1	1	0	1	0	1	0	0

Fig. 3. Two point crossover.



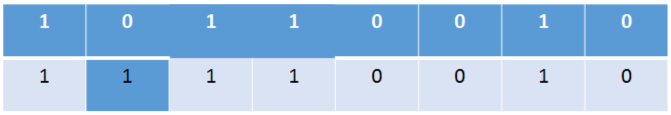
1	0	1	1	0	0	1	0
1	1	0	1	1	1	0	0

1	0	0	1	0	0	0	0
1	1	1	1	1	1	1	0

Fig. 4. Uniform crossover.

B. Mutation

In mutation after crossover at least one bit in each chromosome is changed (Fig. 5) [9]. This is performed to reflect the effect of surrounding in natural genetic process. There are two major types of Mutation i-e Flipping of Bits and Boundary Mutation. In Flipping of Bits one or more bits are converted into 0 to 1 or 1 to 0. In Boundary Mutation randomly upper or lower block in swapped in chromosome [9].



1	0	1	1	0	0	1	0
1	1	1	1	0	0	1	0

Fig. 5. Mutation.

C. Selection

In selection, chromosomes are chosen from the population for generation of new population. The selection is based on fitness value, higher the value more is the chances to be selected. Selection is classified as Roulette-wheel Selection, Tournament Selection; Truncation Selection [8].

D. Fitness Function

This is very important function of Genetic Algorithm because good fitness functions are useful for exploring the search space efficiently and bad fitness functions are confined to local optimum solution. Fitness Function can be categorized as Constant fitness function and Mutable fitness function [9].

Key Selection in cryptography is kind of selection problem and when we consider selection then; the key with highest fitness and randomness is selected. The applications of Genetic Algorithm are also in search heuristic problems, which make the GA a reliable algorithm for key generation and data encryption.

The opinion, which, we are following in this paper, is that if the quality (randomness) of the pseudorandom numbers generated for keys is good then the keys generated will always be non-repeating and purely random and ultimately increase the security and strength of keys.

Our major research question for this research is how can we get benefit of computational intelligence especially the genetic Algorithm to optimize the Cryptosystems? If so what will be the performance of such kind of solutions?

II. Literature Review

With the help of GA most of the research has been done by different researchers in the area of data encryption and key generation. Some of the work is defined in this section.

Jhingran et al. [7] conducted survey on applications of genetic algorithm in the field of cryptography.

Hassan et al. [10] have used the concept of encryption and decryption with the help of GA and RSA. First the key was generated with the help of GA and then generated key was used in RSA to encrypt the data. In this way the strong key was generated that was non-repeating too and this was not easy to break. This algorithm is better in terms of key strength than DES, AES, and RSA, etc. Sindhuja et al. [11] has given a symmetric key cryptosystem by applying GA. Key matrix and text matrix were added to create an additive matrix and then substitution cipher was applied on additive matrix to create the intermediate cipher. Crossover and Mutation were then applied on intermediate cipher to encrypt the data. This method is simple and easy to implement.

Aarti Soni et al. [12] proposed a new algorithm in which pseudorandom number generator was used to generate the key. The random number generator used the current time of computer for random numbers. Then genetic operations were performed on random numbers. Finally selected key was used in AES symmetric algorithm to encrypt the image. The benefits of this algorithm were increased efficiency, less computational time and irregularity of key. The same method of key generation was also followed by Sania Jawed et al. [13] but in this, fitness value was calculated by applying Frequency and Gap test along with hams distance between the two binary keys. This algorithm was implemented in Java technology where 100 chromosomes, 0.5 mutation rate, 2.5 crossover rate were selected for the algorithm.

Narendra K. Pareek et al. [14] used the GA for encryption of gray scale images. The performance analysis of scheme revealed that the algorithm possesses the good statistical results, key sensitivity and can handle the plaintext attack, brute force attack, entropy attack and differential attack. Kirshna et al. [15] proposed cryptographic algorithm by using genetic function. In this algorithm substitution matrix and double point crossover was used to encrypt the data. This algorithm was implemented in Xilinx 13.2 version and verified using Spartan 3e kit. Almarimi et al. [1] dealt with security of electronic data over network. The proposed algorithm integrated the GA and pseudorandom sequence for encryption and decryption of data. Random sequence was obtained by using nonlinear shift register. Time and speed of algorithm was calculated for observing results.

Swati Mishra et al. [8] worked to generate a best fit key which could make code difficult to crack. Fitness of key was calculated by Pearson coefficient of autocorrelation. Two keys public and private were generated by using random number generator, crossover and then mutation. Finally Gap and Frequency tests were applied to select the best sample of key. The process was repeated until there was no best key. C++ programming was used to implement the algorithm and frequency was tested by chi-square test.

Ankit et al [9] generated the key for stream cipher with the help of natural selection process. The genetic operations were repeated until fitness value of any chromosome is less than threshold. Once completed the final selection of key was done through GA. Selected key was unique and non-repeating.

Kalaiselvi et al [16] discussed the need of adaptive and dynamic cryptographic algorithm to reduce computational cost and enhance security. In this paper two enhanced AES cryptosystems were proposed by using GA in SP boxes. AES was modified to accommodate the nonlinear Neural Network in SP network. This scheme ensured the increased security against timing attacks and reduction of computational time.

Subhajit et al [17] encrypted an image by using genetic algorithm. Then statistical test were performed to visualize the feasibility of solution.

The work done by researchers has impressive results but each research work has used some existing cryptographic algorithm in combination with genetic operators. Our motivation is to create novel cryptographic algorithm with the help of Genetic operations, which is easy to implement and secure in terms of key strength and attack time.

III. Proposed Algorithm

The proposed algorithm is named as Genetic Crypto and is divided into three major steps, i.e. Key Generation, Data diffusion and Data Encryption (Fig. 6).

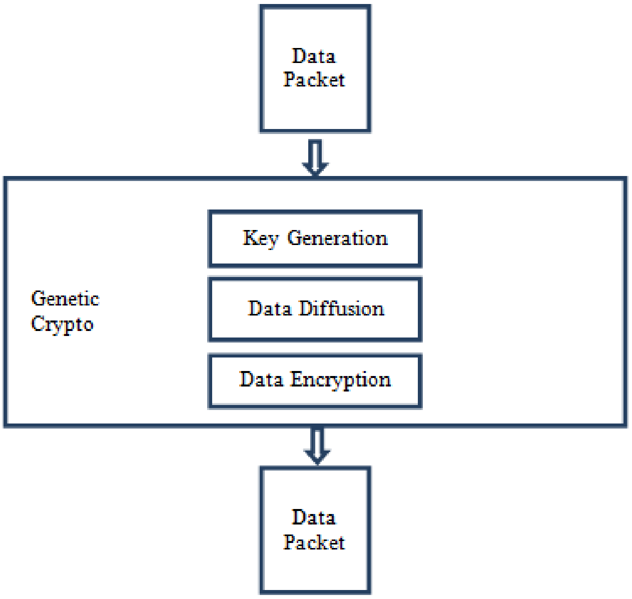


Fig. 6. Genetic crypto flow diagram.

The genetic operators are used in both key generation and data diffusion. Initial population is generated through random number generator. For simplicity one point crossover and bit filliping techniques are used for Crossover and Mutation respectively. Fitness value of key is calculated through Shannon Entropy because entropy is one of important feature of randomness. This algorithm is implemented in C# programing language, .net framework 4.5 in Visual Studio 2012. The interface and example result is shown in Fig. 7.

E. Key Generation: Key will be of 80-128 bits.

1) Sixteen random characters are generated with the help of random number generator from A-Z.

2) Each randomly generated character is converted to binary format (8 bits).

3) The result is stored in 2D array data structure.

4) Sixteen prime random numbers are generated from 0-100.

5) Each randomly generated number is converted to binary format (8 bits).

6) The result is stored in 2D array data structure.

7) Eight random numbers from 1 to 7 are generated for crossover points.

8) The numbers are stored in array data structure.

9) One point crossover is performed by taking one parent from array of random prime number and one parent from array of random characters. The crossover point is identified from the array of random numbers generated in step 1.8.

10) Step 1.9 will be repeated until there is parent left for crossover.

11) For Mutation, bit flipping mutation is used in which first and last bit of each chromosome is inverted; means 0 will be converted to 1 and vice versa.

12) Step 1.11 will be repeated for all the child chromosomes.

13) After Mutation, Fitness function of each chromosome is calculated through Shannon Entropy.

14) Chromosomes with the Shannon Entropy of greater than 0.95 will be merged and selected as key. If there is no any.

15) Chromosome with entropy greater than 0.95 then the whole process will be repeated again until there is no best fit key.

F. *Diffusion of Original Text*

1) Data is converted to binary format.

2) Binary data will be segmented into blocks. Each block size is 8 bits and number of blocks (chromosomes) is size of data/8.

3) The result is stored in 2D array data structure.

4) Eight random numbers from 1 to 7 are generated for crossover points.

5) The numbers are stored in array data structure.

6) One point crossover is performed between adjacent parents in array of binary data. The crossover point is identified from the array of random numbers generated in step 2.5.

7) For Mutation, bit flipping mutation is used in which first and last bit of each chromosome is inverted.

G. *Encryption*

1) Length of key and length of data is calculated first. If any of them has fewer bits than the other, 0s will be appended from left to make the length of data and key equal.

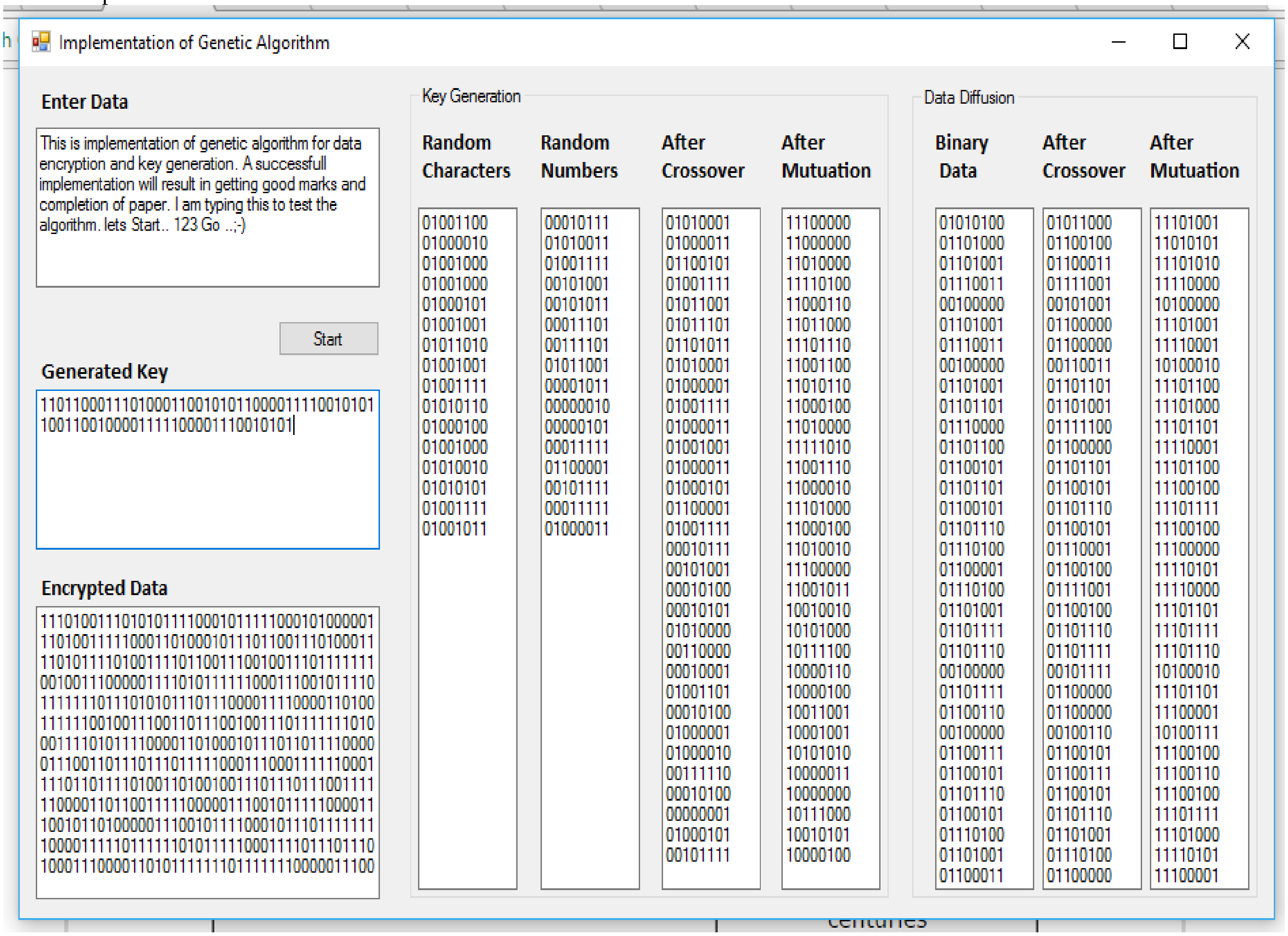
2) Logical XOR operation will be performed between diffused data and key bit wise.

3) The resulting set of bits is encrypted data

Some of the limitations of our work are:

a) Randomness purely depends on the random number generator and it may be pseudo random number generation. It is just limited to 16 characters.

b) Length of key and data is subject to design consideration.



Enter Data

This is implementation of genetic algorithm for data encryption and key generation. A successfull implementation will result in getting good marks and completion of paper. I am typing this to test the algorithm. lets Start.. 123 Go ..;-)

Generated Key

```
11011000111010001100101011000011110010101
1001100100001111100001110010101
```

Encrypted Data

```
11101001110101011110001011111000101000001
11010011111000110100010111011001110100011
11010111101001111011001110010011101111111
00100111000001111010111111000111001011110
11111110111010101110111000011110000110100
11111100100111001101110010011101111111010
00111101011110000110100010111011011110000
01110011011101110111110001110001111110001
11101101111010011010010011101110111001111
11000011011001111110000011100101111000011
10010110100000111001011110001011101111111
10000111110111111010111110001111011101110
10001110000110101111111011111110000011100
```

Key Generation

```
Random      Random      After       After
Characters  Numbers     Crossover   Mutuation

01001100    00010111    01010001    11100000
01000010    01010011    01000011    11000000
01001000    01001111    01100101    11010000
01001000    00101001    01001111    11110100
01000101    00101011    01011001    11000110
01001001    00011101    01011101    11011000
01011010    00111101    01101011    11101110
01001001    01011001    01010001    11001100
01001111    00001011    01000001    11010110
01010110    00000010    01001111    11000100
01000100    00000101    01000011    11010000
01001000    00011111    01001001    11111010
01010010    01100001    01000011    11001110
01010101    00101111    01000101    11000010
01001111    00011111    01100001    11101000
01001011    01000011    01001111    11000100
                        00010111    11010010
                        00101001    11100000
                        00010100    11001011
                        00010101    10010010
                        01010000    10101000
                        00110000    10111100
                        00010001    10000110
                        01001101    10000100
                        00010100    10011001
                        01000001    10001001
                        01000010    10101010
                        00111110    10000011
                        00010100    10000000
                        00000001    10111000
                        01000101    10010101
                        00101111    10000100
```

Data Diffusion

```
Binary      After       After
 Data       Crossover   Mutuation

01010100    01011000    11101001
01101000    01100100    11010101
01101001    01100011    11101010
01110011    01111001    11110000
00100000    00101001    10100000
01101001    01100000    11101001
01110011    01100000    11110001
00100000    00110011    10100010
01101001    01101101    11101100
01101101    01101001    11101000
01110000    01111100    11101101
01101100    01100000    11110001
01100101    01101101    11101100
01101101    01100101    11100100
01100101    01101110    11101111
01101110    01100101    11100100
01110100    01110001    11100000
01100001    01100100    11110101
01110100    01111001    11110000
01101001    01100100    11101101
01101111    01101110    11101111
01101110    01101111    11101110
00100000    00101111    10100010
01101111    01100000    11101101
01100110    01100000    11100001
00100000    00100110    10100111
01100111    01100101    11100100
01100101    01100111    11100110
01101110    01100101    11100100
01100101    01101110    11101111
01110100    01101001    11101000
01101001    01110100    11110101
01100011    01100000    11100001
```

Fig. 7. Results of implementation of proposed algorithm.

IV. Results and Discussions

The proposed algorithm (Genetic Cipher) is compared with DES and AES symmetric key cryptosystems in terms of encryption, decryption time and key strength. The key strength is categorized by key search space size means how many alternative keys can be tried to break the cipher, Attack Scenario means how much time is required by eavesdropper to attack on data. The Encryption and decryption are calculated by implementing the algorithm and key strength is in terms of attack time is calculated with help of GRC [1] Interactive Brute Force key "Search Space" Calculator.

TABLE I. Comparison with Other Algorithms

	DES	AES	Genetic Cipher
Encryption Time	068907 mm	084440 mm	27069 mm
Key Search Space Size	$4.85 * 10^{28}$ Keys	$2.31 * 10^{57}$ Keys	$1.11 * 10^{120}$ Keys
Attack Time (1000 k/s)	15.41 thousand trillion days	7.34 hundred million trillion days	3.53 hundred billion trillion days

Table I shows that Encryption time of DES and AES is 068907mm and 084440 mm respectively while Encryption time of Genetic Cipher is 27069 mm, which is higher than both. The complex cryptographic algorithms with high provision of security are much better than simple algorithm with less security in cryptography. This point is evidenced by measure of key strength. In both categories key search space and attack time the Genetic Cipher requires much higher time to break than DES and AES.

To see performance improvement we consider the encryption time, size of the key search space and attack time. In Fig. 8, we plot the time taken by our algorithm and compare with the time taken by DES and AES. There is a significant improvement as encryption time is lesser, Search space is vast and Attack Time is much higher than AES and DES. In this graph we took log of the Search Space and Attach time in order to improve visibility of the plot.

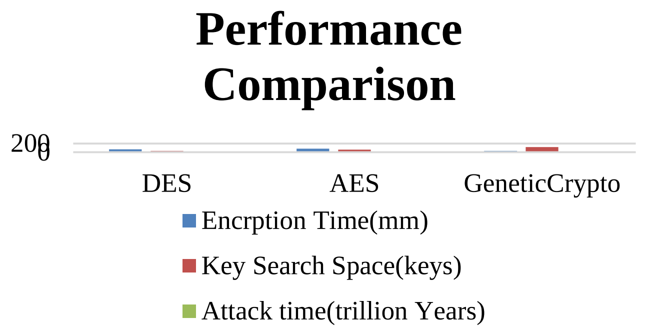


Fig. 8. Performance with respect to DES and AES.

V. Conclusions and Future Work

In this paper we have adopted a new way to encrypt the data i-e using GA. First a key of length between 80 and 128 is generated by applying genetic operations on randomly generated characters and prime numbers. Shannon Entropy is used to calculate the fitness value of each chromosome. After key generation, data is diffused again by applying crossover and mutation on data. At last key and diffused data are XORed for encryption. The result shows that although the proposed algorithm take little longer encryption time than DES and AES but the key strength is better than the other two compared algorithms.

In future we will prepare to improve this algorithm for multimedia encryption like images, video and audio. Efficiency in terms of time will be considered first. From the evaluation point of view, we will compare this genetic cipher with other cryptographic algorithms. Also, we can use more statistical techniques for evaluation of key randomness.

References

[1] A. Almarimi, A. Kumar, I. Almerhag, and N. Elzoghbi, "A NEW APPROACH FOR DATA ENCRYPTION USING GENETIC Original Image Pseudorandom Binary Sequence Generator using GA and Decryption Decrypted Image," Computer (Long. Beach. Calif)., pp. 2–6, 2014.

[2] D. R. Stinson, Cryptography: Theory and Practice, vol. 30. 2005.

[3] J. Daemen and V. Rijmen, The Design of Rijndael: AES - The Advanced Encryption Standard. 2002.

[4] R. L. Rivest, A. Shamir, and L. Adleman, "A method for obtaining digital signatures and public-key cryptosystems," Commun. ACM, vol. 21, no. 2, pp. 120–126, 1978.

[5] W. M. H. Company, Modern Cryptography: Theory and Practice, vol. 170, no. 2. 2003.

[6] D. E. Goldberg, Genetic Algorithms in Search, Optimization, and Machine Learning. 1989.

[7] R. Jhingran and A. Prof, "A Study on Cryptography using Genetic Algorithm Vikas Thada Shivali Dhaka," Int. J. Comput. Appl., vol. 118, no. 20, pp. 975–8887, 2015.

[8] S. Mishra and S. Bali, "Public key cryptography using genetic algorithm."

[9] A. Kumar and K. Chatterjee, "An efficient stream cipher using Genetic Algorithm," 2016 Int. Conf. Wirel. Commun. Signal Process. Netw., pp. 2322–2326, 2016.

[10] A.-K. S. O. Hassan, A. F. Shalash, and N. F. Saudy, "MODIFICATIONS ON RSA CRYPTOSYSTEM USING GENETIC OPTIMIZATION," Int. J. Res. Rev. Appl. Sci., vol. 19, no. 2, p. 150, 2014.

[11] S. K and P. D. S, "A Symmetric Key Encryption Technique Using Genetic Algorithm." .

[12] A. Soni and S. Agrawal, "Using Genetic Algorithm for Symmetric key Generation in Image Encryption," Int. J. Adv. Res. Comput. Eng. Technol., vol. 1, no. 10, pp. 2278–1323, 2012.

[13] S. Jawaid and A. Jamal, "Article: Generating the Best Fit Key in Cryptography using Genetic Algorithm," Int. J. Comput. Appl., vol. 98, no. 20, pp. 33–39, Jul. 2014.

[14] N. K. Pareek and V. Patidar, "Medical image protection using genetic algorithm operations," Soft Comput., vol. 20, no. 2, pp. 763–772, 2014.

[15] G. M. K. and V. Lakshmi, "A Proposed Method for Cryptographic Technique by Using Genetic Function," Int. J. Emerg. Eng. Res. Technol., pp. 1–7, 2015.

[16] K. Kalaiselvi and A. Kumar, "Enhanced AES cryptosystem by using genetic algorithm and neural network in S-box," in 2016 IEEE International Conference on Current Trends in Advanced Computing, ICCTAC 2016, 2016.

[17] S. Das, S. N. Mandal, and N. Ghoshal, "Diffusion and Encryption of Digital Image Using Genetic Algorithm," in Proceedings of the 3rd International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA) 2014, 2015, pp. 729–736.

[1] https://www.grc.com/haystack.htm

Airline Sentiment Visualization, Consumer Loyalty Measurement and Prediction using Twitter Data

Rida Khan, Siddhaling Urolagin
Department of Computer Science
BITS Pilani, Dubai Campus, Academic City
Dubai, United Arab Emirates

***Abstract*—Social media today is an integral part of people's daily routines and the livelihood of some. As a result, it is abundant in user opinions. The analysis of brand specific opinions can inform companies on the level of satisfaction within consumers. This research focus is on analysis of tweets related to airlines based in four regions: Europe, India, Australia and America for consumer loyalty prediction. Sentiment Analysis is carried out using TextBlob analyzer. The tweets are used to calculate and graphically represent the positive, negative mean sentiment scores and a varying mean sentiment score over time for each airline. The terms with complaints and compliments are depicted using visualization methods. A novel method is proposed to measure consumer loyalty using the data gathered from Twitter. Furthermore, consumer loyalty prediction is performed using Twitter data. Three classifiers are employed, namely, Random Forest, Decision Tree and Logistic Regression. A maximum classification accuracy of 99.05% is observed for Random Forest on 10-fold cross validation.**

Keywords*—*Consumer loyalty measurement; consumer loyalty prediction; sentimental visualization; airline consumer analysis

I. Introduction

People are increasingly using social media platforms as a place to review products and share product or service specific opinions [1]. As a result, businesses have become aware of the importance of social media as part of their marketing strategies. It can be used in communicating with consumers on a one-to-one basis and receiving immediate feedback [2]. The data from social media can be used to firstly understand the standing of a company within consumers and in the industry with respect to social media sentiment scores [3]. Secondly, the data can be used to infer consumer feedback, which is essential information for companies. The feedback informs the general public opinion about a company's product or service. People are constantly posting about brands they are using, their satisfaction with products and sharing these opinions with friends or family [4]. Hence, it is important for companies to know what the consumers are saying and whether it is positive or negative. Traditional forms of data collections could be replaced with analysis of social media platforms [5]. Data from surveys is commonly used to compute consumer loyalty measurements. We propose to use sentiment information to deduce loyalty measurement of a consumer towards a brand. The loyalty measurement will help to understand impact of consumer reviews on a brand [6].

Retaining customers and maintaining consumer loyalty is an essential marketing strategy, as customers are becoming increasingly important [7]. A loyal customer is an asset and companies today use a good amount of their marketing tools to maintain it. The currently available consumer loyalty measures use surveys and questionnaires but social media datasets are not used [8]. Since, social media platforms are full of user opinions pertaining to whether they are currently loyal or not and the reasons behind their position, social media datasets should be made an essential part of consumer loyalty measurement calculations [9], [10]. Discerning whether a consumer is loyal includes answering a set of questions, which could be in the form of a survey or questionnaire: How loyal is your customer? How likely is said customer to refer your brands to his friends or family? Is he likely to continue purchasing your products or services? Or is he looking for other potential options? Is he listening to pitches from your competitors? Is he willing to give you feedback and give you time to fix errors? Each of these questions is equally important in determining the loyalty of a certain consumer [11]. Along with these questions, there are many measures available that are used worldwide to calculate consumer loyalty measurement. These would include the Net Promoter Score (NPS) [12], Customer Loyalty Index (CLI), Upselling Ratio (UR) and Repurchase Ratio (RR) [13]. Companies use some or all of these measures in order to understand the current standing of consumers and to work towards retaining any or all of their consumers.

Our focus is on analyzing 'Tweets', which are Twitter [14] user posts of 280 characters or less. Along with the text of the tweet, other aspects are also collected like the username and number of people who have liked the tweet etcetera [15], [16]. Some categories of brands that are active and popular on Twitter include Airlines, Cars, Sports Teams, Entertainment, Finance, Retail and Food Industry among others [17]. This research work is focused on analysis of tweets pertaining to airlines. Our data consisting of tweets is collected using Tweepy [18]. We perform sentiment analysis [19] on tweets using TextBlob [20]. We infer the most common issues customers have with each airline and analyze the varying sentiments of the tweets using various visualization techniques [21]. We developed a new method to measure customer loyalty based on tweets. Furthermore, we carried out customer loyalty prediction based on sentiment data. In Section 2, we have provided the details about the dataset, in Section 3 we discussed the sentiment analysis of airline tweets. The experimental and graphical results are shown in Section 4 and Section 5 covers the conclusion.

II. Datasets

We have collected 'Tweets' for 18 airlines based in four selected regions, which are America, India, Europe and Australia. The specific airlines and tweet count are shown in Table I. Table II shows the sample tweets as an example.

TABLE I. Data Collected

Airlines			
American	***Count***	***European***	***Count***
Delta	6000	**Lufthansa**	507
United	1600	**Air Berlin**	5822
Spirit	1600	**Turkish Air**	4637
Southwest	2600	**KLM**	2404
Jet Blue	3000	**Easy Jet**	1465
Indian	***Count***	***Australian***	***Count***
Indigo	1600	**Virgin**	3800
Air India	1080	**Qantas**	1600
Jet Airways	1100	**Tiger air**	5000
Vistara	1100	**Jet Star**	9994
Spice Jet	2392	**Sharp**	630

TABLE II. Sample Dataset

User name	***Date***	***Tweets***	***Likes***	***Retweets***
sashane_gordon,	2018-04-10 11:33,	Am sorry but I am really disappointed @JetBlue . I was a loyal customer. The way I advertise this airline you would think I work for the place but you guys showed me today that you don't care. Whether my family is dying or not as long as you have the money. Things happen everyday,	0	0
KW_Consultants,	2018-04-09 23:58,	@FlyAirNZ is the best airline I have ever flown! I am a loyal @AmericanAir customer , but the #skycouch stole my heart! https:// twitter.com/flyairnz/status/983283384951754752	1	0
pberrini,	2018-02-28 02:18,	@WorldMark (Michael D Brown) I'm a corporate retiree and worked for organizations where I was empowered to make a difference for the customer as long as I followed 'core values'. I am a LOYAL Southwest Airline s customer for the same reason. I would love to feel same 4 Wyndham.	0	0
sevenpitches,	2018-02-25 04:04,	This is one of the many reasons I am a loyal Delta customer . Proud they are my hometown airline . Thanks, @Delta https:// twitter.com/Delta/status/967391778897891328	3	0
ScottBurges,	2017-12-02 16:54,	Thought you were commenting on the on board staff ! Now they have become a budget airline , there are so many better quality and cheaper alternatives out there, and I was a loyal customer for over 20 years.	1	0

Each airline has a help desk handle; we have collected tweets that users have directed to these particular handles to create Dataset 1. In the case that an airline does not have such an account, we have used their official Twitter handle instead. Tweet collection has been carried out using Tweepy [18], which is a python library used for accessing Twitter API and is widely accessible to all Twitter users. Moreover, we formed Dataset 2 for consumer loyalty analysis by collecting tweets using the search queries "loyal flyer" and "loyal to airline" as well as "left airline" and the response of 10,000 tweets are gathered in CSV form. These 10,000 tweets are collected from 1048 users, out of which 524 have used the term "loyal to airline" or "loyal flyer" and 524 have used the term "left airline". Each collected tweet consists of the ID, permalink, date and time stamp, the text contained within the tweet, username, retweets and likes.

III. Analysis of Airline Tweets

Twitter is an online platform that connects people through a social networking environment. Each user can create an account with a unique handle and post "Tweets" [22]. The main aspect of Twitter is text; however, pictures and videos are also shared. It should, therefore, be vital for companies to ensure any posts regarding their brand are positive. Millions of people discussing and mentioning a brand is only a good thing if the tweets are positive [23]. Sentiment analysis becomes important for identifying positive or negative tweets and determining the consumer voice. It is the use of natural language processing, as well as analysis of text and computational linguistics to study subjective data. It discerns whether a particular text or piece of writing – tweet, for example – is positive, negative or neutral. It uses context of the piece as well as tone, emotion and vocabulary. Sentiment analysis can aid companies in marketing strategies by understanding a general public opinion and in succession improving their customer service [24]. A company can determine the public opinion, analyze the customer satisfaction towards their products and be open to listening to the issues. This analysis can, not only help companies to know how their customers think, but can also aid them in competitor advantage. If companies are aware of the sentiments of competitors, it can aid them in comparison with their own sentiments and they can plan strategies to improve accordingly [25]. This analysis can also help in retaining customers as consumer loyalty can be determined and predicted which can greatly influence a company's organizational decisions [26].

A. Sentiment Information Visualisation

Our focus is on the opinions that users post on Twitter directed to airline Help Desks or to the official Twitter handles of various airline companies. These tweets range from compliments to complaints and issues that various consumers have with any airline. Analyzing the sentiment of tweets has an extra level of complication because the anatomy of tweets includes more textual aspects than an average written piece [27]. There are images, links, emoticons and other forms of media included. Hence, the first step of our analysis is to clean the tweets we have collected [28]. Tokenization is also difficult due to the body of the text. We would need to make sure that the @-mentions, emoticons, links and #hash-tags are preserved as individual tokens and not ignored, since these are equally important aspects of the analysis [27], [28]. In this research we follow the method as shown in Fig. 1. Airline tweets are first gathered from the Twitter API, which are then cleaned and tokenized. We then perform sentiment analysis on the Tweets, giving each tweet a score using TextBlob. Here, a score of 1 indicates most positive and -1 indicates most negative while zero means a neutral tweet or term. This analysis is carried out using a python library called TextBlob, which is used for processing textual data [20].

From each Tweet, its sentiment score TS_i is computed using TextBlob and the tweet is segregated into positive type or negative type as given in (1).

$$T_i = \begin{cases} \text{Positive if } TS_i > 0 \\ \text{Negative if } TS_i < 0 \\ \text{Neutral if } TS_i = 0 \end{cases} \qquad (1)$$

where TS_i is the sentiment score of i^{th} tweet.

We calculate a mean sentiment score, MS_i for each airline using the sentiment score of the airline tweets.

$$MS_i = \frac{1}{n}\sum_{i=1}^{n} TS_i \qquad (2)$$

where TS_i is the sentiment score of i^{th} tweet and n is number of tweets for the airline.

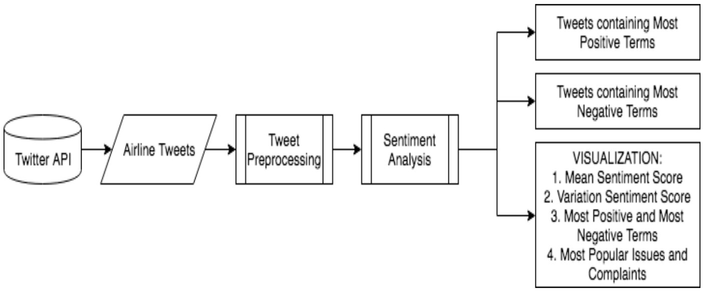

Twitter API
Airline Tweets
Tweet Preprocessing
Sentiment Analysis
Tweets containing Most Positive Terms
Tweets containing Most Negative Terms
VISUALIZATION:
1. Mean Sentiment Score
2. Variation Sentiment Score
3. Most Positive and Most Negative Terms
4. Most Popular Issues and Complaints

Fig. 1. Tweet sentiment analysis.

B. Airline Passenger Loyalty Measurement

Consumer loyalty analysis is carried out using a second dataset, which consists of tweets containing phrases like “loyal flyer”, “left airline” etcetera. The data downloaded with a tweet includes likes, which is the amount of people who have liked a tweet and retweets, which is the number of people who have shared a tweet. This can be used along with the number of followers a user has as well as the sentiment score of their tweets to calculate a loyalty measurement. We collect usernames of passengers who have explicitly stated they are loyal and those who say they are not loyal to an airline. Airline related tweets are gathered for each user. The tweets are subjected to TextBlob to compute the sentiment score, TS_i. The tweets related to a user are segregated to positive type and negative type. From positive tweets, the mean is computed as P^j and from negative tweets; the mean is computed as N^j for the j^{th} person, which is given in (3), (4).

$$P^j = \frac{1}{n}\sum_{i=1}^{n} TS_i \qquad (3)$$

where $TS_i > 0$ and n is number of positive tweets.

$$N^j = \frac{1}{n}\sum_{i=1}^{n} TS_i \qquad (4)$$

where $TS_i < 0$ and n is number of negative tweets.

For the j^{th} person, the mean of his Likes, L^j and Retweets, R^j are also calculated which make the influencer score In^j, and number of followers, F^j is gathered.

$$L^j = \frac{1}{n}\sum_{i=1}^{n} TL_i \qquad (5)$$

where TL_i is number of likes of i^{th} tweet and n is total number of likes for j^{th} person.

$$R^j = \frac{1}{n}\sum_{i=1}^{n} TR_i \qquad (6)$$

where TR_i is number of retweets of i^{th} tweet and n is total number of retweets for j^{th} person.

$$In^j = L^j + R^j \qquad (7)$$

Hence, each user has positive, negative and influence score. Consumer loyalty measurement, CM^j is calculated as given in (8).

$$CM^j = \lfloor P^j + N^j \rfloor \times In^j \times F^j \qquad (8)$$

Fig. 2 shows the method to compute consumer loyalty measurement. The tweets are queried with the search terms “loyal to airline”, “loyal flyer”, “left airline” and searched for usernames. The number of followers for each username is gathered along with the airline related tweets for each user. A sentiment score is computed for the tweets using TextBlob. The tweets are then segregated into positive and negative as given in (1). The positive, negative and influence scores are calculated for each user as given in (3), (4) and (7). The loyalty measurement is computed based on positive, negative, influence and follower scores as given in (8).

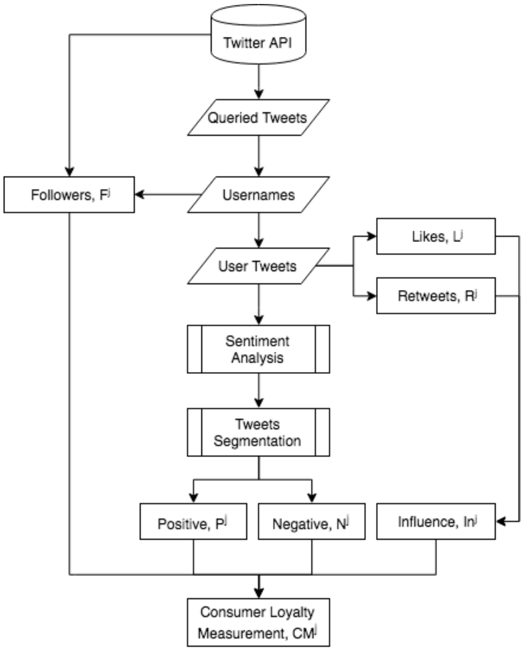


Fig. 2. Consumer loyalty measurement.

IV. Expiremental Results

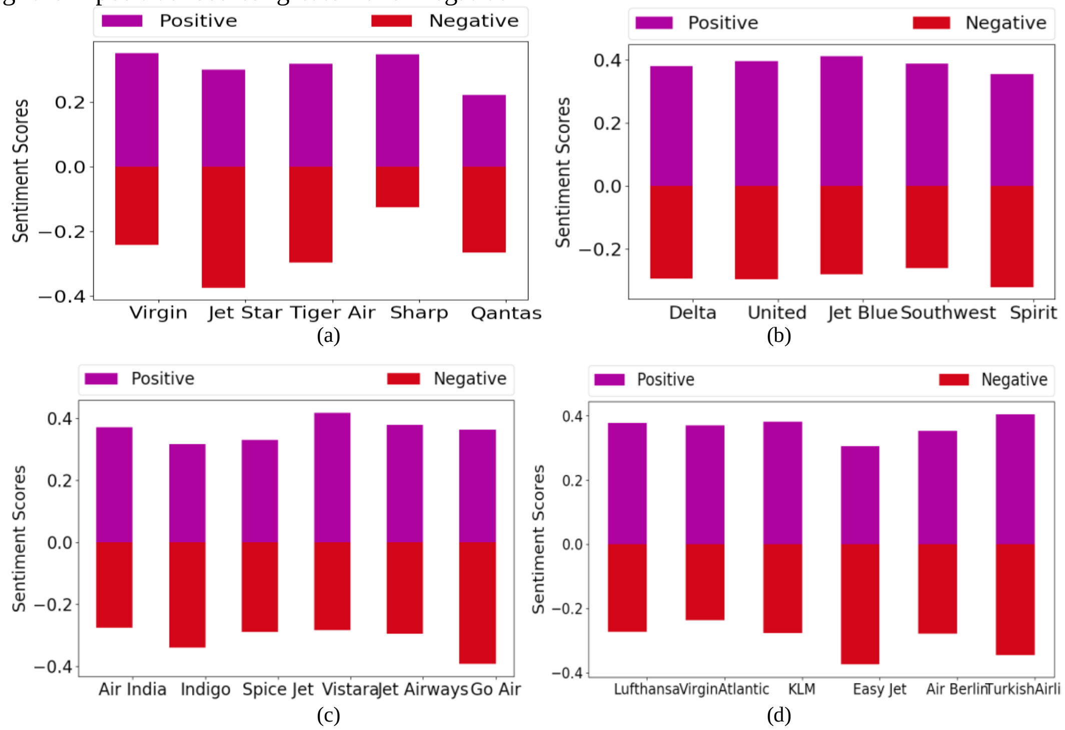
In this research work, we have conducted sentiment analysis, customer loyalty measurement and loyalty prediction on tweets collected from airlines. We have dataset 1 and dataset 2 of tweets from various airlines from four regions, namely India, Europe, America and Australia. Dataset 1 consists of tweets collected from airline handles for region India: 6172 tweets, for region Europe: 14835 tweets, for region America: 13200 tweets and for Australian airlines: 21024 tweets. Searching terms “loyal flyer”, “loyal to airline” and “left airline” forms Dataset 2 and total 10000 tweets are gathered. We use the data to calculate mean sentiment scores for each airline. The airlines can use these depictions to understand areas of improvement, successful strategies and can utilize these insights in retaining customers.

A. *Tweet Sentiment Visualisation*

Sentiment analysis is performed on the tweets from dataset 1 using TextBlob. The tweets are then separated into “Positive tweets” and “Negative tweets”. The mean sentiment score for positive and negative scores is calculated for each airline. We also compute sentiment score over time to depict the variation for selected airlines. From the gathered tweets, we also search for the most frequently occurring positive and negative terms along with the corresponding tweets. Fig. 3(a) graphically represents the positive and negative mean sentiment scores for five Australian airlines. The mean sentiment score is vital for an airline to understand a general consumer opinion about their services at a point in time. We have observed variations in positive and negative sentiments for the various airlines. An airline would want to make sure that their positive sentiment score is greater than their negative sentiment scores. These scores also aid in competitor advantage as an airline can work towards making their positive scores greater and negative scores lesser than their competing airlines. Fig. 3(b), 3(c) and 3(d) show similar results for American, Indian and European airlines, respectively.

The brand sentiment scores over time are important for companies, since they indicate whether customers have been talking about your brand positively and their attitude towards your brand is improving or whether they have been dissatisfied and the score has been reducing. The tweets used to calculate the airline variation score range over a month’s time. The mean sentiment score per week is calculated for each airline. Fig. 4 graphically represents airline variation scores that show the increases or decreases in consumer satisfaction over a time period. These variations can be studied in order to understand the consumer’s satisfaction over a period of time. Fig. 4(a) depicts Jet Airways variation score that starts at 0.15 in Week 1 but decreases down to 0.025 by Week 3.

The most common negative and positive terms for all airlines are counted and a list is made. The negative terms have the highest score of -1.0 and are worst, awful, pathetic, disgusting, terrible and horrible. The positive terms have the highest score of 1.0 and are awesome, excellent, delicious, perfect, superb and wonderful. Fig. 5(a) depicts the most frequently occurring negative terms and their respective frequency. An example of a frequent word with a frequency of 150 is ‘worst’. The tweets with this term can be indicators of areas that need improvement for the respective airline. Fig. 5(b) shows similar results for positive terms. The tweets with positive terms can be indicators of areas that are incurring a positive sentiment and the work in these areas can be maintained in order to keep or improve positive scores and consumer satisfaction. Some of the tweets where these terms occur are shown in Table III with their sentiment score.



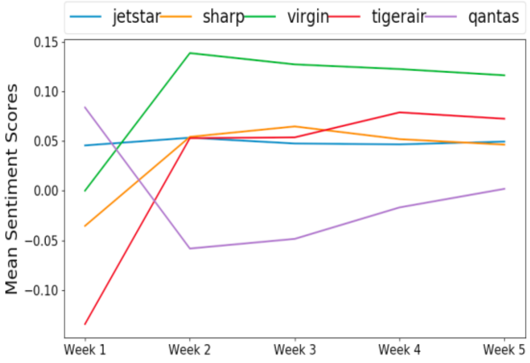
(a) (b) (c) (d)

Fig. 3. Mean sentiment score: (a) Australia (b) America (c) India (d) Europe.

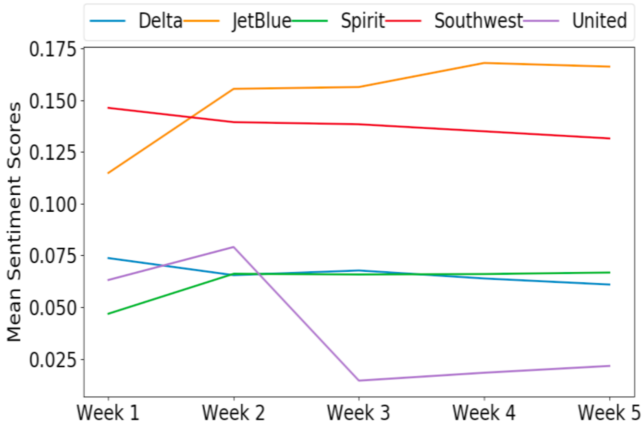
For all airlines, we have represented the most common bigrams in the form of a pie chart in Fig. 6 where each term and its respective frequency are depicted. These bigrams are the two most common terms that occur together most frequently in negative and positive tweets. These terms show the most common problems and the most common praises. Fig. 6(a) shows the most common positive feedback within positive tweets and their respective percentage in terms of frequency within the dataset. The most frequently occurring praise is “customer service” indicating that successful customer service incurs positive sentiment scores. Fig. 6(b) depicts the most common issues found within negative tweets.

TABLE III. TWEETS CORRESPONDING TO POSITIVE AND NEGATIVE TERMS

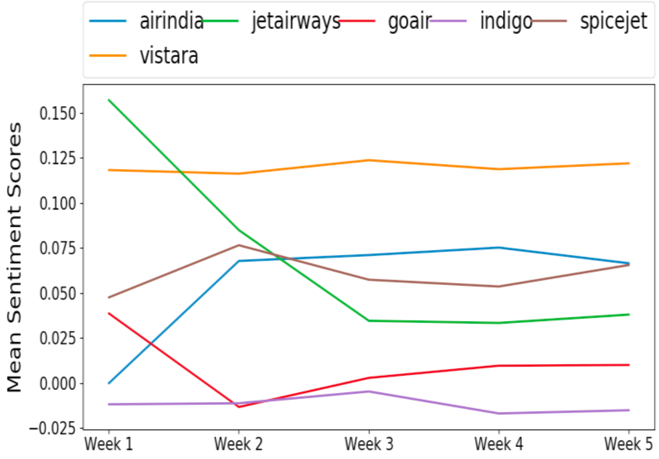
Term	***Tweet***	***Score***
Negative	is the **worst airline** ever My flight been scheduled to leave since 11 am it 3 pm now smh	-1.0
Negative	has no respect for its customers and zero professional pride Terrible service The entire flight is still **waiting for luggage pathetic** You’re losing a customer	-0.475
Negative	And I get to sit in yet another **disgusting seat** No thanks I will never recommend your carrier	-0.55
Negative	My 1st **terrible airline flight** personal items going missing in the plane and cabin crew super unhelpful refused to help find my passport money housekeys shrugging shoulders is not what I expected Conspiring with ground staff disappointed	-0.286115
Positive	Emergency landing in St John s Canada pilot made a **perfect landing** Passenger taken to hospital and is reported to be in good hands bei St John s International Airport YYT	0.56666666
Positive	Such super **delicious food** and awesome hospitality at 38000 ft Just loved it You will surely rule the sky one day Hope to fly frequently with you Thank You	0.55416666
Positive	Flew with for the first time today **superb experience** Great service and hospitality on board Fantastic leg space too Kudos	0.61249999
Positive	Had a **wonderful flight** Never expected such a delicious snack more leg space in domestic flights UK899 See you soon	0.425



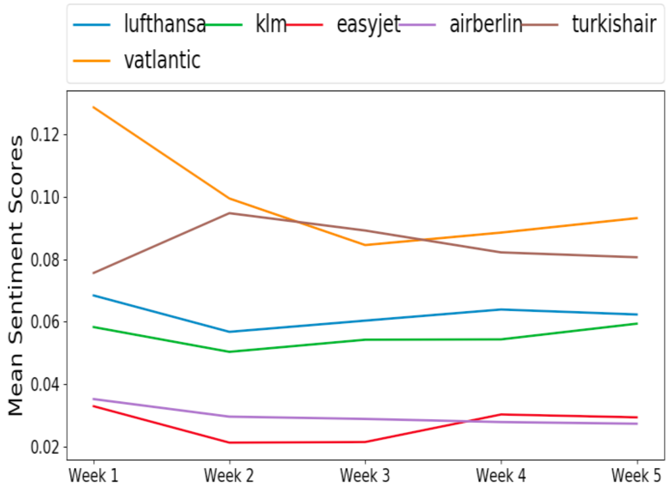
(a)



(b)

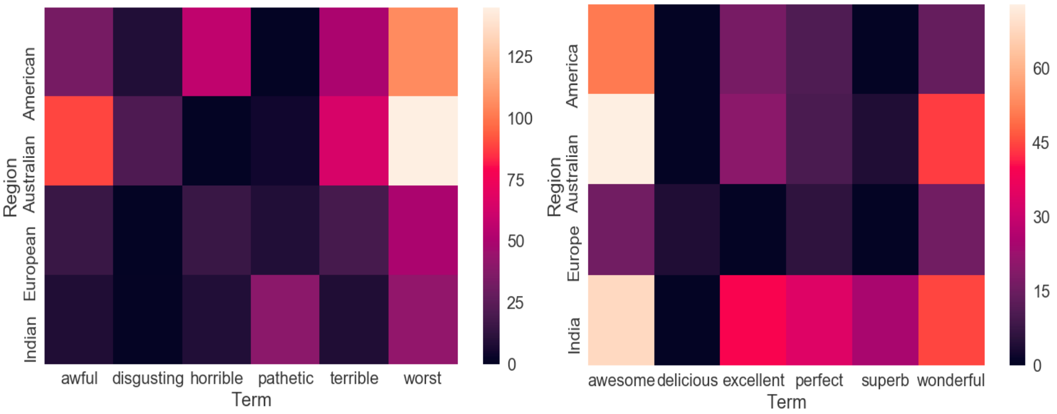


(c)



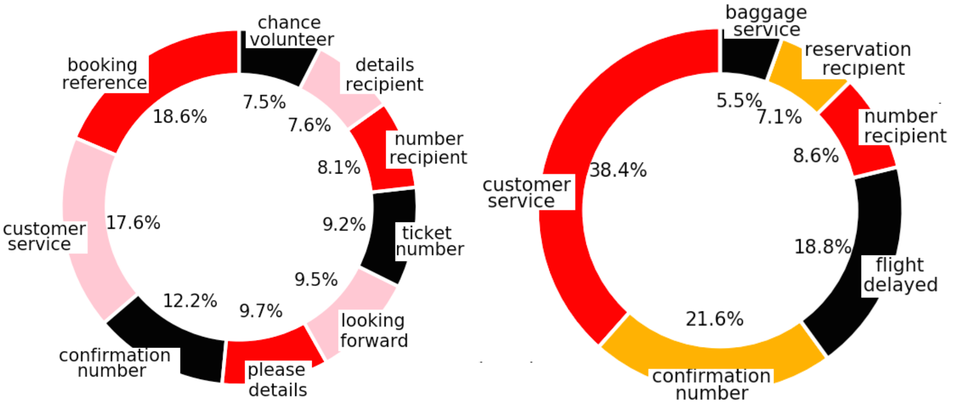
(d)

Fig. 4. Variation sentiment score: (a) Australia, (b) America, (c) Europe, (d) India.



(a) (b)

Fig. 5. Most common terms: (a) Negative terms, (b) Positive terms.



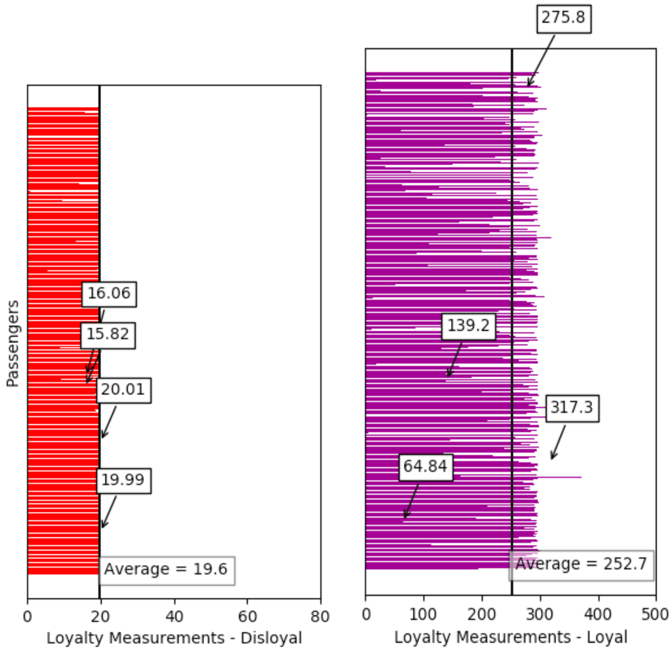
(a) (b)

Fig. 6. (a) Common praises,(b) Common issues.

B. *Airline Passenger Loyalty*

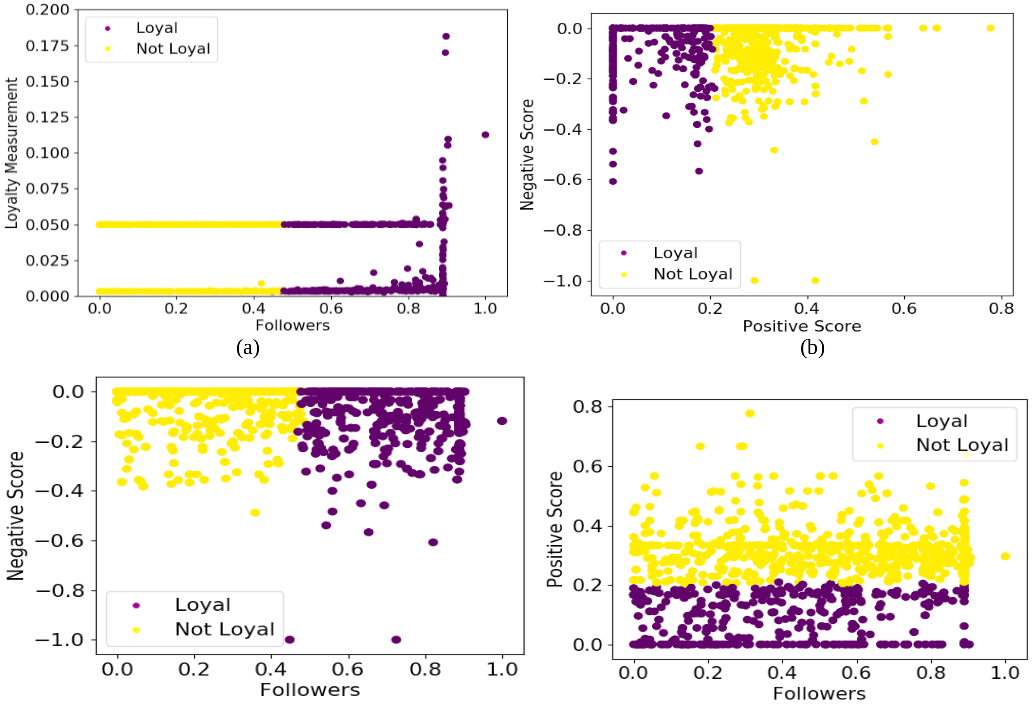
We gathered tweets using search queries such as “loyal flyer”, “loyal to airline” as well as “left airline”. There are 1048 users and 10000 tweets in the dataset 2. From the users, 524 have explicitly said they are loyal to an airline and the other 524 have said they are not loyal or have left an airline. The positive, negative user scores are calculated using (3), (4) along with mean likes and retweets using (5), (6) as described in section 3.2. These values are used to calculate the consumer loyalty measurement as given in section 3.2 using (8). The normalized loyalty measurements are depicted in Fig. 7. The normalization is performed by dividing the difference between maximum and minimum loyalty score.

Fig. 7 represents the consumer loyalty measurements for 524 loyal passengers and 524 disloyal passengers. The passengers who have used the term “left airline” have a loyalty measurement varying between 15 and 21. The passengers who have used the terms “loyal to airline” or “loyal flyer” have a loyalty measurement varying between 250 and 300. These measurements can be used to cluster consumers as loyal or not loyal based on their Twitter data. We used K-Means clustering [29] which is an unsupervised learning algorithm with the number of clusters set to two. The various values used to calculate the loyalty measurement are graphically represented in different combinations using the k-means clusters to depict the loyal and not loyal passengers.



(a) (b)

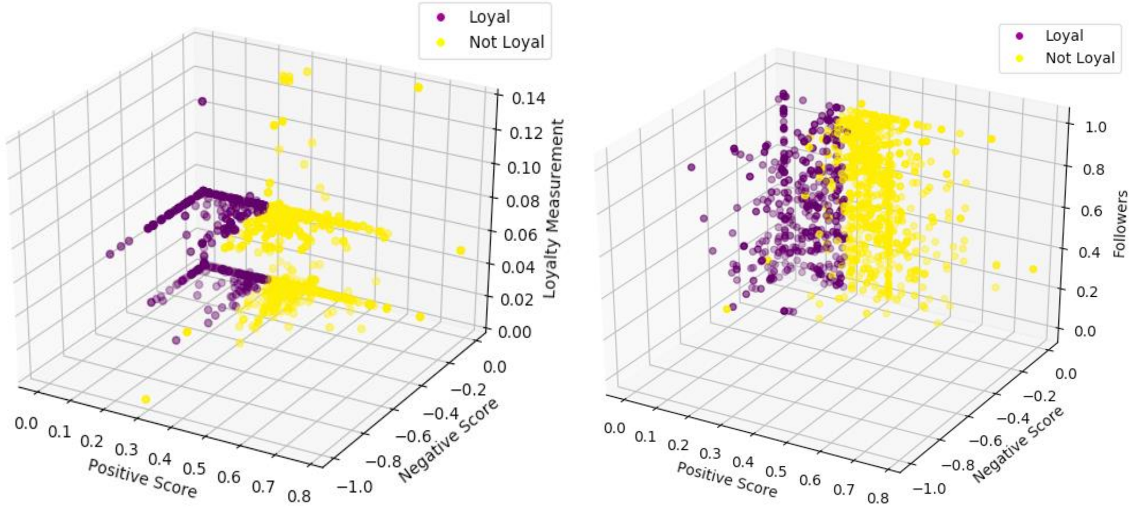
Fig. 7. Airline passenger loyalty measurements: (a) Disloyal passengers, (b) Loyal passengers.



(a) (b)

(c) (d)

Fig. 8. Clustered passengers (a) Loyalty measurement vs followers, (b) Negative score vs positive score, (c) Negative score vs followers, (d) Positive score vs followers.



(a) (b)

Fig. 9. 3D visualizations: (a) Positive score vs negative score vs loyalty measurement, (b) Positive score vs negative score vs followers.

Fig. 8(a) depicts the followers of each passenger in comparison to their normalized loyalty measurement. Fig. 8(a) depicts loyalty measurements along the y-axis and number of followers along the x-axis. This informs an airline about each loyal or disloyal passenger and their influence. Fig. 8(b) depicts the negative score on the y-axis and positive score on the x-axis. Each point represents a user who is either loyal or not loyal. This represents positive versus negative scores with respect to consumer loyalty. Fig. 8(c) and 8(d) represent the negative, positive scores on the y-axis versus the number of followers on the x-axis. Both these figures aid airlines in understanding the influence of users versus their positive and negative scores.

We depict the loyalty measurement and few terms used in calculating this measurement in Fig. 9. Fig. 9(a) depicts positive score on the x-axis, negative score on the y-axis and normalized loyalty measurement on the z-axis. Airlines can understand whether their loyal or disloyal passengers have a positive or negative attitude towards their services at a point in time. Fig. 9(b) depicts positive score on the x-axis, negative score on the y-axis and the number of followers on the z-axis. Both these figures show a clear distinction between the loyal and disloyal passengers. Fig. 9(b) can help inform airlines of the strength of people each loyal and disloyal could influence. The 3D pictorial graphs can be used in vital analysis by airline marketing teams to understand where each passengers stands. Also, these 3D graphs represent the influence of each passenger, negative, positive scores with respect to loyalty.

Next, we carried out consumer loyalty prediction. Previous works that have been carried out for consumer loyalty prediction include surveys (with questionnaires) [30], [31] and airport reports [32] as datasets. Social media is used as a data set but to predict an existing consumer loyalty measure, which is NPS [33]. This recent work can be seen in Table IV.

For consumer loyalty prediction, we used three prediction models, which are Random Forest [34], Decision Trees [35] and Logistic Regression [36] on dataset 2. The model is fitted using tweet related information such as positive sentiment score, negative sentiment score, mean of retweets, mean of likes and the number of followers. Two-class prediction is performed as either loyal or not loyal. The models are tested on 10-fold cross validation [37] and the accuracies are given in Table V. The maximum accuracy of 99.05% is observed for Random Forest.

TABLE IV. RECENT WORK IN CONSUMER LOYALTY PREDICTION

Author	*Year*	*Title*	*Data sets*	*Accuracy*
Mohamed Zaki, Dalia Kandeil, Andy Neely, Janet R. McColl-Kennedy	2016, [30]	The Fallacy of the Net Promoter Score: Customer Loyalty Predictive Model	Surveys	98.00%
Jose Berengueres, Dmitry Efimov	2014, [31]	Airline new customer tier level forecasting for real-time resource allocation of a miles program	Surveys	87.00%
S.T.M. van Velthoven	2014, [33]	Sentiment analysis on social media to predict Net Promoter Score	Social Media	85.67%
Hari Bhaskar Sankaranarayanan, B V Vishwanath, Viral Rathod	2017, [32]	An exploratory analysis for predicting passenger satisfaction at global hub airports using logistic model trees	Airport Reports	80.00%

TABLE V. CONSUMER LOYALTY PREDICTION

Model	**Accuracy**	**Misclassification**
Random Forest	99.05	0.95
Decision Tree	98.97	1.03
Logistic Regression	91.10	8.90

V. CONCLUSION AND FUTURE WORK

In recent years, the tremendous growth of social media is impacting various sectors including businesses. It is vital for any brand today to have a presence on the Internet, one that is memorable for the consumers. In this research, data from social media such as Twitter is gathered for airline industry. We collected airline tweets from four regions namely India, Europe, Australia, America and performed sentiment analysis. We identified the compliments and complaints of customers, variations in sentiment over a period of time and depicted mean sentiments scores using visualization techniques. This analysis provides a general opinion of passengers towards airlines and its variation over time.

Furthermore, we searched tweets with the terms such as “loyal to airline”, “loyal flyer” and “left airline” and collected 10000 tweets. Consumer loyalty analysis is performed on these tweets. We proposed a new method to measure consumer loyalty based on Twitter information such as positive, negative sentiment scores, mean likes, mean retweets and number of followers. Then, consumer loyalty prediction is performed using three classifiers, which are Random Forest, Decision Tree, and Logistic Regression. These classifiers are trained using features collected from Twitter information on 10,000 tweets. All three classifiers are tested using 10-fold cross validation and classification accuracies are collected. A maximum accuracy of 99.05% is observed for Random Forest classifier on 10-fold cross validation. The consumer loyalty analysis helps airline companies to retain consumers and bring in new loyal customers. Moreover, consumer loyalty measure and prediction can be performed for different business sectors.

REFERENCES

[1] M. Lovelin, P. Felciah, R. Anbuselvi, “A study on sentiment analysis of social media reviews”, 2015 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), pp. 1-3, Coimbatore, 2015.

[2] P. Samiei, A. K. Tripathi, “Effect of Social Networks on Online Reviews”, 2014 47th Hawaii International Conference on System Sciences, Waikoloa, HI, 2014, pp. 1444-1453.

[3] F. Zhu and X. Zhang, “Impact of online consumer reviews on sales: The moderating role of product and consumer characteristics”, Journal of Marketing, Vol. 74, no. 2, pp. 133-148, 2010.

[4] Amrita, S.M., Mohan, R., “Application of social media as a marketing promotion tool — A review” in Computational Intelligence and Computing Research (ICCIC), 2016 IEEE International Conference on 15-17 Dec. 2016, Chennai, India

[5] Jamilah, Putu Wuri Handayani, “Analysis on effects of brand community on brand loyalty in the social media: A case study of an online transportation (UBER)”, in International Conference on Advanced Computer Science and Information Systems (ICACSIS), 2016.

[6] Munnukka J., Karjaluoto H., Mahlamäki T., Hokkanen V. “Effects of Social Media on Consumers’ Sports Brand Experiences and Loyalty.” In: Stieler M. (eds) Creating Marketing Magic and Innovative Future Marketing Trends. Developments in Marketing Science: Proceedings of the Academy of Marketing Science. Springer, Cham

[7] J. b. Shao, Z. z. Li and M. y. Hu, “The impact of online reviews on consumers' purchase decisions in online shopping”, 2014 International Conference on Management Science & Engineering 21th Annual Conference Proceedings, Helsinki, 2014, pp. 287-293.

[8] Jansen, Bernard J., M. Zhang, K. Sobel, A. Chowdury, “Twitter power: Tweets as electronic word of mouth”, Journal of the American society for information science and technology 60, no. 11, 2009, Pages 2169-2188.

[9] Baumöl, U., Hollebeek, L. & Jung, R. “Dynamics of Customer Interaction on social media platforms.” Electron Markets (2016) August 2016, Volume 26, Issue 3, pp 199–202

[10] R. Dowling Grahame, Uncles Mark, “Do Customer Loyalty Programs Really Work? Sloan Management Review”, vol. 38, no. 4, pp. 71-82, Sep. 1997

[11] Fan, Yingren, “Research on improving the customer loyalty of retail enterprises, in Artificial Intelligence”, Management Science and

Electronic Commerce (AIMSEC), 2011 2nd International Conference on 8-10 Aug. 2011

[12] Jacob K. Eskildsen , Kai Kristensen, “The accuracy of the Net Promoter Score under different distributional assumptions” in Quality, Reliability, Risk, Maintenance, and Safety Engineering (ICQR2MSE), 2011 International Conference on 17-19 June 2011

[13] Choi Sang Long, Raha Khalafinezhad, “Customer Satisfaction and Loyalty: A Literature Review” in the Perspective of Customer Relationship Management. J. Appl. Bus. Fin. Res. 1 (1): 06-13.

[14] Twitter (2017) Twitter Developer Documentation. [online] Twitter Developer Documentation. Available at: https://dev.twitter.com/rest/public/rate-limiting [Accessed 10th March 2018].

[15] Neethu, M. S., Rajasree, R, “Sentiment analysis in twitter using machine learning techniques”, in Computing, Communications and Networking Technologies (ICCCNT), 2013 Fourth International Conference on 4-6 July 2013, Tiruchengode, India, 2013.

[16] Gupta, B., Negi, M., Vishwakarma, K., Rawat, G., Badhani, P., “Study of Twitter Sentiment Analysis using Machine Learning Algorithms on Python” in International Journal of Computer Applications, Volume 165 – No.9 on May 2017

[17] Kerns C. “Understanding Brands on Twitter.” In: Trendology. Palgrave Macmillan, New York, 2014, Pages 39-85.

[18] Tweepy (2017) Streaming With Tweepy — tweepy 3.5.0 documentation. [online] Tweepy.readthedocs.io. Available at: http://tweepy.readthedocs.io/en/v3.5.0/streaming_how_to.html [Accessed 10th March 2018].

[19] Mäntylä, M. V., Graziotin, D., Kuutila, M., “The Evolution of Sentiment Analysis in Computer Science Review”, Volume 27, February 2018, Pages 16-32

[20] Loria, S., Keen, P., Honnibal, M., Yankovsky, R.; Karesh, D., Dempsey, E. Textblob: Simplified Text Processing. Available online: https://textblob.readthedocs.org/en/dev/ (accessed on 10th Match 2018).

[21] Bird, S., Klein, E., Loper, E. “Natural Language Processing with Python”, O'Reilly Media, 2009.

[22] Sarlan, A., Nadam, C. and S. Basri. “Twitter sentiment analysis” in Information Technology and Multimedia (ICIMU), International Conference on 18-20 Nov. 2014, Putrajaya, Malaysia, 2014.

[23] Jo Mackiewicz, Dave Yeats, Thomas Thornton, “The Impact of Review Environment on Review Credibilit”, Professional Communication IEEE Transactions on, vol. 59, pp. 71-88, 2016, ISSN 0361-1434.

[24] Chen LS., Yang TY.K. “Increasing Customer Loyalty in Internet Marketing.” In: Pan JS., Snasel V., Corchado E., Abraham A., Wang SL. (eds) Intelligent Data analysis and its Applications, Volume II. Advances in Intelligent Systems and Computing, vol 298. Springer, Cham.

[25] Isah, H., Trundle, P., Neagu, D. “Social media analysis for product safety using text mining and sentiment analysis” in Computational Intelligence (UKCI), 2014 14th UK Workshop on 8-10 Sept. 2014, Bradford, UK.

[26] Cuthbertson, R. & Laine, A. J Target Meas Anal Mark “The Role of CRM within retail loyalty marketing.” October 2003, Volume 12, Issue 3, pp 290–304.

[27] Bao Y., Quan C., Wang L., Ren F. “The Role of Pre-processing in Twitter Sentiment Analysis.” In: Huang DS., Jo KH., Wang L. (eds) Intelligent Computing Methodologies. ICIC 2014. Lecture Notes in Computer Science, vol 8589. Springer, Cham.

[28] Hu, G., Bhargava, P., Fuhrmann, F., Ellinger, S., Spasojevic, N. “Analyzing Users’ Sentiment Towards Popular Consumer Industries and Brands on Twitter” in Data Mining Workshops (ICDMW), 2017 IEEE International Conference on 18-21 Nov. 2017, New Orleans, LA, USA.

[29] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman and A. Y. Wu, “An efficient k-means clustering algorithm: analysis and implementation”, in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 7, pp. 881-892, Jul 2002.

[30] M. Zaki, D. Kandeil, A. Neely, J. R. McColl-Kennedy, “The Fallacy of the Net Promoter Score: Customer Loyalty Predictive Model”, Cambridge Service Alliance 2016, October 2016.

[31] Berengueres, J. & Efimov, “Airline new customer tier level forecasting for real-time resource allocation of a miles program”, D. Journal of Big Data, December, vol. 1, issue 3, 2014.

[32] H. B. Sankaranarayanan, B. V. Vishwanath and V. Rathod, “An exploratory analysis for predicting passenger satisfaction at global hub airports using logistic model trees”, 2016 Second International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN), Kolkata, 2016, pp. 285-290.

[33] S.T.M. van Velthoven, “Sentiment analysis on social media to predict Net Promoter Score”, in partial fulfillment of the requirements for the degree of Master of Science in Operations Management and Logistics Eindhoven, October 2014.

[34] M. S. Alam and S. T. Vuong, “Random Forest Classification for Detecting Android Malware”, 2013 IEEE International Conference on Green Computing and Communications and IEEE Internet of Things and IEEE Cyber, Physical and Social Computing, Beijing, 2013, pp. 663-669.

[35] Linna Li and Xuemin Zhang, “Study of data mining algorithm based on decision tree”, 2010 International Conference On Computer Design and Applications, Qinhuangdao, 2010, pp. V1-155-V1-158, 2010.

[36] S. T. Indra, L. Wikarsa and R. Turang, “Using logistic regression method to classify tweets into the selected topics”, 2016 International Conference on Advanced Computer Science and Information Systems (ICACSIS), Malang, 2016, pp. 385-390.

[37] S. Yadav and S. Shukla, “Analysis of k-Fold Cross-Validation over Hold-Out Validation on Colossal Datasets for Quality Classification”, 2016 IEEE 6th International Conference on Advanced Computing (IACC), Bhimavaram, 2016, pp. 78-83, 2016.

A Proposal for a Technological Solution to Improve user Experience in a Shopping Center based on Indoor Geolocation Services

Luinel Andrade, Johan Quintero
School of Computing
Central University of Venezuela
Caracas, Venezuela

Eric Gamess
Mathematical, Computing, &
Information Sciences
Jacksonville State University
Jacksonville, Alabama, USA

Antonio Russoniello
School of Computing
Central University of Venezuela
Caracas, Venezuela

***Abstract*—For shopping centers, mobile devices and their associated technologies represent great business opportunities and a way to improve the user experience within their facilities. These types of constructions are usually quite large, multi-story, and with a significant number of shops, services, where the visitors may find themselves having difficulties to have a complete and up-to-date list of the stores, determine which stores and services are those that meet the characteristics or specifications they seek, know the location of the shops or how to reach them. This research studies and contemplates different technologies, tools, and approaches for the development of a technological solution for shopping centers that offers in its functionality a geolocation system in the interior spaces of the buildings. Our technological solution includes a mobile application for the Android operating system implemented by using the native development approach, and a web application for managing data, where the contents and settings of the mobile application will be obtained following the client/server model through a private API. It is worth to mention that the already mentioned system of geolocation in interiors is implemented using WiFi technology and the different Access Points installed in the shopping center, through which users can obtain their position, locate the stores or services of their interest, and receive indications on how to reach them.**

Keywords*—*Mobile application; web application; geolocation; shopping center; WiFi; access points

I. Introduction

As time goes by, technology is evolving rapidly, thus increasing the solutions that seek to facilitate the day-to-day life of people. It is a fact that technologies such as mobile devices and web applications are part of the daily occupations of people, in both the personal and work activities, and influence numerous aspects of their life such as personal relationships, entertainment, education, and economic activities, among others.

Mobile devices evolve and produce changes in society through the different technological features that they include and the diverse applications that are available to users, generating new trends and behavior models that are increasingly becoming more relevant for organizations and companies from different sectors of activity, creating great business opportunities when taking advantage to them. For these reasons, shopping centers should not only focus on increasing their sales and the number of visitors, but also improve the experience of visitors within their facilities, by providing a service that allows the interaction to be more personal and immediate, generating good memories into the people, and getting closer to their consumers, by incorporating the daily technological resources that they carry and use.

Shopping centers are usually very large constructions, with lot of space, several stories and a large number of shops, services and parking spaces, where visitors may have difficulties to know which stores are in the place and their location, which stores and services are those that meet the characteristics or specifications they seek, and much more. The first objective of this research is to contemplate and analyze various technologies, before developing a mobile application, through which users can see information of a shopping center, browse directories of stores and services, consult maps, have a geolocation system in interiors, and see the various promotions, sales and events offered. Additionally, we also develop a web application that allows the administrative staff of the commercial center to manage the data, content and settings that will be displayed in the mobile application. It is worth to mention that the research was done for a mall in Venezuela that we will not identify for privacy reasons.

The rest of this paper is structured as follows. In Section II, we review the previous work. An introduction to mobile devices and the Android operating system is done in Section III and Section IV, respectively. The different implementation approaches for a mobile application are presented in Section V, while Section VI discusses the development tools. The architecture of our solution is presented in Section VII, with more details of the web application in Section VIII, and the mobile application in Section IX. Section X is dedicated to the evaluation and validation of our work. Finally, Section XI concludes the paper and gives directions for future development.

II. Related Work

Larcomaps is a mobile application developed for the Larcomar shopping center, located in the district of Miraflores in Lima, Peru. It has both Android and iOS versions. This application provides the map of the facilities of the shopping center, with all its stores, parking, etc. In addition, it offers a routing service that shows users the path they must follow to reach a certain destination, such as a store, a bathroom, a food court, or any other internal location of the shopping center. Another important feature of this application is to allow users to add a reminder of the place where they parked their car.

Plaza Las Americas, the biggest shopping center in San Juan, Puerto Rico, has its own mobile application developed for the Android and iOS operating systems. Through the Plaza-Mall App, users are able to access the stores' directory, information about the mall's services, maps, promotions, and offers of the stores, among other functions. Users can purchase gift cards from the mall directly from the application, and also have the possibility to log-in through Facebook or register to receive offers and information through emails. To facilitate the location of their vehicle, users can save the information of where they parked, by taking a photo, recording a voice message, or marking the location with a pin.

Indoor Location [1] is an application developed by a group of students from the Complutense University of Madrid, Madrid, Spain, which allows indoor geolocation of mobile devices using WiFi networks. It is developed with Java and uses Swing and AWT (Abstract Window Toolkit) as graphic libraries, which makes it a completely cross-platform application.

ProMotion App [2] is an application that presents all the stores and shops affiliated with the system to users, where they can see promotions of the stores through notifications when physically approaching them. The promotions notifications appear in the upper part of the screen of the mobile devices when they are within a range of 0-30 meters from the associated stores, using Bluetooth Low Energy technology and Beacons devices as a location method.

III. Mobile Devices

A large number of electronic devices are currently classified as mobile devices, for example: cell phones, portable game consoles, tablets, digital agendas, calculators, digital cameras, among others. With this diversity of devices, it is difficult to determine which characteristics must be possessed to be considered as mobile devices. Some essential aspects that they fulfill are: (1) they are small devices that can usually be transported in the pocket of the owner or in a small bag, (2) they have processing capacity, permanent or intermittent connection to a network, and memory, and (3) they are usually associated with personal use and have a high capacity for interactions through the screen or the keyboard.

This research focuses on two types of mobile devices that are mobile phones (specifically smartphones) and tablets.

A. *Operating Systems for Mobile Devices*

In a few words, it could be said that an operating system (OS) is a program that manages the hardware of a computer. It also provides basic application programs and acts as an intermediary between users and the computer's hardware [3].

A mobile device also has its own OS, which is loaded and executed when the device is turned on, and in charge of managing the resources of the system, both software and hardware, thus allowing communications between the equipment and users. Mobile operating systems are focused on mobility, wireless connectivity, multimedia formats and the optimal management of processing, storage, and energy consumption [4].

- Android: It is an operating system and a software platform, based on Linux for mobile phones, although it is also used for other devices such as tablets, wearable devices, music players, and even netbooks. Older versions of Android allow users to program applications in a Java environment, on a Dalvik virtual machine that is a variation of the Java virtual machine with compilation at runtime [5]. Newer versions of Android are based on Android Runtime (ART), which uses ahead-of-time (AOT) compilation to entirely compile the application bytecode into machine code upon the installation of an application. It was developed by Android Inc., a company that was later bought by Google in 2005, but it was not until 2008 when it began to become popular, thanks to the Open Handset Alliance consortium, which decided to join the project, promote free software, and develop open standards for mobile devices.

- iOS: It is an operating system developed by Apple Inc. for its devices such as iPhone, iPod Touch, and iPad. Since it is a proprietary operating system, it cannot be installed on devices of other brands. The first version of iOS was presented in 2007 along with the first iPhone. It is derived from MacOS, which was also developed by Apple Inc. for its range of Macintosh computers since 2002.

- Windows Phone: Windows Phone is a mobile operating system developed by Microsoft for its smartphones that was first released in October 2010. It was introduced as the successor of Windows Mobile 5, 6, and Zune, and has a number of changes that make it primarily aimed at the consumer market rather than the business market. Microsoft dropped it in the first quarter of 2015, to focus on the mobile operating system called Windows 10 Mobile, which primary focus is the unification with Windows 10, its PC counterpart.

- Blackberry OS: It is a proprietary operating system embedded in the wide range of mobile phones of the Canadian company BlackBerry Limited, formerly called Research In Motion (RIM). BlackBerry OS has a core based on the Java Virtual Machine (JVM). It is written in Java and C ++.

- Firefox OS: It is an open-source mobile operating system, based on HTML5, with a Linux kernel, for several platforms. It is composed of 3 layers, which are: Gonk (the core), Gecko (the web rendering engine), and Gaia (the user interface). At the end of 2015, the Mozilla Foundation concluded the development of the Firefox OS system for mobile phones and announced the end of its development. The main obstacles encountered in the development of the system were commercial. The system was not sold and the costs exceeded the benefits.

IV. Android Operating System

In this section, we focus on the Android OS because we selected it for the development of the mobile application. Android is an operating system and a software platform based on the Linux kernel [6]. It is open-source, and can also be freely extended to incorporate new cutting-edge technologies that are emerging. It has been written for mobile devices and one of the principles of its design is to give developers the ability to create mobile applications that take full advantage of all the tools that modern mobile devices can offer.

A. *Architecture of Android*

Android is a stack of open-source software based on Linux created for a wide variety of devices. Fig. 1 shows the main components of the Android architecture [7].

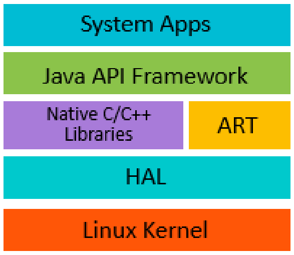


Fig. 1. Architecture of android.

- Linux Kernel: The kernel is the basis of an operating system. It provides fundamental system services such as security, memory management, process management, network functions, and driver management [7].

- Hardware Abstraction Layer (HAL): Provides standard hardware interfaces of the device to the Java API Framework. The HAL consists of several library modules and each of these implements an interface for a specific type of hardware component, such as the camera or Bluetooth modules [7].

- Android Runtime: Android runtime (ART) is the managed runtime used by applications and some system services on Android. ART replaced the Dalvik Virtual Machine (DVM) since version 5.0 of Android. Both were specifically created for this platform and are compatible with each other.

- Native C/C++ Libraries: Many components and central services of the Android system, such as ART and HAL, are based on native code that requires native libraries written in C and C ++. The Android platform provides the Java API Framework layer to expose the functionality of some of these native libraries to applications [7].

- Java API Framework: The Java API Framework layer is a set of APIs written in Java where the full set of Android OS functions is available. These APIs are the foundations needed to create Android applications, simplifying the reuse of system components and services, such as the following [7]: View System, Resource Manager, Notifications Manager, Activities Manager, and Content Providers.

- System Applications: Android includes a set of applications for camera, email, SMS messaging, calendar, contacts, Internet browsing, among other elements. These system applications provide key functionalities that developers can access from their own applications.

B. *Android Applications and their Components*

A mobile application, also known as an app, can be defined as a computer application designed to be executed on mobile devices to perform tasks, allowing users to satisfy needs of different types such as educational, professional, entertainment, health, etc. Mobile applications are found and run at the last level of the Android architecture and use services, APIs, and libraries of the lower levels. The file format used in Android to install applications is the .apk (Android Package). This .apk format is a variant of the Java .jar format.

Next, we present the different components related to an Android application.

- Activity: It serves as an entry point for the interaction of a user with an application, since it is the window in which the application displays its user interface. Also, it is fundamental for the navigation within an application or between applications.

- Intent: It is an object that can be used to request an action from another application component. Intents facilitate communication between components, and there are three fundamental use cases that are: (1) to start an activity, (2) to start a service, and (3) to deliver a broadcast.

- Service: It is an application component that can perform long-running operations in the background, and that does not provide a user interface. It can be started by another application component, and be run in the background, even if the user switches to another application.

- Content Providers: They manage the access to a structured set of data. They encapsulate the data and provide mechanisms to define data security. Content providers are the standard interface that connects data in one process with code that is executed in another process [8].

V. Development Approaches for Mobile Applications

Choosing a development approach for a mobile application is one of the decisions that most influence the final result of the product, since it involves taking into account many aspects such as budget, project deadlines, application functionality, type of users to which the app is directed, among others. In the following subsections, the different development approaches are described, to give an idea of the advantages and disadvantages of each one.

A. Native Applications

It is a mobile application that is targeted to be installed and executed in a specific operating system. This type of applications can freely access all the APIs offered by the OS and, in many cases, has unique functions and features that are typical of the platform on which it is run.

The development process is usually similar for different operating systems, but the SDK (Software Development Kit) is specific to each platform, and each mobile OS has its own tools, languages, formats, and distribution channels [9]. The main disadvantage of this approach is that the mobile application developed for a specific operating system cannot be used in another platform, so that the same application would carry a different development and maintenance process in each OS.

In turn, the fact that an application is developed for a specific operating system is an advantage, since it allows to take better advantage of its specific hardware such as GPS, accelerometer, camera, compass, among others, because their implementations are developed in native programming languages for each OS and accessed through proprietary APIs of each system, hence obtaining a better performance compared to other development approaches.

B. Web Applications

These are applications that run in a web browser, and can be adapted and optimized for any mobile device. Modern mobile devices have powerful browsers that support the new functionalities of the HyperText Markup Language version 5 (HTML5), Cascading Style Sheets (CSSs), and the JavaScript (JS) language.

The main advantage of using web-based mobile applications is that they are multiplatform, so once developed, they can be distributed to different operating systems, which helps to minimize development costs when comparing to the development of a native application. Another advantage is that there is no need to control the version of the application that is published or the subsequent updates in the clients, since new versions or updates automatically reach all users, regardless of the OS or device they have.

Among the disadvantages of this development approach [10] is that an important part of the functionality of the device is being lost by not developing natively with the SDKs and APIs available for each platform. That is, in web applications, many of the features of the device cannot be used or only partially.

C. Hybrid Applications

A hybrid application is one that combines native development with web technology, where a large part of the application is developed with web technologies such as HTML5, CSS, and JavaScript, for multiple platforms, maintaining direct access to native APIs if needed; then it is wrapped into a native container using a packager, which acts as an intermediary and translates all the instructions so that the operating system of the device can understand them when executing the application. The native portion of the application uses operating system APIs to create a built-in HTML search engine that functions as a bridge between the browser and the device's APIs. This bridge allows the hybrid application to take advantage of all the features offered by modern devices [9].

The main advantage of hybrid applications is that, similarly to web-application, they are cross-platform applications so that they can be executed in different operating systems. Also, due to the aforementioned, they allow the reuse of code, so there is no need to write a different application for each system. Another important advantage is the "partial" direct access to native APIs and resources of the operating system.

However, one of the main disadvantages is that it does not have full access to all hardware resources and APIs available due to the limitation of the framework used to develop the hybrid app. Additionally, its performance is lower than the one of a native application since the latter run at a lowest level than hybrid ones, because they are developed directly for a specific operating system. Instead, a hybrid application has to be packaged, and each instruction has to go through a translation process before being executed.

D. Comparing the Development Approaches

Native applications stand out for their better performance and for having better access to the APIs and mobile device functionalities. However, they are not multiplatform and tend to have more expensive development processes in effort, time, and resources. Web-based applications are cross-platform, have a simpler and less expensive development process, and their updates are faster since the application resides on a server and reaches all users when the resources are requested. However, their biggest disadvantage is that they do not take advantage of all the functionalities of the devices. Hybrid applications are halfway between the two approaches mentioned above, offering some of the advantages that each one can have, especially if the developers want the application to be used in various operating systems, even though, most likely, the level of performance is lower.

The approach chosen by us to develop our application to improve user experience in a shopping center is the native one. Among some of the reasons for this choice is to have a user interface closely linked to the design trends of Android, and that we need to use native APIs to take advantage of certain functionalities of mobile devices related to WiFi.

VI. Development Tools

In this section, we present a series of tools and technologies that are important and necessary to develop the technological solution in which the study is focused. To understand a little more why some of the technologies and development tools that follow are considered, it is worth to remember that the mobile application to be developed makes use of an indoor geolocation system, and the web application that supports the creation and administration of the contents uses a Content Management System (CMS).

A. Web Application

Regarding the development of web applications, a type of abstraction is made to differentiate and separate two parts of the system that are: Front-End and Back-End.

- Front-End: This represents the processes, functionalities, and technologies of the web application that run on the client's side, that is, the processes and technologies that run or are on the side of the web browser. It is usually associated with the visual part of the application, such as: the structure, styles, colors, animations, and effects.

  The technologies used for their development are: HTML5, CSS, and JavaScript; where in a summarized way, HTML5 is used to define the structure, CSS to assign the styles, and JavaScript to bring dynamism to the web application.

- Back-End: This part of the application represents the processes that are executed on the server side, which are not visible to users, such as the processing of data, requests, etc. In addition, it allows the handling and manipulation of the processed data which implement the business logic, that is, the implementation of the functions, connections, and queries to the database, the management of sessions, among others.

  Some of the programming languages that allow the development of this part of the application are PHP (PHP Hypertext Preprocessor), Python, .NET, Java, Node.js, etc. Developers can also use frameworks, which allow not only the development of the application’s Back-End, but also the development of the Front-End.

B. *Frameworks*

They are software structures made up of customizable and interchangeable components for the development of an application. These frameworks have a set of tools, libraries, conventions, and good practices that seek to encapsulate repetitive tasks in generic reusable modules.

Currently, frameworks are used mostly in the development of web and mobile applications. Some examples of frameworks for web development are: Laravel, CakePHP, Symfony, Zend, CodeIgniter, Yii2, Ruby on Rails, among others. On the other hand, examples of frameworks for mobile development are: Ionic, jQuery Mobile, PhoneGap, and React-Native.

For this project, we decided to build the web application with Laravel. It is an open-source framework that was officially launched in 2011, multiplatform, and that allows the development of web applications and services with the PHP programming language.

Some of the advantages of this framework that have motivated our choice are that it has abundant documentation on the web, a large community of users, forums, it is modular and with a wide variety of package systems and drivers to extend its functionality. Also, it has different ways for accessing relational databases, and it provides utilities that benefit the application’s deployment and maintenance. For example, it supplies a command line interface tool called Artisan that allows performing tasks such as data migrations, it has an ORM (Object-Relational Mapping) called Eloquent that makes the interaction with databases totally oriented to objects, and finally, it is adaptable to the MVC (Model-View-Controller) pattern.

C. *Database Management System (DBMS)*

A DBMS [11] (Database Management System) is a coordinated set of programs, procedures, languages, etc., which provides the necessary means to describe and manipulate data stored in a database, guaranteeing their security. For our project, we selected MySQL that was originally developed by MySQL AB. This company was purchased by SUN Microsystems in 2008 and, in turn, SUN Microsystems was acquired by Oracle Corporation in 2010. We used the community version under the GNU General Public License.

Some of the reasons why we decided to work with this DBMS are: (1) it is an open-source database, (2) it is supported and compatible with multiple programming languages, and (3) it is very stable in applications where there is low concurrency in the modification of data but a high level of data reading, which fit to our needs of data management in both the web application and the mobile application.

D. *REST Application Programming Interface (API)*

The REST API [12] is an architecture for the exchange and manipulation of data. Currently, it is the most frequently used because it is the most logical and efficient architectural style in the creation of APIs for Internet services. REST allows the creation of applications and services that can be used by any client that understands HTTP to obtain data or generate operations on that data in various formats such as XML and JSON (JavaScript Object Notation).

A reason that motivates the utilization of REST is that it makes use of the global addressing scheme based on Uniform Resource Identifiers (URIs) which identify resources or conceptual objects. The representation of such objects is distributed through messages in the web.

E. *Java*

We used Java as a programming language to develop our mobile application, because it is the official language used to develop native mobile applications in the Android operating system. Java is a language with the following characteristics [13]: simple, object-oriented, distributed, robust, secure, multi-threaded, portable, neutral architecture, and high performance. The Java Virtual Machine (JVM) is a virtual environment that is used to interpret the Java bytecode, since Java was made to run on any platform without recompiling the binaries. That is, the bytecode can be obtained in any architecture and operating system, and be executed in any other system with a JVM.

F. *Integrated Development Environment (IDE)*

It is an application that contains a set of programming tools to facilitate the development of applications, that is, it is made up of a source code editor, a compiler, a debugger and tools to simplify the construction of Graphical User Interfaces (GUIs). We selected Android Studio, the official IDE for the development of Android applications, for the implementation of our mobile application. It is based on the IntelliJ IDEA software from JetBrains. We chose Android Studio since it provides good tools for the fast creation of applications for all types of Android devices. In addition, it has a powerful code editor, the developer tools of IntelliJ, and even more features that increase productivity during the development of Android applications.

G. Geolocation with WiFi

Geolocation can be defined as the ability to calculate and obtain the geographical location of an object, such as a smartphone. Some of the technologies that can be used to carry out the geolocation are: GPS (Global Positioning System), RFID (Radio Frequency Identification), and WPS (WiFi-based Positioning System).

1) WiFi

We decided to choose the WiFi technology to implement the functions related to geolocation. WiFi is one of the communication technologies with the greatest presence in the world. It is based on the 802.11 standard of the Institute of Electrical and Electronics Engineers (IEEE), which has conquered the market since the first approval of the 802.11b specification in 1999 [14].

One of the main elements in WiFi networks in infrastructure mode are the Access Points (APs), which allow communications between the wired network and wireless devices. An AP receives and sends data, both through Ethernet cables and wirelessly. Other functions associated with these equipments are the access control, security, and isolation of the ambient noise or improvement of the signal/noise ratio [15]. The range of WiFi devices will depend on many factors, such as walls, ceilings, obstacles, etc.; and there may be other factors that, in turn, increase the range of the signal, such as the installation of WiFi repeaters.

2) Motivation to use WiFi Technologies for Geolocation

WiFi has advantages over the GPS geolocation system, because the latter does not work in interior spaces since GPS receivers cannot receive the signal from the satellites.

Another advantage of WiFi is that it is a widespread and daily use technology presents in many places such as homes, offices, schools, libraries, shopping centers, squares, etc. In the case of shopping centers, it is very common that there is a large number of APs inside the building, whether they are owned by the stores or they are part of the network infrastructure of the building, which makes it viable to have access to this technology. This high presence of APs within the facilities of those constructions allows data to be collected from the WiFi networks, and once a relationship between a group of WiFi networks and a certain geographical point is established, the location of users can be determined to support the indoor positioning system required by our mobile application.

3) Method of the k-Nearest Neighbors

Most methods of geolocation using WiFi technology are based on the signal strength with which the APs are detected by mobile devices. Depending on the algorithm used, the precision and complexity of the approach vary. Some of the methods to carry out the geolocation using WiFi networks are: power vector, power triangulation, proximity heuristics [6], method of the k-nearest neighbors [16], [17], probabilistic methods [15], and location using RTT through ping messages [1].

We selected the method of the k-nearest neighbors to develop the geolocation functionalities of the mobile application since, according to our research, it is the most appropriate for factors such as: (1) implementation, (2) calculation cost, and (3) precision. Below, we present more information about the method in details.

The k-nearest neighbor method is based on the definition of a metric in the space of cells covered electro-magnetically by the APs. Each cell or area of the space considered to perform the geolocation is assigned a value for every AP that represents the power of its signal received in that place. When the power is measured at an arbitrary point, the distance is defined as the difference between the initially recorded power during the fingerprinting process and the measured power. The absolute value of this difference will be considered as a distance [15].

Once a measurement is made in a specific place, the method looks for a point that minimizes the sum of the squares of the distance. If there were no measurement errors, there should be a point for which the quadratic sum of distances is zero. However, since errors always exist, the point for which the quadratic sum of the distance is minimum is considered to be the best approximation that can be obtained.

The fingerprinting process should be done as a preliminary phase to the implementation of the k-nearest neighbor method. To do so, a map of the place is required, and it must be divided into sectors or areas that have uniform measurements. The locations of the APs and the locations where the measurements will be made must be recorded on the map. Then the building must be traversed, and in each measurement location, the values of the strength of the signal received from each accessible AP must be kept. It is recommended to take different measurements for each AP from the most possible centered position of each of the sectors, and each geographical orientation, keeping the height of the device constant.

Depending on the distance between the measuring points that are taken during the fingerprinting phase, the precision of the algorithm will vary. This process is performed once and serves to calculate the position of all users. When there are structural changes in the building or considerable changes in the infrastructure of the wireless network (e.g., new positions for some APs, new APs, eliminated APs), there is a need to repeat the fingerprinting process. Once the signals have been scanned, each sector must be assigned a characteristic measurement that represents the signal of each AP in that location.

To get the position of a device, the measurement of the power of the signal must be made in the device for each AP, obtaining as a result a vector $x_1, x_2, \ldots, x_N$, where N is the number of APs. It is assumed then that $P_{Aj,k}$ is the power that was measured from AP k in region Aj.

Consider the sum (1), the estimation of the point where the signal measurement was made will be the region for which S_{Aj} is minimum.

$$S_{Aj} = \sqrt{\sum_{k=1}^{N} (P_{Aj,k} - x_k)^2} \qquad (1)$$

VII. Architecture for the Proposed Solution

Fig. 2 depicts the general architecture of the technological solution, where the position of each element to be integrated is specified.

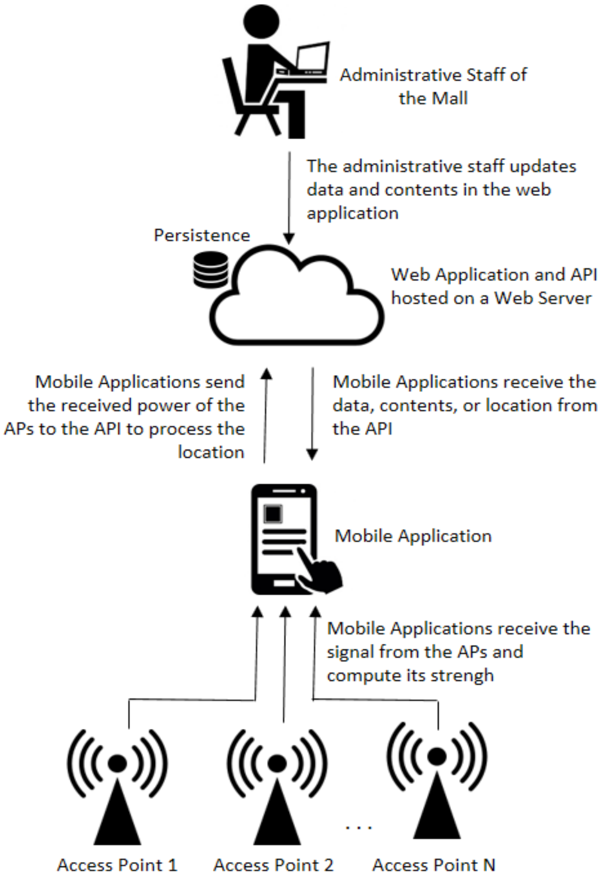

Administrative Staff of the Mall
The administrative staff updates data and contents in the web application
Persistence
Web Application and API hosted on a Web Server
Mobile Applications send the received power of the APs to the API to process the location
Mobile Applications receive the data, contents, or location from the API
Mobile Application
Mobile Applications receive the signal from the APs and compute its strengh
. . .
Access Point 1
Access Point 2
Access Point N

Fig. 2. General architecture of the solution.

VIII. Introduction and Analysis of the Web Application

The web application for the management of the data, contents, and configurations that the mobile application consults was developed using the Laravel framework and MySQL DBMS for the persistence of data. In general, the functionalities offered by this application are the management of the data to be consumed by the mobile application through the API. These data include: global information of the shopping center, information about shops, services, categories, promotions, events, and data related to the geolocation system such as maps, sectors, access points, and measurements done during the fingerprinting process. In addition, the web application allows the management of users and users' profiles, inherent of its administration.

A. Development of the Web Application

To understand how the development process was made using the Laravel framework, it is worth to mention that all the modules had a similar development. Hence, the general process to carry out the coding is explained below.

Once configured the connection to the MySQL DBMS, the models can be created from the command line using the commands from Artisan. The models were created in the folder "/app;" which are PHP files with information of the tables of the database to be used for the respective model. Then, the migration file must be created. Migration files are found in the path "/database/migrations" and allow the creation of the structure of a table of the database. This type of files facilitates a version control on the tables of the database.

The controllers can be created using commands from Artisan, in order to handle the events generated by users and invoke the requests to the models when they are made on the information. These controllers are PHP files that are generated in the path "/app/Http/Controllers." In the header of those files, it is necessary to import the model of the represented entity and any other models that are required to be used in that controller.

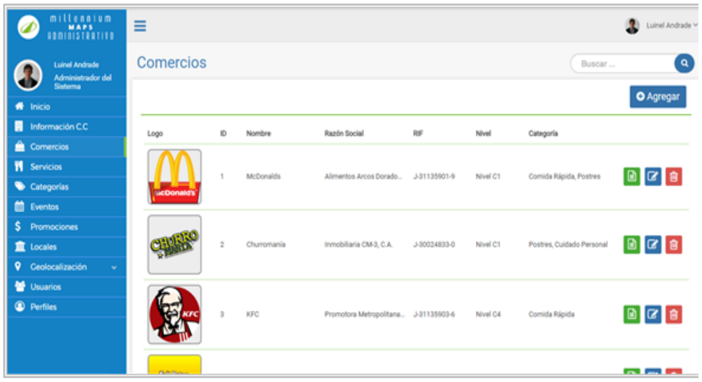
Once having the controllers, it must be ensured that the application has the necessary routes configured. The routes of the application are defined in the file "routes.php" located in the path "/app/Http." When accessing a route from a browser, an HTTP request is sent to the aforementioned file. An error is generated if the route does not exist. Otherwise, it redirects to the controller that contains the logic of the application. Instead of completely defining the logic of the requests in the file "routes.php," it is also possible that this behavior is organized through classes of type "Controller". Controllers work with the GET, POST, PUT, and DELETE requests, which can be grouped and manipulated logically in a class. The methods of the controllers are associated with the HTTP requests in the following way: GET (index, create, show, edit), POST (store), PUT (update), and DELETE (destroy). When using a type of route called "resource" in the file "routes.php," Laravel internally creates the application routes related to a specific resource and assigns them to the controller's methods.

The logic of the application is coded in the different methods of the controller. Some of the different actions that are carried out in these methods are: process the requests generated by the users, interact with the model for the manipulation of the data, validate data in the forms, validate the users' permissions on the different modules, send data received from the model to the views, etc.

The views can be understanding as the public part of the system that users will be able to see. They are implemented using HTML to define the structure, CSS to assign the styles, and JavaScript to apply the dynamism. In addition, in the case of Laravel, a templating engine called Blade is also used. The views are located in the path: "/resources/views".

In the framework, template files can also be used and are usually the files of the main views, which have the segments of code that are repeated in more than one view, such as the header, the main menu, and the footer. The idea is to keep common code in a specific place, instead of repeating it in several views.

As an example, Fig. 3 depicts the view shown to users when running method "index()" in the Commerce module (i.e. the directory of stores that are stored in the database).

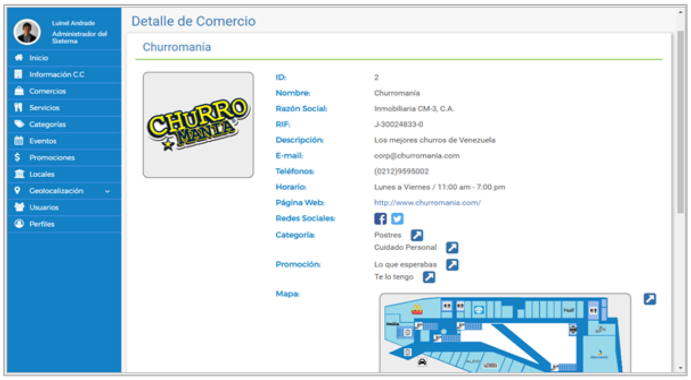


Comercios

Logo	ID	Nombre	Razón Social	RIF	Nivel	Categoría
	1	McDonalds	Alimentos Arcos Dorado...	J-31135901-9	Nivel C1	Comida Rápida, Postres
	2	Churromania	Inmobiliaria CM-3, C.A.	J-30024833-0	Nivel C1	Postres, Cuidado Personal
	3	KFC	Promotora Metropolitana...	J-31135903-6	Nivel C4	Comida Rápida

Fig. 3. Directory of the stores.

Fig. 4 depicts the view shown to users when running method “show()” in the Commerce module (i.e. the specific details associated with a store).



Detalle de Comercio

Churromanía

ID: 2
Nombre: Churromanía
Razón Social: Inmobiliaria CM-3, C.A.
RIF: J-30024833-0
Descripción: Los mejores churros de Venezuela
E-mail: corp@churromania.com
Teléfonos: (0212)9595002
Horario: Lunes a Viernes / 11:00 am - 7:00 pm
Página Web: http://www.churromania.com/
Redes Sociales:
Categoría: Postres
Cuidado Personal
Promoción: Lo que esperabas
Te lo tengo
Mapa:

Fig. 4. Detailed view of a specific store.

B. Development of the API

Once developed all the modules of the web application, it was straightforward to write the API since it was only necessary to develop new controllers that will validate a key called API Key, through a middleware. In the API of this project, it was only necessary to implement query methods on the database. The queries are made through GET requests of the HTTP protocol.

In the API for routes, one of the most important functionality is done by the method to get the geolocation in the interior space of the shopping center. This method returns the sector of the shopping center in which a user is located and is available in the path “api/med/locate”. In this function, the method of k-neighbors is implemented to compute the user’s location, according to the strength of the signal received from the APs.

To get the position of a user, the mobile application sends the information (MAC Address, signal strength, etc.) of the reachable access points to the API. The API receives the request, analyzes and processes the received data, before returning the `<Map-ID>` and `<Sector-ID>`, of the actual location.

IX. General Analysis of the Mobile Application

The mobile application was developed using Android Studio as an IDE and Java as the programming language, since Java is the official language to develop applications for Android. For the development of the application, we used Git for version control, and the GitLab platform which is based on Git and allows private repositories for free.

At a glance, the application has a set of functionalities that allow users to see the list of shops of the mall, the promotions of these shops, the events and services offered by the shopping center, as well as the maps of each floor. In addition, it has a function to record the parking ID where the user’s vehicle is parked, and a geolocation system so that the user can be located inside the shopping center.

A. Development of the Mobile Application

For the development process, different types of files were used. Java classes allowed to manage the behavior and functionalities of the application. The layout files, with “XML” extension, represent the user interface of the application. It is worth to mention that the layouts are associated with the Java classes so that, from the code, it is simple to get access to the elements of each view.

In addition to the aforementioned files, other files are used to represent certain resources that the application needs such as: images, strings, colors. To represent the images, a directory called “drawable” is used. For the usage of certain text strings defined in the project, an XML file called “strings.xml” is used. The “colors.xml” file allows the establishment of the colors.

The file “AndroidManifest.xml” was also used. In this file, the Android Manifest provides essential information about the application to the operating system so it can execute the code of the application. Finally, the gradle files represent the settings for the compilation of the application, the minimum SDK on which the application can run, the version of the SDK with which the application was compiled, its dependencies, among others.

To obtain all the information that is shown in each module of the application, requests are made to the API, using Retrofit, which is a REST client for Android and Java. These requests are made in a file called “APIServices.java”, where the type of request is specified (GET, POST, PUT, or DELETE), the URI, the name of the function to which the call will be made from the code, the object to receive the response, and the parameters that will be passed to the request. Upon receiving the response of these requests, they are stored in an object according to the models defined in the applications which are made up of the properties and their respective getters and setters.

As for the modules of the application, some of them list elements such as: stores, categories, promotions, events, and services; in each of these lists, when selecting a specific element, it is presented/shown in details.

Fig. 5 depicts the screens of the lists of shops, promotions, and events. For each of the previous category, the elements are shown, with certain “filtering” options, for a better selection.

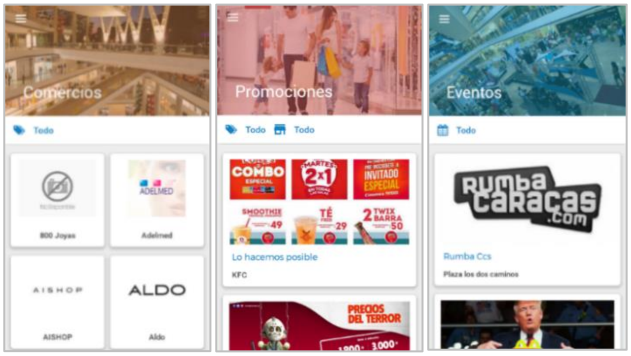
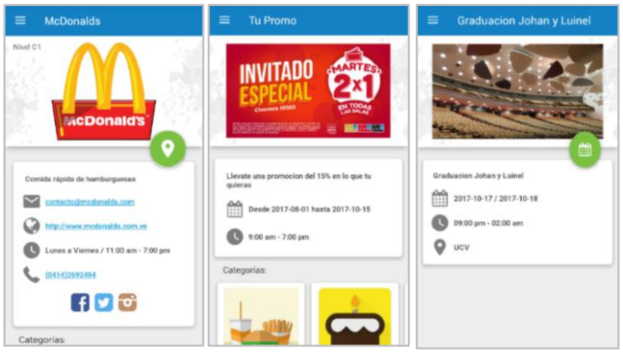


Fig. 5. Screenshot of the lists of shops, promotions and events.

Fig. 6 shows the information associated with a store: Store Info, Promotions, and Events. In “Store Info”, a general description of the store is presented, and includes: (1) the floor where the store is situated, (2) a brief description of the store, (3) its electronic mail, (4) the URL of its web page, (5) its opening hours, (6) its telephone number, (7) its social networks, and (8) the categories to which it belongs. At the level of a promotion, the information includes: (1) a brief description of the promotion, (2) the range of dates of the promotion, (3) the range of hours of the promotion, and (4) the categories to which it belongs. At the level of an event, the information includes: (1) a brief description of the event, (2) the range of dates of the event, (3) the range of hours for the event, and the place where the event will be celebrated. In addition, the details of the shop and the event have an additional option. In the case of the shop, users can view the location of the shop in the maps of the shopping center. In the case of the event, users can add this event to the calendar of their mobile device.


McDonalds

Nivel C1

Comida rápida de hamburguesas

contacto@mcdonalds.com

http://www.mcdonalds.com.ve

Lunes a Viernes / 11:00 am - 7:00 pm

(0414)2692494

Categorias:

Tu Promo

Llevate una promocion del 15% en lo que tu quieras

Desde 2017-08-01 hasta 2017-10-15

9:00 am - 7:00 pm

Categorias:

Graduacion Johan y Luinel

Graduacion Johan y Luinel

2017-10-17 / 2017-10-18

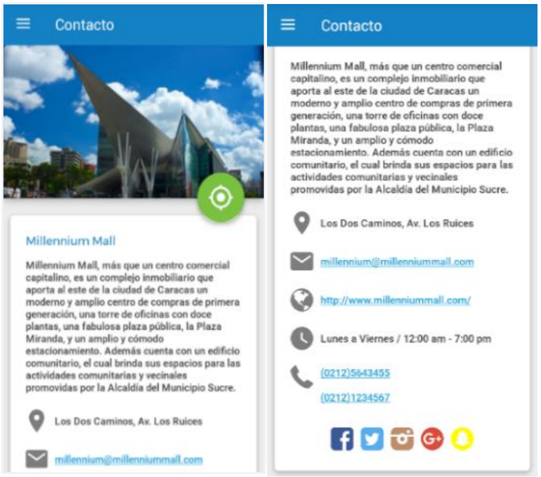
09:00 pm - 02:00 am

UCV

Fig. 6. Detailed screenshot of a store info, promotions, and events.

Furthermore, the application has modules such as: “Contact”, “About the Application”, and “Your Parking Position”. The “Contact” module has general information about the shopping center, including contact information. The “About the Application” module provides useful information about the developed application (the authors). The “Your Parking Position” module allows users to view, modify, and eliminate the parking position where their vehicle is stationed.

Fig. 7 shows the “Contact” module, where the general information of the shopping center is given, which includes: (1) a brief description, (2) the address, (3) email, (4) web page, (5) opening hours, (6) telephone numbers, and (7) social networks. In addition, users have the option to locate the shopping center on Google Maps.


Contacto

Millennium Mall

Millennium Mall, más que un centro comercial capitalino, es un complejo inmobiliario que aporta al este de la ciudad de Caracas un moderno y amplio centro de compras de primera generación, una torre de oficinas con doce plantas, una fabulosa plaza pública, la Plaza Miranda, y un amplio y cómodo estacionamiento. Además cuenta con un edificio comunitario, el cual brinda sus espacios para las actividades comunitarias y vecinales promovidas por la Alcaldía del Municipio Sucre.

Los Dos Caminos, Av. Los Ruices

millennium@millenniummall.com

Contacto

Millennium Mall, más que un centro comercial capitalino, es un complejo inmobiliario que aporta al este de la ciudad de Caracas un moderno y amplio centro de compras de primera generación, una torre de oficinas con doce plantas, una fabulosa plaza pública, la Plaza Miranda, y un amplio y cómodo estacionamiento. Además cuenta con un edificio comunitario, el cual brinda sus espacios para las actividades comunitarias y vecinales promovidas por la Alcaldía del Municipio Sucre.

Los Dos Caminos, Av. Los Ruices

millennium@millenniummall.com

http://www.millenniummall.com/

Lunes a Viernes / 12:00 am - 7:00 pm

(0212)5643455

(0212)1234567

Fig. 7. Screenshot of the contact module.

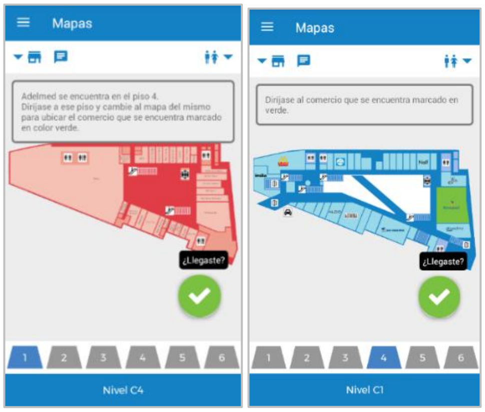
Finally, the “Maps” module was developed, which represents the most complex part of the application. This module has two operation modes: (1) the user is inside the facilities of the shopping center, and (2) the user is outside the shopping center. When the user of the mobile device is outside the shopping center, the “Maps” module displays a message to inform the user that he/she is outside the shopping center facilities, and therefore he/she cannot be located within it. However, the user can still make use of the other functionalities of this module such as seeing the maps of each floor of the shopping center, etc.

The function of “How to Get to a Store” consists in showing the list of shops of the shopping center, and when the user selects a specific store, an emergent message is shown with the instructions on how to get to the required shop. The function “List of Services” consists in showing the list of services of the shopping center. When the user selects a specific service, he/she is shown a map per floor that has only the requested service.

Fig. 8 depicts the “Maps” module, which shows the directions on how to get to two different shops (Adelmed and Mercantil). Additionally, users have the option to see the list of shops, the available services, and there are six buttons to change the current floor.

When the user is inside the facilities of the shopping center, the “Maps” module works practically in the same way as the previous case, with the difference that in this case, when the map module is started, the map of the floor where the user is located is shown, and the actual sector of the user is marked in green on the map.

Once the Maps module is started, a capture of the neighboring access points is made every eight seconds, to get the MAC address, the SSID, and the strength of the signal of each access point. Subsequently, these data are sent to the API, and the response of the request is the sector where the user is located, at the time of each capture.



Adelmed se encuentra en el piso 4.
Dirijase a ese piso y cambie al mapa del mismo para ubicar el comercio que se encuentra marcado en color verde.

Dirijase al comercio que se encuentra marcado en verde.

Fig. 8. Screenshot of the maps module with indications on how to reach a specific store.

In each capture, the sector where the user is located is obtained, and the picture of the corresponding map is shown, with this sector in green. In the next capture, if the new sector that is obtained as a result of the API is in a different floor, a popup message is shown to inform the user of the new floor, and the picture of the map of the new floor is loaded, with the new sector in green. Fig. 9 depicts the map of the 4th floor (Level C1) with the actual sector of the user shown in green.

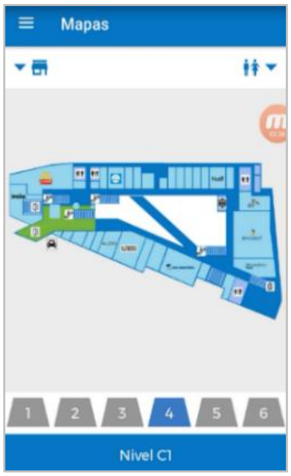


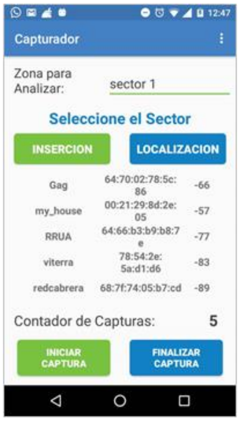
Fig. 9. Screenshot of the map of the fourth floor (level c1) where the green sector specifies the actual position of the user.

B. *Development of Tools for the Process of Fingerprinting*

In addition to the development of the web application and the mobile application in which the present research work is focused, we also developed a second mobile application for Android OS and a program in C++ to support the fingerprinting process that must be done to feed the database, and be able to process the algorithm of the nearest k-neighbors. The objective of this second mobile application is to facilitate and speedup the recollection of data from the APs in the different sectors of the shopping center.

Its usage is straightforward and consists in identifying the sector to be captured, before selecting if the data is collected for insertions in the database or for testing the geolocation functionality in the API. For each reachable AP, the data captured are: SSID, MAC, and signal strength. Fig. 10 shows a screenshot of the application.

Once the fingerprinting process is over, the results of the capture of the mobile device must be transferred to a PC. In the PC, the C++ program reads the capture, groups the results by AP, and computes the average of the signal for each AP. Then, the C++ program stores the results in a file with a specific format with four data: (1) the MAC of the AP, (2) the SSID of the AP, (3) the average of the strength of the signal, and (4) a recommendation of whether or not the information associated with the AP must be stored in the DB. This decision is based on the stability of the signal captured from the AP. That is, if the standard derivation is low, the AP is considered stable, and the C++ program will recommend the insertion of its data in the DB.



Zona para Analizar: sector 1

Seleccione el Sector

Gag	64:70:02:78:5c:86	-66
my_house	00:21:29:8d:2e:05	-57
RRUA	64:66:b3:b9:b8:7e	-77
viterra	78:54:2e:5a:d1:d6	-83
redcabrera	68:7f:74:05:b7:cd	-89

Contador de Capturas: 5

Fig. 10. Screenshot of the mobile application for the process of fingerprinting.

X. Tests and Analysis of Results

In this section, we present the tests made to the mobile application and the results obtained, in order to evaluate the level of usability, the level of quality, and the compliance with the specific objectives and functions for which it was created.

A. *Acceptance Tests*

A technical survey was applied to a group of 50 end-users of different genders, social and educational level, in order to assess the usability and proper functioning of the app, and determines if it meets the requirements established in the project. The survey consisted of twenty questions that are listed below:

1) *How do you rate the combination of colors used for the application?*

2) *How do you evaluate the distribution of the elements in the application?*

3) *How do you assess the application design?*

4) *How do you rate the navigation in the application?*

5) *How do you evaluate the fluency when running the application?*

6) *Do you consider that the application is easy and intuitive to use?*

7) *How do you evaluate the splash screen shown when launching the application?*

8) Does the application's menu displays its options clearly and legibly?

9) Do you consider that the elements shown in the Home of the application are the indicated?

10)Do you consider that "Your Parking Position" is an important functionality for the application? That is, the functionality of letting the user write down his/her parking slot, and receiving directions on how to get back to the car.

11)Do you consider that it is easy and intuitive to handle the functionality of "Your Parking Position"?

12)Do you consider that the list of shops, promotions, events, and services clearly shows each element and what they refer to?

13)Is the filtering option easy to manage in the list of shops, promotions, events, and services?

14)Do you consider that the information shown for the elements (stores, promotions, events, services) is adequate?

15)Is the option of "location" can be clearly understood in the Trade Detail, Service Detail and Contact?

16)Is the option "Add to the Calendar" is clearly understood in the information of events?

17)Is it clear that the texts that represent emails, web pages, and telephone numbers are clickable?

18)Do you consider that it is easy and intuitive to use the Maps module in the application?

19)Is it easy and intuitive to understand the option of "services" and "shops" of the Maps functionality?

20)Do you consider that the indications about how to reach a certain shop of the "Maps" functionality are clear?

Fig. 11 and 12 depict the results of the survey.

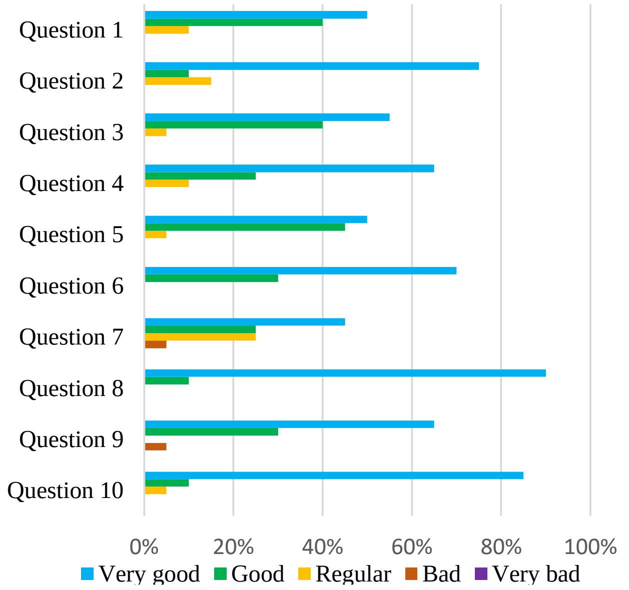


Fig. 11. Results of Questions 1 to 10 of the survey.

In general, the results of our survey show that the answers given by the people who assessed the application were positive; hence, we can say that it complies with the established functional requirements and with the scope defined in the beginning. However, it is worth to mention that in some of the questions asked, some regular responses were obtained. Although it does not represent the majority of the opinions, it is important to consider correcting these aspects of the application as future work, or before the application passes to a production phase, since they are an indication that the application can be improved.

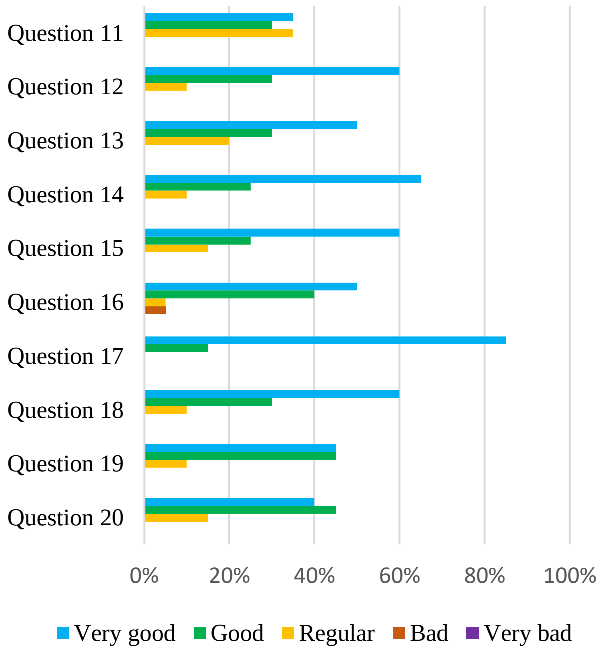


Fig. 12. Results of Questions 11 to 20 of the survey.

B. Geolocation Tests

To test the k-neighbor algorithm as a WiFi geolocation method and for purposes of the present investigation, we decided to use only two levels of maps of the selected shopping center (Level C1 and Level C2), where Level C1 was divided into eight sectors (sector 1 to sector 8), and Level C2 was divided into seven sectors (sector 9 to sector 15). As stated previously, the work was done in a mall of the Venezuela that we will not identify for privacy reasons.

To generate the database used by the algorithm in the fingerprinting process, thirty samples of the signal of each AP were taken in the centroid position of each one of the sectors. Through the C++ program that we developed, we analyzed the collected data and selected the APs that were more stable in the captures for processing and insertion in the database through the web application.

Once the fingerprinting process was completed, several tests were carried out within the shopping center facilities to assess its correct operation. The tests were conducted in two different scenarios. In the first scenario, the shopping center was closed, that is, the corridors were completely free, and there were no people in it. In the second scenario, the shopping center was opened to the public, with many visitors in the corridors. The idea of the second scenario was to evaluate the alteration made by the mobile devices of the numerous visitors over the geolocation service.

For each sector, we did twenty geolocation tests. Five of these tests were performed statically in the approximate position of the centroid of the sector. The remaining fifteen tests were made with a mobile device in movement within the sector. In the testing process, a result was considered correct if the API returned the sector where the user was exactly positioned. In the border of two sections, we considered that the result was correct if the algorithm returned any one of the two adjacent sectors. On the other hand, a result was considered as incorrect if the API returned a sector of another level with respect to the current level of the user, or when the returned sector was in the current level of the user, but did not correspond to its actual sector.

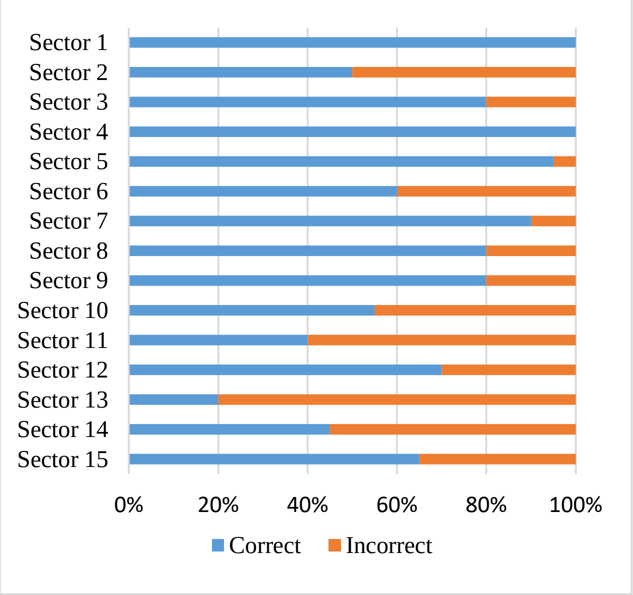


Fig. 13. Results obtained by the geolocation system when the shopping center was closed.

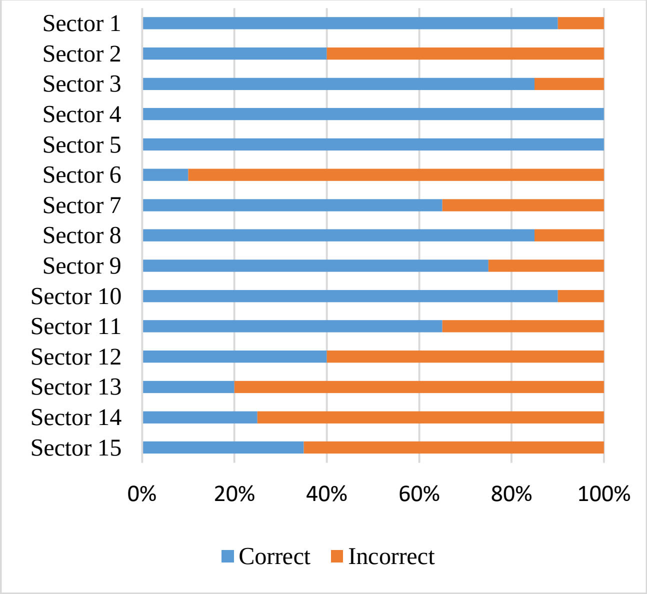


Fig. 14. Results obtained by the geolocation system when the shopping center was open.

Fig. 13 and Fig. 14 show the results of the geolocation tests in the interior spaces of the shopping center when it was closed and opened to the public, respectively.

For the geolocation tests carried out, the general percentage of success was 69.0% when the shopping center was closed, and 61.7% when the shopping center was open. It is important to mention that the selected shopping center did not have its own WiFi network. We worked with the APs installed by the shops, which had different characteristics (different brands, standards, antennas, signal power, etc.). Better results could be achieved if all the considered APs had the same characteristics, and if we knew their exact position. In some APs, it is even possible to change the firmware, and thus modify the power with which the signal is sent according to the needs that are presented by the structure of the building.

XI. Conclusions and Future Work

For companies and organizations, mobile applications represent an added value and a high advance in technologies. In the case of shopping centers, they offer an increase in productivity and greater business opportunities, representing a very important support in the relationship between the mall and its visitors, expanding the network of benefits for both parties.

In the present research, a series of concepts, tools and technologies were presented that served as a solid base for the successful development of a technological solution for shopping centers that allows the diffusion and management of data, such as contact information, directory of shops, promotions, events, services, maps, and a geolocation system based on WiFi. We developed a solution for a shopping center of Venezuela that includes a mobile application for mobile devices with Android OS, a web application implemented using the Laravel framework for data and content management to be consulted by the app, and a private API as a communication mechanism between both applications. It is worth mentioning that the APs that were used in the location service, that belong to the shops within the facilities of the shopping center, are also part of the architecture of the solution. Thanks to these APs, the functionality of geolocation in interior spaces were implemented, by using the algorithm of the nearest k-neighbors. Even if the research was done for a specific mall in Venezuela, it could be reproduced in any shopping center of the world.

In this work, we could verify how certain factors are vital to guarantee a precise geolocation system. Based on this acquired experience, some recommendations can be given to improve the results: (1) pre-establish the location of the APs avoiding obstacles that could interfere their signal, (2) use APs of similar characteristics in terms of power of transmission and subsequently modify these parameters to fine-tune the coverage footprint of each device, (3) perform a monitoring and control of the APs to take action against possible changes in their transmission power, (4) evaluate the use of wireless repeaters in specific cases where the signal suffers attenuation.

As future work, we propose to extend our work so we can have WiFi coverage in the internal parking spaces of the shopping center. Also, we plan to consider other network technologies such as Bluetooth and Beacon devices, and make a comparative study regarding the accuracy of the geolocation system according to the selected technology.

REFERENCES

[1] P. Mulas, A. González and R. Rivera, "Localización de dispositivos móviles en interiores usando redes wireless", Faculty of Computing, Complutense University of Madrid, Madrid, Spain, 2007.

[2] P. Díaz and E. Alvarado, "Desarrollo de soluciones web y móvil para la integración de marketing de proximidad y gestión publicitaria con dispositivos Beacon como tecnología base", School of Computing, Central University of Venezuela, Caracas, Venezuela, 2016.

[3] A. Silberschatz, P. Baer Galvin, and G. Gagne, "Fundamentos de sistemas operativos". Seventh edition, McGraw-Hill, 2005, p. 3.

[4] G. M. Ramírez Villegas, "Seguridad en aplicaciones móviles", Unidad Educativa Nacional Abierta y a Distancia, April 2013, http://datateca.unad.edu.co/contenidos/233016/EXE_SAM/leccin_1_sistemas_operativos_moviles.html.

[5] M. Báez, Á. Borrego, J. Cordero, L. Cruz, M. González, et al. "Introducción a Android", Complutense University of Madrid, E.M.E. Editorial, Madrid, Spain.

[6] A. Chico, "Diseño y desarrollo de un sistema de posicionamiento en interiores basado en WiFi con tecnología Android", University Carlos III of Madrid, Madrid, Spain, December 2009, pp. 59-62.

[7] Developer Android, "Platform architecture", https://developer.android.com/guide/platform/index.html.

[8] Developer Android, "Content providers", https://developer.android.com/guide/topics/providers/content-providers.html.

[9] IBM Corporation, "El desarrollo de aplicaciones móviles nativas, Web o Híbridas", New York, United States of America, April 2012.

[10] P. Rincón, "Aplicaciones móviles nativas con consumo de APIs online, estudio comparado con aplicaciones web móviles en iOS y Android y caso práctico 'Native Client' para Wordpress", University Carlos III of Madrid, Madrid, Spain, July 2012.

[11] M. Piattini and A. Miguel, "Fundamentos y modelos de base de datos". Segunda Edición, Editorial RA-MA, Madrid, Spain, 1999.

[12] A. Richardson, "Automating and testing a REST API: a case study in API testing using: Java, REST assured, Postman, tracks, URL and HTTP proxies, compendium developments Ltd", 1 edition, August 2017.

[13] G. Torres, "Espacios virtuales de experimentación cooperativa. Caso de estudio: Laboratorio virtual de cinemática", Center for Research in Information Technology and Systems, Autonomous University of the State of Hidalgo, Pachuca, Mexico, 2001.

[14] M. El Yaagoubi, "Acceso a internet vía WiFi-WiMax", Department of Electronic Technology, University Carlos III of Madrid, Madrid, Spain, 2012, p.31.

[15] D. Cohen, L. Cohen, G. Faillace, M. Gianatelli, V. Haber et al. "La localización utilizando WiFi (802.11b-g), diferentes algoritmos", Faculty of Engineering, University of Palermo, Buenos Aires, Argentina, 2005, pp. 5-13.

[16] H. Nascimento, E. Rodrigues, F. Cavalcanti, and A. Paiva, "An algorithm based on Bayes inference and k-nearest neighbor for 3D WLAN indoor positioning", XXXIV Brazilian Symposium on Telecommunication, September 2016, Santarem, Brazil.

[17] P. Torteeka and X. Chundi, "Indoor positioning based on WiFi fingerprint technique using Fuzzy k-nearest neighbor", in Proceedings of the 2014 11th International Bhurban Conference on Applied Sciences and Technology (IBCAST 2014), January 2014, Islamabad, Pakistan.

Open-Domain Neural Conversational Agents: The Step Towards Artificial General Intelligence

Sasa Arsovski, Sze Hui Wong, Adrian David Cheok
Imagineering Institute
Iskandar Puteri, Malaysia

***Abstract*—Development of conversational agents started half century ago and since then it has transformed into a technology that is accessible in various aspects in everyday life. This paper presents a survey current state-of-the-art in the open domain neural conversational agent research and future research directions towards Artificial General Intelligence (AGI) creation. In order to create a conversational agent which is able to pass the Turing Test, numerous research efforts are focused on open-domain dialogue system. This paper will present latest research in domain of Neural Network reasoning and logical association, sentiment analysis and real-time learning approaches applied to open domain neural conversational agents. As an effort to provide future research directions, current cutting-edge approaches applied to open domain neural conversational agents, current cutting-edge approaches in rationale generation and the state-of-the-art research directions in alternative training methods will be discussed in this paper.**

Keywords*—*Artificial intelligence; deep learning; neural networks; open domain chatbots; conversational agents

I. Introduction

Artificial General Intelligence (AGI) is achieved when machines are capable to conduct intellectual tasks [1]. Hence it is undeniable that being able to interact with machines in natural language is an important feature of Artificial General Intelligence. Human's ability to communicate and conduct cognitive behavior in natural language play an important role in completing most intellectual tasks. The oldest conversations agents created could be traced back to the success of ELIZA [2] which was first designed with the aim to be a virtual therapist. However due the lack of human knowledge and basic coding, it failed to reach farther than a short conversation. The next conversational agents that followed the creation of ELIZA have improved ELIZA model as a base. Conversational agent *Parry*, also known as “ELIZA with attitude” was designed to simulate a patient with schizophrenia. *Jabberwacky* [3], is the one of the earliest attempts to create a chatbot via human interaction. *Alicebot* [4], an improved version of ELIZA, took the attention of conversational agents research. *Alicebot* uses heuristical pattern matching rules to understand questions and make reply to the human's input. Winning the Loebner Prize, for *Alicebot* was a great success at its time of launch. However, due to its occasional glitch in exposing its mechanistic aspects in short conversations, *Alicebot* was unable to pass the Turing Test. The earliest precursor of the commercial conversational agent in the market would be *SmarterChild* [5], but it is more widely known as the precursor to the various well-known conversational agents in the industry such Apple's Siri or Samsung's S voice.

The first few years of the 21st century witnessed creations of conversational agents which do not only rely on natural language processing. For example, IDM's Watson integrated Machine Learning methodologies along with natural language processing to pull intelligent insights from data. Watson was the only conversational agent with implemented Artificial Intelligence (AI) technologies to gain wide commercial pick up. Apple's Siri incorporates speech recognition to create a better user experience via voice command. Siri's speech recognition uses deep neural network to rank the confidence level of voice inputs and makes a response accordingly. The success of Siri subsequently inspired the creation of other intelligent personal assistants like Microsoft's Cortana [6] and Amazon's Alexa [7]. Despite the success of these chatbots, they are basically goal-oriented dialogue system [8], which sole purpose is to help people solve day-to-day problems using natural language processing. In order to create a conversational agent which is able to pass the Turing Test and move AI research towards full AGI, more research effort towards open-domain dialogue system are presented in [9]. Open-domain conversational agent, Microsoft Tay bot is designed to imitate the way teenagers communicate in America. Tay was created to learn through interactions with human users on Twitter, but it was suspended shortly after subsequent controversy where the bot began to post inflammatory and offensive tweets through Twitter [10].

After almost 50 years since the introduction of chatbots and numerous surveys over the past several decades, chatbot technology is still in the wrap with much more room for improvement. The improvement of computer hardware as well as the use of smartphones, which increased the availability of data, have acted as the catalyst to recent advancement in AI research such as neural conversational agents. The introduction of Sequence to Sequence (Seq2Seq) Long Short Term Memory (LSTM) Neural Network framework [11] has greatly improved Natural Language Processing applications such as neural machine translation [12]. Researchers apply the same methodology to generate neural conversational agents [13].

Seq2Seq is a neural network model that uses one LSTM, a special kind of RNN to read the input sequence encoder, for every time step and the results is a vector. Then it uses another LSTM to extract the output sequence from that vector-decoder. The goal of *LSTM* is to estimate the conditional probability $p(y_1, ..., y_{T'}|x_1, ..., x_T)$, were $(x_1, ..., x_T)$ is an input sequence and $(y_1, ..., y_{T'})$ is its corresponding output sequence whose length T' may differ from T. The *LSTM* computes this conditional probability by first obtaining the

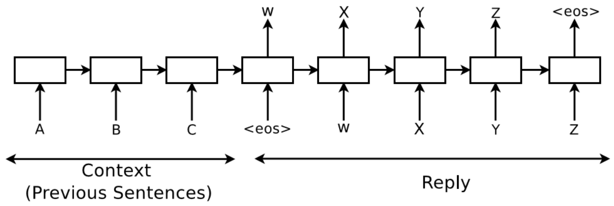


Fig. 1. *seq2seq* model [11].

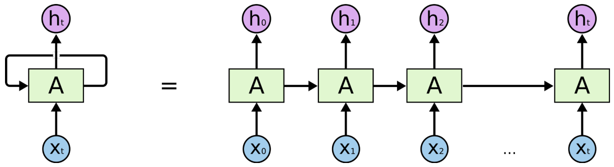


Fig. 2. Recurrent Neural Network (RNN) [14].

fixed dimensional representation v of the input sequence $(x_1, ..., x_T)$ given by the last hidden state of the *LSTM*, and then computing the probability of $y_1, ..., y_{T'}$ with a standard *LSTM* language model formulation whose initial hidden state is set to the representation v of $x_1, ..., x_T$:

$$p(y_1, ..., y_{T'}|x_1, ..., x_T) = \prod_{t=1}^{T'} p(y_t|v, y_1, ..., y_{t-1}) \quad (1)$$

In this equation, each $p(y_t|v, y_1, ..., y_{t-1})$ distribution is represented with a softmax over all the words in the vocabulary. The overall scheme is outlined in Figure 1, where the shown *LSTM* computes the representation of $A, B, C, < EOS >$ (sequences of the input data) and then uses this representation to compute the probability of $W, X, Y, Z, < EOS >$ (sequences of the output data) as shown in Fig. 1. Numerous of neural conversational agent based on [13] have been developed and unprecedented results have been observed.

The focus of this paper will be on open-domain, deep neural network based conversational agents. Success of Seq2Seq model applied to Neural Machine Translation and neural conversational agent [13] initiated the state-of-the-art research in the dialogue generation. This review paper presents the current state of the art open-domain generative neural conversational agents based on Seq2Seq model. Future directions in neural conversational agent research towards creation of a full AGI will also be discussed in this paper.

This paper is organized as follows: analysis of open-domain neural conversational agents is presented in the first section, neural conversational agent with reasoning and logical association is presented in the second section, sentiment analysis and reasoning networks in conversational agents are presented in the third section, neural conversational agent with real-time learning component are analyzed in the fourth section, the fifth section presents the discussion and future work.

II. Current State of the Art Neural Conversational Agents

Progress in Text and Natural Language Processing has always been a crucial component in the pursue of Artificial General Intelligence. Despite the groundbreaking achievement in Neural Machine Translation [12], researchers are still a distance away from achieving conversational agents that are able to conduct open-ended conversations that are indistinguishable from a human. Most open-ended neural network conversational agents research utilize Sequence to Sequence model (Seq2Seq) [11] which uses a multi-layered Long Short-Term Memory (LSTM) to map the input sequence to a vector of a fixed dimensional, and then another deep LSTM to decode the target sequence from the vector hence enabling conversational models to be trained end-to-end, thus decreasing the number of hand-crafted rules [13]. This method as proposed in [13] has enabled generation of simple conversations when is trained on a large conversational dataset. This method has not only successfully extracted knowledge from a domain specific dataset but has also demonstrated simple form of common sense reasoning when trained on a large and general domain dataset of movie subtitles [15]. The success of this research inspired many subsequent open-domain conversational agent research with the model proposed by [13] as the baseline model.

Despite its success as a dialogue generation and machine translation model, one significant weakness of RNN (Fig. 2.) based architectures such as Seq2Seq is its tendency to generate generic, non-meaningful and non-diverse responses due to its maximum likelihood estimation mechanism in dialogue generation [16]. Maximum likelihood mechanism is the procedure of finding the parameters value of given statistic which maximise likelihood distribution [17]. Another weakness would be its tendency to be caught in an infinite repetitive loop [18]. Hence, there is numerous research extensions aimed to overcome the major shortcomings of RNN based architectures.

A. Latent Variable Hierarchical Recurrent Encoder-Decoder (VHRED)

VHRED uses a hierarchical generation process in order to exploit the with-in sequence structure in utterances and is trained using a variational lower bound on the log-likelihood. The proposed method generates a response as a sample of the continuous latent variable which is a representation of the high-level semantic content of the response. The model then generates the response word by word conditioned on the latent variable [19]. The study has demonstrated that humans prefer generic responses generated by LSTM for short contexts while for long contexts humans prefer the semantically richer responses generated by the proposed model. The results of VHRED model are shown in Table I.

B. Diversity Promoting Model

Diversity promoting model proposed Maximum Mutual Information (MMI) approach [20] as the objective function in neural network models. MMI approach replace traditional objective function such as the likelihood of output (response) to given input(message). The research in [20] demonstrated that the proposed MMI models produce more diverse, interesting, and appropriate responses, yielding substantive gains in BLEU scores on two conversational datasets and in human evaluations [21]. The BLEU score algorithm compares the N-grams of two text fragments and counts the number of matches, the similarity score of these texts is a function of the number of matches [22].

TABLE I. TWITTER EXAMPLES FOR THE NEURAL NETWORK MODELS. THE TOKEN INDICATES A CHANGE OF TURN [19]

Context	Response
jamaica blood!! ¡url¿ lol. yes [name]!. [name]! what's poppin in toronto homie	**VHRED**: nothing much. just chillin. what's good with you? **LSTM**: something!!!!!!! **HRED**: something!!!!!!
reaches for [name]! - approaches him warily - hey there buddy. - swings him into the air - how you doin? tilts his head and your wariness and reaches for your face [name].	**VHRED**: - looks up at you and smiles - you're so cute. - kisses your cheek - **LSTM**: - smiles - i'm fine thank you. how are you? **HRED**: - grins, looking at you as i look at you -
my math teacher over here talkn bout nickels nd i'm tryna make big money i need to be in calculus lol	**VHRED**: i need to go to the gym and get my hair done lol **LSTM**: lol **HRED**: how u been
i think i want an ipad2. but i already have an ibook. want vs. need. if you have an ibook. then you deserve an ipad!	**VHRED**: that's what i'm thinking. i'm going to buy a new one. **LSTM**: i want one!!!! **HRED**: i don't know what to do with it. i don't know what to do
ah echt? dan anrtwoord ik volgende keer via twitter als je me smst. haha doe maar p. goed bezig hoor myrthe d goed bezig? die snap ik even niet p	**VHRED**: haha, ik heb er geen zin in de something, maar ik heb er geen something meer something **LSTM**: haha something something something **HRED**: something something something something something something something something

C. Deep Reinforcement Learning for Dialogue Generation

Researchers in [18] proposed neural network model that is able to generate coherent and interesting dialogues by applying deep reinforcement learning to model future reward in chatbot dialogue. Model is able to recognize future direction of conversation. Presented research shows that the proposed algorithm generates more interactive responses and manages to foster a more sustained conversation in dialogue simulation. Hence marking the first step towards learning a neural conversational model based on the long-term success of dialogues. The examples of generated conversation by RN and MMI model [18] are given in Table II.

As mentioned above, the maximum likelihood estimation technique utilized in RNN based architectures such as Seq2Seq often end up generating redundant and meaningless responses such as "I don't know" or "Maybe" due to the high frequencies of these generic responses in the training datasets. Technique, known as "mutual information" generate a list of possible responses based on an input source, known as the mutual information score from a pre-trained Seq2Seq model. This mutual information score will then be used as a reward and back-propagated to the encoder-decoder model, tailoring it to generate sequences with higher rewards using reinforcement learning technique. This method has demonstrated its success in suppressing undesirable generic responses and promoted diversity in responses. Longer sentences are also observed as a result of this proposed model, bridging state-of-the-art dialogue generation research a step closer towards creating conversational agents that are indistinguishable from a human speaker.

III. NEURAL CONVERSATIONAL AGENTS WITH REASONING AND LOGICAL ASSOCIATION COMPONENTS

Despite the progress in generating conversations using deep neural network models, most of these work produce little or no ability to form clear and human like reasoning, hence generating responses that are less human like, adhering human users from conducting elaborated and continuous meaningful conversations. Research presented by [21] and [18] might have improved dialogue generation model presented in [13] but they are still unable to produce conversations with elaborated logical associations and reasoning ability. The work presented in [19] has significantly improved the quality of conversations by producing semantically richer responses especially conversations in longer contexts. Despite its success in answer generation based on the context of the input, the model is still lack to conduct continuous conversations with reasoning and association ability.

A. Sentiment Analysis and Reasoning Networks in Neural Conversational Agents

Progress in sentiment analysis and reasoning ability of a conversational agent is crucial towards creating more human-like conversations. However, it is important for the agent to be able to understand the underlying basis of the input, which could be a question, statement with differing sentiment or objectives in order to perform accurate reasoning and logical association. Therefore, a good sentiment analysis component in the conversational agent model plays an important role to compensate the lack of transparencies of most complex neural models in natural language processing (NLP). The ideal complex neural conversational model should yield improved performances and offer interpretable rationales for answer predictions. Current cutting-edge approach presented in [23], incorporate rationale generation as an integral part of the learning problem. This approach limit the models to extractive rationales by limiting the rationales to be subsets of words from input text that are short and coherent or must alone suffice for prediction as a substitute of the original text. Rationale generation must be learned in an unsupervised manner, hence a model with rationales is trained on the same data as the original neural models. The model demonstrated in [23], trained in an end-to-end manner, gives rise to quality rationales in absence of any explicit rationale annotations. To minimize the ambiguity of what counts as rationales, as well as difficulty in evaluation of rationale selections, the model focuses on the two following domains: (i) concerns with multi-aspect sentiment analysis, (ii) concerns with the problem of retrieving related questions. The model obtained high performance on both tasks, and, for the sentiment prediction task, the model achieved an accuracy of 96%, a significantly higher accuracy compared to bigram support vector machine (SVM) [24] and neural attention baseline. The approach was evaluated on multi-aspect sentiment analysis against a manually annotated test case, and it outperformed attention-based baseline by a significant

TABLE II. SAMPLED RESPONSES GENERATED FROM THE MUTUAL INFORMATION MODELS AND THE RL MODEL [18]

Input Message	Mutual Information Model	RL
How old are you?	I'm 16.	I'm 16, why are you asking?
What's your full name?	I have no idea.	What's yours?
I don't want to go home tonight.	Really?	Why?
Do you have any feelings for me?	I don't know what you are talking about.	I don't want to hurt your feelings.
How much time do you have here?	Not long enough. Sorry, sir	Ten seconds
Shall we get started?	Of course!	Yes. We've got a lot of work to do here.
Do you play football?	No, I don't	Yes. I love football!
We'd have to talk to him	I mean, he's a good guy	About what?
How come you never say it?	Because I don't want to hurt you.	I don't think it's a good idea to say it

margin.

Another approach to create conversational agent that is able to perform logical associations and reasoning is by incorporating Reasoning Networks (RN) in neural conversational agent model. An example of RN is demonstrated by authors in [25], achieved state of the art performance on a challenging visual Question-Answer (QA) dataset CLEVR. This is a good example where additional intelligence emerges from the composition of simple modules. Relational reasoning is a principal component of generally intelligent behavior but has proven difficult for neural networks to learn. In the work [25], authors describe how to use Relation Networks (RNs) as a simple plug-and-play module to solve problems that fundamentally hinge on relational reasoning. Authors stated that RNs deep learning architecture can discover and learn how to reason about entities and their relations. Reasoning NNs are a potential venue for a breakthrough towards AI that is now hopeful to be found. Reasoning Networks can be used for a variety of problems that can benefit from structure learning and exploitation, such as rich scene understanding in Reinforcement Learning (RL) agents, modelling social networks, and abstract problem-solving.

IV. JOINT MODEL BASED TRAINING TO IMPROVE PERFORMANCE OF NEURAL CONVERSATIONAL AGENTS

In order to improve question answering systems researchers analyze how communication between humans rely on both, question asking as well as question answering and the intrinsic connections between the two tasks [26]. We will present a few important incentives towards question generation as well. Questions are usually asked to access knowledge of others or direct one's own information-seeking behavior. According to authors in [27], the incentives to teach machines to ask questions are:

- improving the acquisitions of information in intelligent systems by asking appropriate questions based on the situation;
- improving the ability of machines to answer questions by teaching machines how to ask questions;
- create a system that is able to solve abstractive tasks such as question asking instead of extractive tasks such as question answering;
- providing practical solution on many possible applications on questions asking mechanisms about documents.

Current state of the art question answering systems based on deep neural network rely on the two main information retrieval (IR) models; generative model and discriminative model. While generative models are theoretically sound and successful in modelling features, they suffer from difficulty in leveraging relevancy of the signals from other channels like links and clicks, whereas discriminative models lack a principled way of obtaining useful features or gathering helpful signals from massive unlabeled data. Efforts have been invested in solving these shortcomings via information retrieval generative adversarial networks (IRGAN) model. IRGAN model takes advantage and characteristics of both models where the generative model acts as an attacker to the current discrimination model, generating difficult examples for the discriminative model in an adversarial way by minimizing its discrimination objective. Authors in [28], achieve state of the art results using a IRGAN model.

Several research have utilized question generation as a tool to improve the efficacy of question answering models [29], [30], [26]. Authors in [30] simultaneously train the model by alternating input data between question answering (QA) and question generation (QG), both in the same model. The trained model is then used to generate questions and answers with the hypothesis that a good question generation helps models to improve QA performance. Authors in [26] incorporates the probabilistic correlation of QA and QG by leveraging the correlation to guide the training process of both models. Authors in [26] randomly initialize the parameters in both QA and QG models with a combination of fan-in and fan-out. Parameters of words of word embedding matrices are shared in QA and QG model. However, both QA and QG models use two different embedding matrices for question words and answer words in order to learn question and answer specific word meanings. This training framework successfully shown that by exploiting the "duality" of QA and QG improves both QA and QG. Another joint approach for QA and QG is presented in [29]. Authors uses approach presented in [30] to prove hypothesis that good question generation can improve QA performance. Authors in [29] leverages convolutional neural network (CNN) and Recurrent Neural Network (RNN) for question generation in order to cover both question generation approaches; retrieval based and generation based. Authors demonstrated that the question generation method successfully improve existing question answering systems. Both results for [30] and [26] are shown in Tables 3 and 4, respectively.

The methodology proposed by [30] might have not achieve state of the art results achieved by selected QA models, it nonetheless demonstrated the effectiveness of joint training between QA and QG, especially in efficacy of abstractive tasks like QG. Whereas [26] demonstrated the "duality" of both QG and QA models are able to improve both tasks, authors in [29] manage to significantly improve QA using a good question generation approach and feed QA system

TABLE III. EXAMPLES OF QA BEHAVIOUR CHANGES POSSIBLY INDUCED BY JOINT TRAINING. GOLD ANSWERS CORRESPOND TO TEXT SPANS IN GREEN. IN BOTH THE POSITIVE AND THE NEGATIVE CASES, THE ANSWERS PRODUCED BY THE JOINT MODEL ARE HIGHLY RELATED (AND THUS PRESUMABLY INFLUENCED) BY THE GENERATED QUESTIONS. [30]

Positive	Document	*in the 1960 election to choose his successor, eisenhower endorsed his own vice president , republican richard nixon against democrat john f. kennedy.*
	Q gold	*who did eisenhower endorse for president in 1960 ?*
	Q gen	*what was the name of eisenhower own vice president ?*
	Answer	A-gen: *john f. kennedy* JointQA: *richard nixon*
Negative	Document	*in 1870 , tesla moved to karlovac , to attend school at the higher real gymnasium , where he was profoundly influenced by a math teacher martin sekulic*
	Q gold	*why did tesla go to karlovac ?*
	Q gen	*what did tesla do at the higher real gymnasium ?*
	Answer	A-gen: *to attend school at the higher real gymnasium* JointQA: *he was profoundly influenced by a math teacher martin sekulic*

TABLE IV. SAMPLED EXAMPLES FROM THE SQUAD DATASET [26]

question	**correct answer**	**question generated by Dual QG**	**question generated by Basic GQ**
what's the name of the green space north of the center of newcastle ?	*Another green space in Newcastle is the Town Moor , lying immediately north of the city centre.*	*what is the name of the green building in the city ?*	*what is the name of the city of new haven ?*
for what purpose do organisms make peroxide and superoxide ?	*Parts of the immune system of higher organisms create peroxide, superoxide, and singlet oxygen to destroy invading microbes.*	*what is the purpose of the immune system?*	*what is the main function of the immune system ?*
how much money was spent on other festivities in the bay area to help celebrate the coming super bowl 50?	*In addition, there are 2 million worth of other ancillary events, including a week - long event at the Santa Clara Convention Center, a beer, wine and food festival at Bellomy Field at Santa Clara University, and a pep rally.*	*how much of the beer is in the santa monica convention center?*	*what is the name of the beer in the santa monica center ?*

with generated questions. These state of the art research paves an optimistic direction for an alternative training method to further improve the current state of the art open domain neural conversational agents.

V. DISCUSSION AND FUTURE WORK

Current state-of-the-art open-domain neural conversational agents have largely utilized Seq2Seq architecture along with several optimization methodologies which have produced unprecedented results. Most of these results have not been achieved by other open-domain dialogue generation models. Despite the accomplishments of these state of the art research, these conversational agents are lacking in several aspects which hinder these agents from creating conversational experiences that are indistinguishable from humans, which is the key in creating a system that is able to pass the Turing Test. Aspects that are lacking in current state of the art conversational agents are:

- Agents that are able to learn new information in real time and “remember” information in future conversation sessions.
- Agents that are able to conduct vigorous reasoning and logical associations.
- Agents that are able to select appropriate emotion in responses.
- Agents with personalities and individuality.
- Agents that are able to respond with appropriate timing.
- Multi-modal agents that are able to connect computer vision (sense of sight), audio processing (sense of hearing) as an input for a dialog generation.

In order to achieve full Artificial General Intelligence (AGI), a combination of approaches need to be considered for future research. Most likely an multi-modal solution will be the direction of future research in creating conversational agents closest to full AGI, such is presented in [31]. This research has shown that a single multi-modal deep learning model is able to jointly learn a number of large-scale tasks from multiple domains, allowing as many parameters as possible to be shared. This allows transfer learning from tasks with large amount of data to the ones that are limited with data, hence benefiting tasks with less data by joint training.

While there are several approaches which tackle each of above listed problems, full AGI will most likely be achieved with the combination of different approaches with different aims in a single multi-modal architecture. There are a few researches that specifically address above mentioned problems, which has the potential in being part of a multi-modal model that will be able to create the most human-like conversational agent.

Authors in [32] presents an conversational agent with unforgettable characters who exhibit various salient emotions in conversations. The research has achieved the goal by focusing on humanizing artificial character of conversational agents. Often times, despite how human-like responses are able to be generated by conversational agents, the neutral mono-personality answers at certain situations which are inappropriate due to the lack of the emotional consideration generated easily gave away non-human like traits.

Authors in [23] utilize encoder-generator framework which is trained in an end-to-end manner to rise rationales quality in absence of explicit rationale annotations. The results from this research could be used to build a conversational agent that is able to reason and rationalize without supervision.

A. Train an Agent Based on Pre-defined Personality

Another approach would be personality and individual goal based conversational agents. Current approach rely largely on available data-sets composed of information from multiple personalities, mood and goals. Hence most conversational agents trained this way produce an amalgamation of all possible responses given one input. Despite plenty of implementations aimed to improve this aspect, Seq2Seq based architectures still generating response with an assumption that there is only one correct response for each input, when in fact, is not the case for open-domain conversations. In open-domain conversations, many temporal or permanent factors influence the outcome of response. Temporal factors are such as the agent's mood, temporal interests, goals and desires due to external exposure and experience. Whereas permanent factors are: agent's basic personality and temperament, intelligence maturity such as the agent's knowledge base and reasoning and association ability. Personality based conversational agents are trained with pre-defined temporal conditioning such as personality, goals and desires. Hence the agent will be able to learn interests, personality and desire, hence reducing the high dimensional space of context variables which in turn alleviating the curse of dimensionality by lowering latent semantic space.

B. Future Direction of Research in Neural Conversational Agents with Reasoning Component

The approach proposed by [25] could be incorporated into future dialogue generation models combined with sentiment analysis and attention mechanisms (to filter unimportant relations, thus bound the otherwise quadratic complexity of the number of considered pairwise relation), to create a truly powerful open-ended conversational agent which truly understand input sources at each sentence input level as well as the sentiment of the entire conversation as a whole. Connecting the conversational agent that is able to perform robust sentiment analysis and reasoning to a real-time scalable ontology could be an important direction in the field of research in open-ended dialogue generation models.

C. Neural Conversational Agent with Real-Time Learning Component

Modelling the concept of a malleable knowledge database, conversational agent will able to learn from conversations. With the ability to learn from conversation, neural conversational agents will able to conduct more human-to-human like conversations and improve logical associations and reasoning ability. Implementation of an ontology to an conversational agent architecture will enable logical associations, reasoning and knowledge saving. These knowledge can be saved as data and updated in real-time. Conversational Agents with Real-Time Learning component model will be able to conduct more realistic human like conversations.

VI. Conclusion

This review paper explored current state-of-the-art in open domain neural conversational agent, as an effort in evaluating current stage in research towards achieving Artificial General Intelligence. Ever since the success of Seq2Seq model in creating state of the art generative neural conversational agent, numerous of research efforts has been invested in this direction with the aim of improving open domain neural conversational agents. Despite the great leap in breakthroughs of research in this area, current cutting edge open domain neural conversational agents are still far from being indistinguishable from a human speaker. In this paper latest research in the field of Neural Network reasoning and logical association, sentiment analysis and real-time learning approaches applied to open domain neural conversational agents are presented. Future research directions in the domain of neural conversational agents and present state of the art research directions in alternative training methods are also explored at the end of this paper as an effort to discuss the directions needed to be addressed in order to move open domain neural conversational agent research one step closer towards full AGI.

References

[1] B. Goertzel and P. Wang, "A foundational architecture for artificial general intelligence," *Advances in Artificial General Intelligence: Concepts, Architectures and Algorithms*, vol. 6, p. 36, 2007.

[2] o. Weizenbaum, "Eliza-a computer program for the study of natural language communication between man and machine," *Communications of the ACM*, vol. 9, no. 1, pp. 36–45, 1966.

[3] R. Carpenter and J. Freeman, "Computing machinery and the individual: the personal turing test," *Computing, Accessed September*, vol. 22, p. 2009, 2005.

[4] R. S. Wallace, "The anatomy of alice," in *Parsing the Turing Test*. Springer, 2009, pp. 181–210.

[5] R. Khan and A. Das, "Introduction to chatbots," in *Build Better Chatbots*. Springer, 2018, pp. 1–11.

[6] L. Chris, "Why cortana assistant can help microsoft in the smartphone market. the street," 2014.

[7] R. Crist, "Amazon alexa: Device compatibility, how-tos and much more," *CNET. com*, 2016.

[8] K. Crockett, O. James, and Z. Bandar, "Goal orientated conversational agents: applications to benefit society," in *KES International Symposium on Agent and Multi-Agent Systems: Technologies and Applications*. Springer, 2011, pp. 16–25.

[9] R. Higashinaka, K. Imamura, T. Meguro, C. Miyazaki, N. Kobayashi, H. Sugiyama, T. Hirano, T. Makino, and Y. Matsuo, "Towards an open-domain conversational system fully based on natural language processing," in *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, 2014, pp. 928–939.

[10] J. Wakefield, "Microsoft chatbot is taught to swear on twitter," Mar 2016. [Online]. Available: http://www.bbc.com/news/technology-35890188

[11] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in neural information processing systems*, 2014, pp. 3104–3112.

[12] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey *et al.*, "Google's neural machine translation system: Bridging the gap between human and machine translation," *arXiv preprint arXiv:1609.08144*, 2016.

[13] O. Vinyals and Q. Le, "A neural conversational model," *arXiv preprint arXiv:1506.05869*, 2015.

[14] "Understanding lstm networks." [Online]. Available: http://colah.github.io/posts/2015-08-Understanding-LSTMs/

[15] C. Danescu-Niculescu-Mizil and L. Lee, "Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs." in *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, ACL 2011*, 2011.

[16] I. V. Serban, A. Sordoni, Y. Bengio, A. C. Courville, and J. Pineau, "Building end-to-end dialogue systems using generative hierarchical neural network models." in *AAAI*, vol. 16, 2016, pp. 3776–3784.

[17] J. W. Harris and H. Stöcker, *Handbook of mathematics and computational science*. Springer Science & Business Media, 1998.

[18] J. Li, W. Monroe, A. Ritter, M. Galley, J. Gao, and D. Jurafsky, "Deep reinforcement learning for dialogue generation," *arXiv preprint arXiv:1606.01541*, 2016.

[19] I. V. Serban, A. Sordoni, R. Lowe, L. Charlin, J. Pineau, A. C. Courville, and Y. Bengio, "A hierarchical latent variable encoder-decoder model for generating dialogues." in *AAAI*, 2017, pp. 3295–3301.

[20] L. Bahl, P. Brown, P. De Souza, and R. Mercer, "Maximum mutual information estimation of hidden markov model parameters for speech recognition," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'86.*, vol. 11. IEEE, 1986, pp. 49–52.

[21] J. Li, M. Galley, C. Brockett, J. Gao, and B. Dolan, "A diversity-promoting objective function for neural conversation models," *arXiv preprint arXiv:1510.03055*, 2015.

[22] P. Koehn, J. Martin, R. Mihalcea, C. Monz, and T. Pedersen, "Proceedings of the acl workshop on building and using parallel texts," in *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, 2005.

[23] T. Lei, R. Barzilay, and T. Jaakkola, "Rationalizing neural predictions," *arXiv preprint arXiv:1606.04155*, 2016.

[24] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proceedings of the fifth annual workshop on Computational learning theory*. ACM, 1992, pp. 144–152.

[25] A. Santoro, D. Raposo, D. G. Barrett, M. Malinowski, R. Pascanu, P. Battaglia, and T. Lillicrap, "A simple neural network module for relational reasoning," in *Advances in neural information processing systems*, 2017, pp. 4974–4983.

[26] D. Tang, N. Duan, T. Qin, and M. Zhou, "Question answering and question generation as dual tasks," *arXiv preprint arXiv:1706.02027*, 2017.

[27] X. Yuan, T. Wang, C. Gulcehre, A. Sordoni, P. Bachman, S. Subramanian, S. Zhang, and A. Trischler, "Machine comprehension by text-to-text neural question generation," *arXiv preprint arXiv:1705.02012*, 2017.

[28] J. Wang, L. Yu, W. Zhang, Y. Gong, Y. Xu, B. Wang, P. Zhang, and D. Zhang, "Irgan: A minimax game for unifying generative and discriminative information retrieval models," in *Proceedings of the 40th International ACM SIGIR conference on Research and Development in Information Retrieval*. ACM, 2017, pp. 515–524.

[29] N. Duan, D. Tang, P. Chen, and M. Zhou, "Question generation for question answering," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 866–874.

[30] T. Wang, X. Yuan, and A. Trischler, "A joint model for question answering and question generation," *arXiv preprint arXiv:1706.01450*, 2017.

[31] L. Kaiser, A. N. Gomez, N. Shazeer, A. Vaswani, N. Parmar, L. Jones, and J. Uszkoreit, "One model to learn them all," *arXiv preprint arXiv:1706.05137*, 2017.

[32] J.-C. Heudin, "Emotion selection in a multi-personality conversational agent." in *ICAART (2)*, 2017, pp. 34–41.

Formal Analysis and Verification of Agent-Oriented Supply-Chain Management

Muhammad Zubair Shoukat, Muhammad Atif,
Imran Riaz Hasrat
Department of Computer Science
and Information Technology
The University of Lahore
Lahore, Pakistan

Nadia Mushtaq
Department of Software Engineering
The University of Lahore
Lahore, Pakistan

Ijaz Ahmed
Department of Computer and
Information Sciences DWC
Higher College of Technology
Dubai, United Arabe of Emirates

***Abstract*—Managing various relationships among the supply chain processes is known as Supply Chain Management (SCM). SCM is the oversight of finance, information and material as they move in the flow from different suppliers to manufacturer, wholesaler, retailer and customers. The main problem with such software architecture is coordination and reliability while performing activities. Moreover, continuously changing market makes this coordination challenging. For example failure of production facilities, irregularities in meeting deadlines, unavailability of workers at required times. However, in the Agent-Oriented Supply-Chain Management described in [Mark S. Fox, Mihai Barbuceanu, and Rune Teigen “Agent-Oriented Supply-Chain Management”. The International Journal of Flexible Manufacturing Systems, 12 (2000)] the proposed solution claims a remarkable coordination on the basis of an agent-oriented software architecture. In this paper, we formally specify architecture and verify it using model checking. We use UPPAAL to formally specify the agents’ behaviour involved in SCM. By model-checking, we prove that the given SCM’s architecture partially fulfills its functional requirements.**

***Keywords*—Supply chain management; agent-oriented supply-chain; model checking; formal specification and verification**

I. Introduction

Supply chain is a system of organizing activities, people, resources and information involved during the movement of raw material or finished goods from supplier to customer. Managing various relationships among the supply chain processes is known as supply chain management (SCM). Supply chain management (SCM) is the oversight of finance, information and material as they move in the flow from different suppliers to manufacturer, wholesaler, retailer and customers. Supply chain management (SCM) software architecture maintains coordination among and within companies. The main problem with such software architecture is the coordination and reliability while performing activities but the drastically changing market makes the coordination complex.

Supply Chain is not a chain of businesses rather it is a relationship of multiple businesses [1]. It represents a new way of managing the relationship and associated businesses. So, there is need to build standardized methods to put supply chain management (SCM) in practice.

Nirupam Julka et al. in [2] propose a unified framework for monitoring, modeling and management of supply chain. The proposed framework implements various activities of supply chain like production process, enterprise, business knowledge and data. It presents all the activities as an intelligent and unified function. Various software agents are used to compete activities. This framework helps to evaluate and analyze the different business behaviours according to different circumstances faced in supply chain management.

In [3], certain issues regarding agent-oriented supply chain management are investigated and for those issues respective solutions are presented. It is claimed that the proposed solution can handle the complex tasks and interruptions caused by some unexpected events. Our target in this paper is to study the proposed solution for formal specification and verification Formal verification offers a large potential to provide correctness measuring techniques [4]. We apply model checking as formal analysis by using a tool-set UPPAAL.

During the past few years, many automatic verifications and modeling tools for real-time and hybrid structures [5], [6], [7], [8] and [9] have been developed.

The main contribution of this research is a formally described Agent based supply chain management system given in [3] with a set of formal and informal requirements. We prove that the given construction of agent-oriented architecture doesn’t meet certain functional requirements. The results are given in the form of message sequence charts.

Structure of the Paper: In Section II, we describe the behaviour of agents which are participants of the agent-orients supply chain management. The behaviour of these participants is formally specified and explained in Section III. Functional requirements are described in Section IV which are specified as formulas in Section IV-A. Results of model-checking are there in Section V. Section VI provides limitations used to develop formal models and we conclude this paper in Section VII.

II. Agent-Oriented Supply Chain Management System

In Fig. 1, the basic architecture is shown which tells about the customer conversation with the logistics agent. The process starts when a customer agent place an order, the logistics agent receives the proposal and acknowledges the customer about the received order. Logistics receives the order and

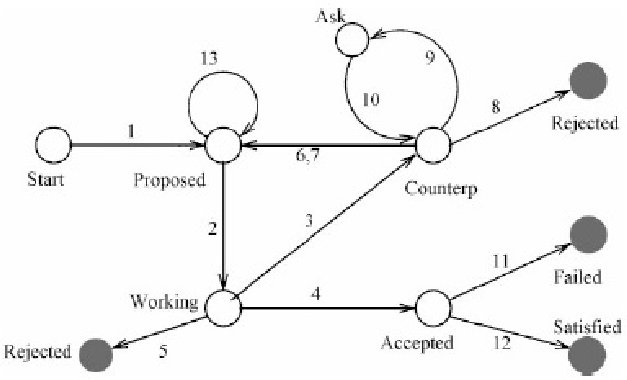


Fig. 1. Customer conversation [3].

tries to decompose the order into different activities if the order decomposed the next process starts if decomposition not possible the process ends. In case of decomposition the next process will be the ranking of contractors on the bases of the activities that are formed after decomposing the order. This process contains two steps: one is formation of large teams and the other is formation of small teams.

In large teams all the contractors that are interested in performing the activities are involved each activity have at least one contractor. After this small teams formed in which one contractor assigned to the one activity. In this stage small team should work cooperatively and inform to the logistics if they have any problem that make it impossible for them to complete their task or activities assigned to them. In case of decomposition is failed, the logistics acknowledges negatively so that the customer may change the proposal. The possible changes can be on the basis of requirements, time or contractors' availability. If logistics again cannot handle a modified proposal then it goes to rejected state.

After the work completion Logistics hand over the work to the customer and customer state is accepted then. If the customer is satisfied then feedback is provided by going on satisfied state. In this way, multiple agents can be considered while placing an order in the agent based supply chain management system.

A. *Customer and Logistics Agents*

- Customer Agent: The customer agent sends proposal to the logistics agent and goes to working state. After the processing a proposal the logistic agents acknowledge to the customer agent. The customers may go to a rejected state or can ask for counter proposal if the order cannot be decomposed. If customer satisfied then it goes to satisfied state otherwise on failed state.
- Logistics Agent: The logistics agent receives the proposal and works on it. Logistics agent also informs the customer that working has been initiated on the proposal. Logistics agent decomposes the order or proposal, rank the contractors and also creates the teams that are able to perform the activities according to customer need. Logistics agent also negotiates with customer if there is delays in work or if the decomposition not possible logistics negotiate with different proposal.

III. Formal Specifications

Our formal specification in UPPAAL covers the following participants or processes, i.e. the *Customer*, the *Logistics* and the *Small Team*. The main process is the *Logistics* process. The logistics process receives and sends messages to the other processes to communicate with them. The customer can sends order to any logistics process using handshaking channels. After reception of order the logistics decomposes the order and sends it to the small team. The small team communicates with logistics process and committed to complete that task after checking its schedule. We gives a brief description of the formal specifications model checking of main processes in our explanation of the architecture.

We specify all the concurrent processes of Agent-Oriented Supply-Chain Management. The *Customer*, *Logistics* and *Small Team* are the processes or participants in the given model. These processes of the software architecture are modeled as parallel processes.

A. *Channels*

This software architecture uses thirteen channels. To model the functionality of Agent-Oriented Supply-Chain Management the following channels are used for one-to-one communication or for broadcasting:

1) *Proposal:* This channel is used by a customer to send some proposal to logistic agents.
2) *Order:* If proposal is accepted then this channel is used place order for selected items.
3) *Reject;* This channel is used to acknowledges a customer if some order can be processed or not.
4) *Success:* A completion of order is conveyed through this channel.
5) *Failed:* A team uses this channel if some order cannot be completed with certain time.
6) *Complete:* If a task is completed successfully then an acknowledgment is sent by a small team through this channel.
7) *NegT1:* This channel is used to send task to the small team 1 by the logistics.
8) *NegT2:* This channel is used to send task to the small team 2 by the logistics.
9) *Commited:* This channel is used to send acknowledgment to the logistics by the small team if the team is interested and willing to work.
10) *Alternative:* This channel is used to send acknowledgment to the logistics by the small team if the team has some issues in the proposal and needs alternative which is received from the logistics.
11) *NewT1:* This channel is used to send task to the small team 1 after the new contractor is assigned to the existing task.
12) *NewT2:* This channel is used to send task to the small team 2 after the new contractor is assigned to the existing task.
13) *Change:* This channel is used to send acknowledgment to the small team if the new contractor is not available for the existing task.

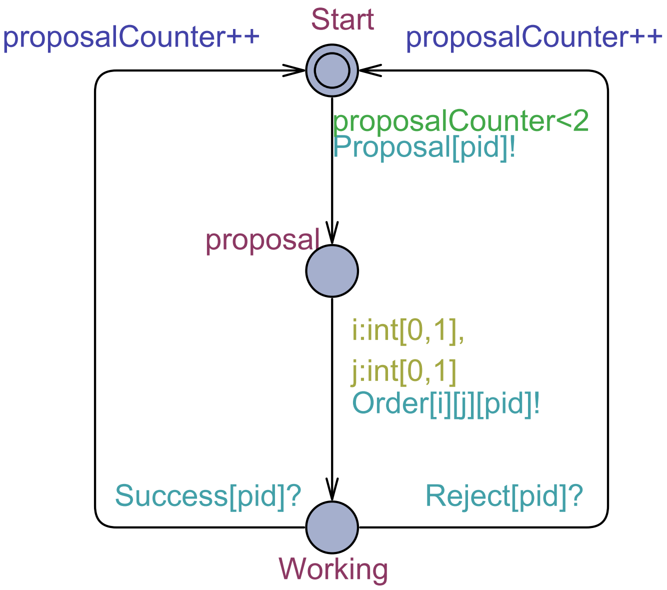


Fig. 2. The customer process.

B. Behaviour of a Customer Process

In Fig. 2, the automaton for *Customer* process is illustrated. The initial state is named as *Start*. The *Customer* process has three states. The first state is *Start* state, second is *Proposal* state and the third state is *Working* state. There are four major actions in this process described below:

1) Sending proposal to the *Logistics*.
2) Sending Order to the *Logistics* and waiting for the response.
3) Going to the start state through rejected path if response is negative from *Logistics*.
4) Going to the start state through success path if order is successfully completed.

First of all, the channel *Proposal* transfers a value to some logistic agent which is originated by a customer process. The logistic agents receives these values while synchronization of channel *Proposal[x]*. Here ‘x’ represents process ID (pid), i.e., the customer ID sending a proposal.

Secondly, after sending proposal the customer sends order using channel *Order[2][2]* which is received by the logistics at channel *Order[i][j][cus_id]* and goes to the *Working* state. There are two values ’i’ and ’j’ that are sent by the *Customer* for the activities that a customer needs. If *i=0* and *j=1* the *Customer* needs activity *j*, if *i=1* and *j=0* the customer needs activity *i* and if both *i* and *j* are 1 the customer needs both the activities.

At the end if the customer receives error message from the logistics that the order cannot be processed or teams fail to work then it goes to *Start* state using *Reject* channel, if the work is successfully done it goes to *Start* state using *Success* channel. On initial state means that it is ready for the next proposal. There is a counter *proposalCounter* for the proposals sent by a *Customer*.

C. Automaton for the Logistics Process

In Fig. 3 the automaton for *Logistics* process is illustrated. The initial state is named as *Start*. There are five major actions in this process described below:

1) Receiving proposal from the *Customer* and decompose it.
2) Forming small team of contractors that will execute the activities.
3) Providing alternative if small team has issue in the order.
4) Providing alternative contractor if one team needs alternative and other one ready to work.
5) Providing alternative contractor if one team fails to complete its work and other one successfully completed work.

The Fig. 4 shows that the *Logistics* process receives proposal using synchronization channel *Proposal[x]*. These values are process ID of *Customer* describes that which customer sends order. After receiving proposal the *Logistics* receives order from the *Customer* using channel *Order[i][j][cus_id]?*. There are two values ’i’ and ’j’ that are received by the *Logistics* are the activities that customer needs. If *i=0* and *j=1* the *Customer* needs activity *j*, if *i=1* and *j=0* the *Customer* needs activity *i* and if both *i* and *j* are 1 the *Customer* needs both the activities.

After receiving order a *Logistic agent* tries to rank contractors according to the activities a *Customer* demands. For example if customer needs A1 activity then contractor that can perform A1 activity is not available then the order is rejected and *Logistics* goes to *Start* state, ready to receive new order and acknowledges the *Customer*. Similarly, if *Customer* needs A1 and A2 activities contractors for both activities should be available. If contractors successfully ranked *Logistics* assign activities to contractors and goes to *ContractorRanked* State.

Fig. 5 shows the next part of the *Logistics* process. After ranking the contractors *Logistics* waits for the response from the small team whether or not they will accept the contract. This is done by sending each activity to that small team which is available and willing to do work. For this purpose *NegT1[pid][0]!* and *NegT2[pid][1]!* channels are used for *SmallTeam(0)* and *SmallTeam(1)*, respectively where *[pid]* is the process id of *Logistics* sending order and *[0]* and *[1]* values describe the pid of *SmallTeam* to which *Logistics* are sending order. The response from the *SmallTeam* can be of three types *Logistics* receives it on *SmallTeam* state which is as follows:

1) Both the *Small Teams* are ready to do work or committed.
2) *Small Team 1* needs alternative and *Small Team 2* ready to do work.
3) *Small Team 1* ready to do work and *Small Team 2* needs alternative.

If both the *Small Teams* are ready to do work. Then the *Logistics* receive the response using channels *Commited[pid][c]?* from both teams. *Commitcount++* is used to count the commit response, if the value in *Commitcount* is 2 it means both teams committed in case of *Customer* needs one activity the value of *Commitcount* will be 1. If *Small Team 1* needs alternative

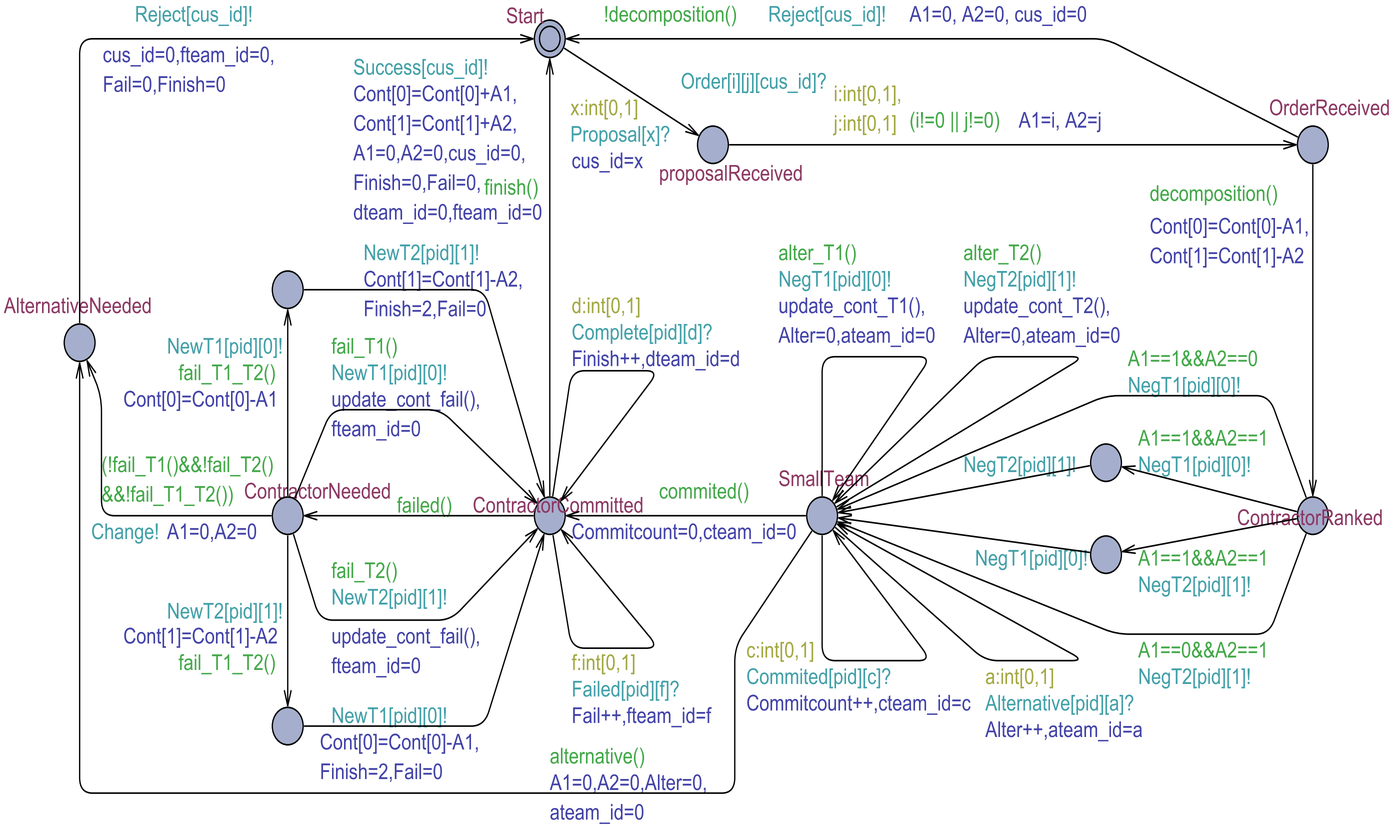


Fig. 3. The logistics process.

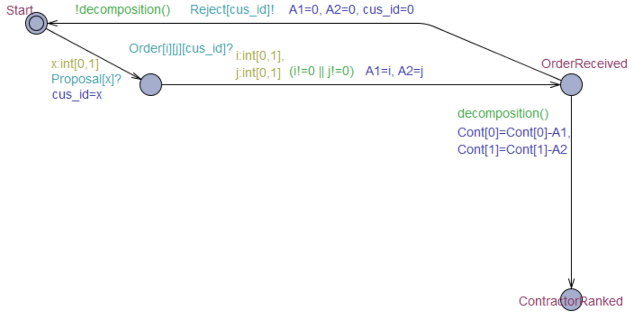


Fig. 4. The logistics process (a).

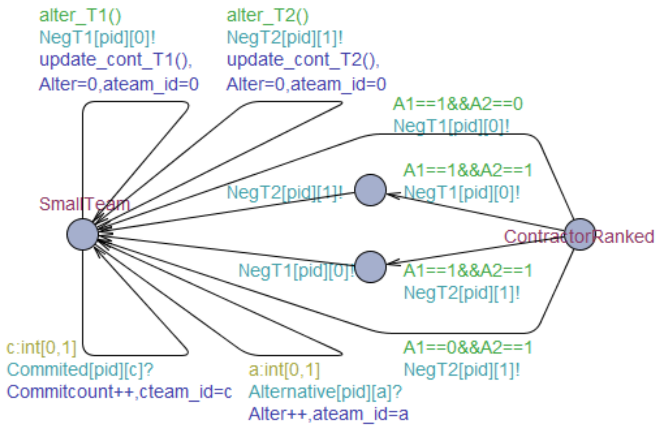


Fig. 5. The logistics process (b).

and *Small Team 2* ready to do work or vice versa then *Commited[pid][c]?* for committed and *Alternative[pid][a]?* for alternative response is used. *Commitcount++* used to count the commit response and *Alter++* used to count the alternative response.

When one team is committed and other needs alternative then *Logistics* checks for the available contractors willing to work and assign activity for which *Small Team* needs an alternative. This is done by using *NegT1[pid][0]!* and *NegT2[pid][1]!* for *Small Team 1* and *Small Team 2*, respectively.

Fig. 6 shows the next procedure after small team formation. If both the teams need alternative in case *Customer* needs both activity then process goes to *AlternativeNeeded* state. Moreover, it goes to *Start* state for negative acknowledgment to customer using channel *Reject[cus_id]!* after which it becomes ready for receiving new order. In case of small Teams are committed to work then process goes to *Contractorcommitted* state. At this state *Logistics* checks whether small teams have complete their work or not. If the teams complete their work successfully then the respective logistic agent goes to start state using channel *Success[cus_id]* and acknowledges the *Customer* accordingly. The number of failed teams are counted

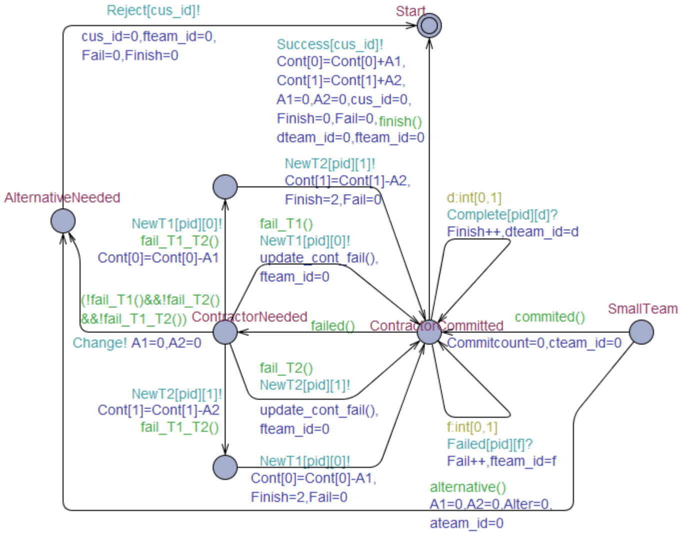


Fig. 6. The logistics process (c).

by *Fail*.

When one team is successful and other team fails then *Logistics* checks for the available contractors that are willing to do work and assign activity for which *Small Team* fails. This is done by using *NewT1[pid][0]!* and *NewT2[pid][1]!* for *Small Team 1* and *Small Team 2*, respectively. If the contractor is available activity assigns to that contractor and after completing work *Logistics* process goes to *ContractorCommitted* state and further more at *Start* state. In case of contractors are not available then process goes to *AlternativeNeeded* using channel *Change!* this transition further more goes to start state using channel *Rejected[cus_id]!*.

When we use UPPAAL system models, functions can be declared within the procedure or process. We can pass parameters in functions and functions can also have return type. The *Logistics Process* have various functions and are used at different transitions to perform its functionality.

1. *Decomposition():* function used as guard and checks the contractor against activities. *A1* and *A2* are the activities. If customer needs both activities vales of *A1* and *A2* will be 1, if customer needs one activity then the value of *A1* and *A2* will be 1 according to the activity that customer needs. This guard prevents to take action if the contractors will not available against the activity which customer needs and take action that goes to start state so logistics can receive new order.

2. *Committed():* function is also used as guard and checks the small teams response. If customer needs both the activities then both the teams should be committed to work if not, guard will prevent to goes to *ContractorCommitted* state. If customer needs one activity then the small team against that activity should be committed. This Function uses an integer variable *CommitCount* for counting the response form the the teams and compares with number of teams and activities.

3. *Alternative():* function is also used to check the small team response. It works same as *committed()* function but the difference is that in case of both activities, if both the teams need alternative this guard will allow to go to *Alternative-Needed* state through alternative transition. And if customer needs one activity then the small team against that activity can ask for alternate. This Function uses an integer variable *Alter* for counting the response form the the teams and compares with number of teams and activities.

4. *Alter_T2():* function is used to check the availability of contractors. If the customer needs both the activities and 1st team committed and 2nd team needs alternative, then contractor for 2nd team should be available otherwise this guard will prevents to take further action and wait for the availability. Statement *Cont[1]¿0* checks the availability. This function uses *Commitcount* and *Alter* variables to check which teams needs alternative.

5. *Alter_T1():* function is also used to check the availability of contractors. If the customer needs both the activities and 1st team needs alternative and 2nd team is committed, then contractor for 1st team should be available otherwise this guard will prevent to take further actions and waits for the availability. Statement *Cont[0]¿0* checks the availability. This function uses *Commitcount* and *Alter* variables to check which teams needs alternative.

6. *Finish():* function is used when small team complete their work successfully after commitment. In case of customer needs one activity, variable *Finish* value will be 1 and function allows to finish the work and *Logistics* goes to *Start* state to take new order. If customer needs both activities value of *Finish* will be 2.

7. *Failed():* function is used if the small team fails to complete work after commitment. Both the teams can be failed or may be one team fail and other complete the task then this function will allows to take action and goes to *ContractorNeeded* state. *Fail* and *Finish* variables are used to check which teams are failed or has completed their work. If value of *Fail* is 2 then both the teams failed, if value of *Fail* and *Finish* is 1 then one team has completed and other has finished the work.

8. *Fail_T2():* function is used to check the availability of contractors in case of one team finishes its work and other team completes its work successfully. If the customer needs both the activities and 1st team finished work and 2nd team failed, then contractor for 2nd team should be available for replacement otherwise this guard will prevents and goes to *AlternativeNeeded* state. In case if a customer needs only 2nd activity and small team fails to complete work then contractor against that activity should be available for replacement. Statement *Cont[1]¿0* checks the availability. This function uses *fteam_id* variable to check which teams is failed.

9. *Fail_T1():* function works same as *fail_T2()* difference is that if the customer needs both the activities and 1st team failed and 2nd team finished, then contractor for 1st team should be available for replacement otherwise this guard will prevents and goes to *AlternativeNeeded* state. And if customer needs only 1st activity and small team fails to complete work then contractor against that activity should be available for replacement. Statement *Cont[0]¿0* checks the availability.

10. *Fail_T1_T2():* function is used if customer needs both activities and both are failed to complete their work after

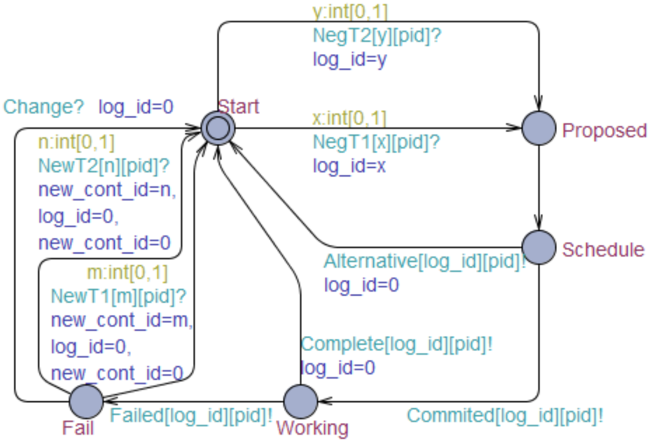


Fig. 7. The small team process.

commitment then both contractors for 1st team and 2nd team should be available for replacement otherwise this guard will prevents and goes to *AlternativeNeeded* state. Statement *Cont[0]¿0* and *Cont[1]¿0* checks the availability.

11. *Update_cont_T1()* and *update_cont_T2():* functions are used to update contractors. If customer needs both activities and one team is committed and other needs alternative. The function used *cteam_id*, *ateam_id* and *Alter* variables to check which team is committed and which one needs alternative.

12. *Update_cont_fail:* function is used to update contractors in case of customer needs both activities and one team has completed its work successfully and other team failed to work. This function uses same variables as previous functions.

D. The Automaton for the Small Team Process

Fig. 7 the automaton for small team process is illustrated. The initial state is named as *Start*. The *Logistics* process has five states. The first state is *Start* state, second is *Proposed*, third is *Schedule*, fourth is *Working* and the fifth state is *Fail* state. There are four major actions in this process described below:

1) Receiving proposal from the logistics.
2) Sending acknowledgment to the logistics if needs some changing or alternative proposal.
3) Sending acknowledgment to the logistics if ready to perform activity and goes to *Working* state.
4) If fails to complete activity goes to fail state and waiting for new contractor who is willing to complete that activity.

The channel *NegT1[x][pid]?* and *NegT2[y][pid]?* receives value from the logistics process, received values are the tasks assigned to the team 1 and team 2 received from channels *NegT1[x][pid]?* and *NegT2[y][pid]?*, respectively. The small team checks its schedule if team has no issue and willing to work. Then this small team acknowledges the logistics using *Commited[log_id][pid]!* channel to go to *Working* state. If there are issues in the proposal like small team has not enough time or could not perform that activity on time, small team acknowledges the logistics that it needs alternative for that task using *Alternative[log_id][pid]!* channel and goes to *Start* state for receiving new or alternative proposal.

After committing small team starts working on the task. If small team fails to complete its task it sends negative acknowledgment to the logistics that it needs new contractor and goes to *Fail* state waiting for new contractor to be assigned by the logistics and this is done by using *Failed[log_id][pid]!* channel. If the contractor is available and willing to work small team is assigned that contractor using *NewT1[m][pid]?* and *NewT2[n][pid]?* channels for team 1 and team 2, respectively otherwise goes to *Start* state after receiving response from the logistics using *Change?* channel.

IV. Functional Requirements

According to [3], we extract the following functional requirements:

R1: Deadlock freedom. No deadlock when a customer needs to places an order. In other words, deadlock can occur only when there are no more orders.

R2: If customer sends order, logistic agents eventually acknowledge it.

R3: A customer is in working state after paying an order.

R4: If logistics agent is in OrderReceived state if it receives an order.

R5: Every order decomposed by some logistic agent results in formulating a small team.

A. Formal Specification of Requirements

In this section, we describe formal specification of the requirements. The customer process sends order and then increases its counter, i.e., known as *proposalCounter*. This increment continues up to two it means the customer can send maximum 2 orders. So, according to the R1 requirement, there is deadlock only when there are no more orders to send by the customers. The formula of R1 requirement is given below.

```
A[] deadlock imply(Customer(0).
     proposalCounter==2&&Customer(1).
     proposalCounter==2))
```

When customer sends order the logistics agent receives and acknowledges it with a message either the given order is workable or not. The formula of to represent this requirement is:

```
E<> forall (i:id_t) forall (j:id_t)
(Customer(i).Working &&Logistics(j).OrderReceived)
```

Formula describes that Customer(0) and Customer(1) sends proposal to Logistics(0) and Logistics(1) and vice versa. The logistics acknowledges the customer.

When customer sends order the logistic agents receives and acknowledges the customer at that time customer goes to *Working* state. For example, when customer(0) sending order definitely goes to *working* state. The formula of this requirement is.

```
forall (i:id_t)Customer(i).proposal -->
           Customer(i).Working
```

According to the R4 requirement, when a logistics agent receives proposal it goes to *OrderReceived* state. The formula of the requirement is given below.

```
forall (i:id_t) Logistics(i).proposalReceived
            --> Logistics(i).OrderReceived
```

According to the R5 requirement, every order decomposed by some logistics agent formulates a small team. The formula of this requirement is given below.

```
forall (i:id_t) Logistics(i).ContractorRanked
            --> Logistics(i).SmallTeam
```

V. Results

To analyse features specified in the above section, we use the verifier, a feature of UPPAAL model checker. Ultimate results are derived in query section of verifier feature and presented in Table I. In query section, the feature is written and its consequences are to be revealed in the status section. The outcomes are in the form of “Satisfied” and “Not Satisfied” of property. We verify our system model for,

Total Number of Processes = 3

Order Sending Limit = 2

Activity Demand Limit = 2

TABLE I. Results

Requirement	Status	Computational Time
R1	Not Satisfied, 131 states	0.125 sec
R2	Satisfied, 28,180 KB	0.015 sec
R3	Satisfied, 138 states	0.539 sec
R4	Satisfied, 1623 states	0.562 sec
R5	Not Satisfied, 32,204 KB	0.032 sec

R1: This requirement is violated and not satisfied. According to the requirement system should be deadlock free or deadlock can occur only when there are no more orders to send. But there is a scenario in which this requirement is not satisfied. When a small team needs alternative there is no more contractor available against that activity at that state deadlock occurs. The counter example for requirement R1 generated by UPPAAL is shown in Fig. 8.

R5 requirement is not satisfied and according to this requirement upon decomposing an order by logistic agents, small team is formed. If a customer needs both activities then upon decomposition if one small team needs alternative but there contractors are unavailable pertaining to that activity then small team is not formed, so this requirements is not satisfied. The counter examples for the requirement is shown in Fig. 9.

VI. Limitations and Challenges

There are some obstructions for authentication of intended Agent-Oriented Supply-Chain Management. We restrict the number of orders to two. We also restrict the number of activities to two and the contractors against those activities. A customer can send maximum two orders and demands for maximum two or minimum one activity. These limitations reduce the state space because the model generates a huge state space. The machines are used in our verification have limited resources for memory and speed. These limitations are also used due to limited memory of machine. The machine can crash during execution of query verification phase. We perform some computations on the machine with 4GB RAM, core i3(3rd Gen) Laptop.

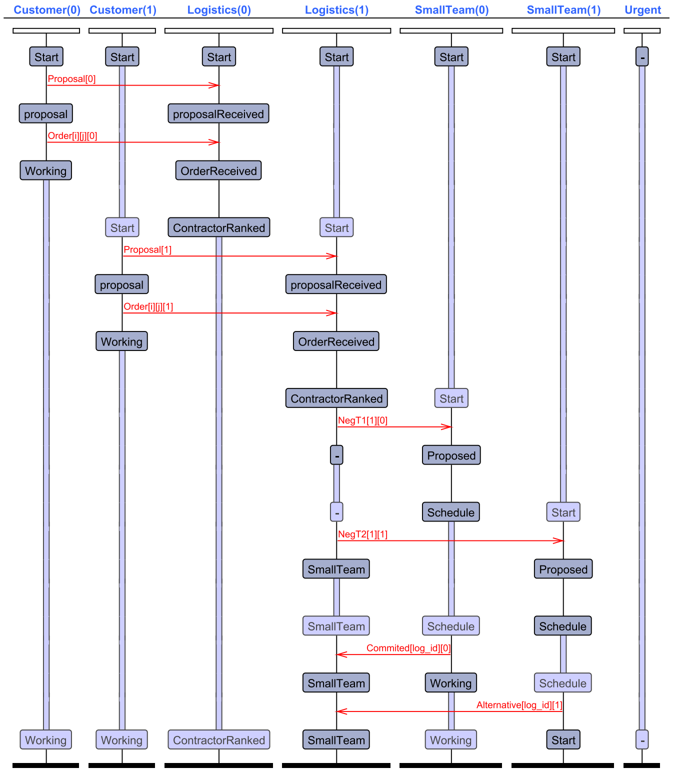


Fig. 8. Trace for requirement R1.

VII. Conclusion

We formalized Agent-Oriented Supply-Chain Management as specified in [3] in UPPAAL model checker. We then formalized functional requirements of the architecture and verified them by model checking. Results show that the given architecture partially fulfills its functional requirements. Proof the results are presents in the form for message sequence charts. The given protocol is verified with limited number of logistic agents, orders and customers.

References

[1] M. C. C. Douglas M. Lambert, “Issues in supply chain management,” *Industrial Marketing Management*, vol. 29, p. 19, 2000.

[2] I. K. Nirupam Julka, Rajagopalan Srinivasan, “Agent-based supply chain management-1: framework,” *Computers and Chemical Engineering*, vol. 26, p. 15, 2002.

[3] R. T. Mark S. Fox, Mihai Barbuceanu, “Agent-oriented supply-chain management,” in *The International Journal of Flexible Manufacturing Systems, 12*. Kluwer Academic Publishers, 2000, pp. 165–188.

[4] C. Baier, J.-P. Katoen *et al.*, “Principles of model checking, vol. 26202649,” *MIT Press Cambridge*, vol. 26, p. 58, 2008.

[5] P. R. D’Argenio, J.-P. Katoen, T. C. Ruys, and J. Tretmans, *The bounded retransmission protocol must be on time!* Springer, 1997.

[6] H. Lonn and P. Pettersson, “Formal verification of a tdma protocol start-up mechanism,” in *Fault-Tolerant Systems, 1997. Proceedings., Pacific Rim International Symposium on*. IEEE, 1997, pp. 235–242.

[7] P. Pettersson, *Modelling and verification of real-time systems using timed automata: theory and practice*. Citeseer, 1999.

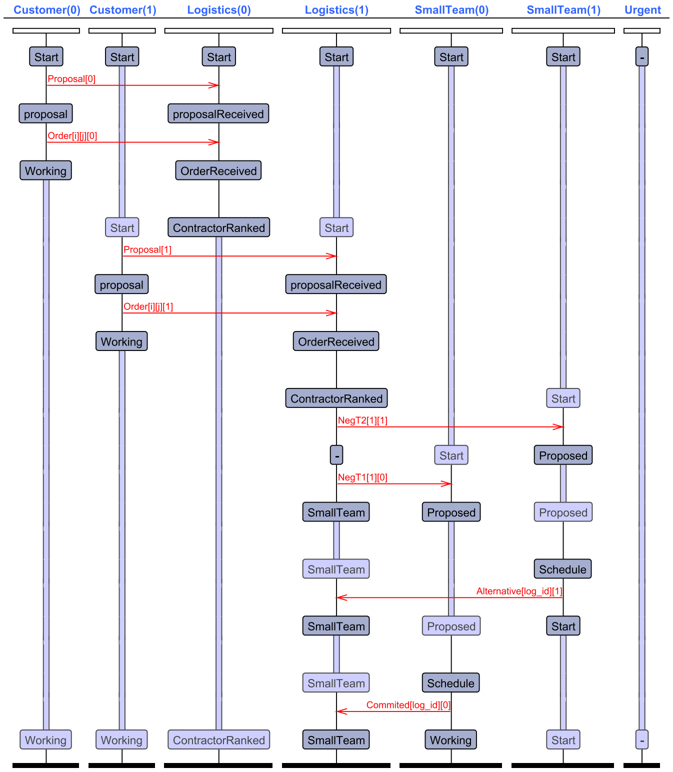


Fig. 9. Trace for requirement R7.

[8] W. Yi, P. Pettersson, and M. Daniels, “Automatic verification of real-time communicating systems by constraint- solving.” in *FORTE*, vol. 6. Citeseer, 1994, pp. 243–258.

[9] M. Atif, “Formal modeling and verification of distributed failure detectors,” *Faculty of Mathematics and Computer Science, TU/e*, vol. 10, 2011.

Effect of TCP Buffer Size on the Internet Applications

Imtiaz A. Halepoto[1], Nazar H. Phulpoto[2], Adnan Manzoor[2], Sohail A. Memon[3], Umair A. Qadir[2]
[1]Department of Computer Systems Engineering, QUEST Nawabshah, Pakistan
[2]Department of Information Technology, QUEST Nawabshah, Pakistan
[3]Department of Mathematics, SALU Khairpur, Pakistan

***Abstract*—The development of applications, such as online video streaming, collaborative writing, VoIP, text and video messengers is increasing. The number of such TCP-based applications is increasing due to the increasing availability of the Internet. The TCP protocol, works at the 4th layer of the Internet model and provides many services such as congestion control, reliable communication, error detection and correction. Many new protocols have been proposed such as stream control transmission protocol (SCTP) with more features compared to the TCP. However, due to the wide deployment, TCP is still the most widely used. TCP creates the segments and transmit to the receiver. In order to prevent the errors TCP saves the segments into the sender buffer. Similarly, the data is also saved at the receiver buffer before its transmission to the application layer. The selection of TCP sender and receiver buffer may be varied. It is very important because many applications work on the smart-phones that are equipped with a small amount of memory. In many applications such as online video streaming, some errors are possible and it is not necessary to retransmit the data. In such case, a small buffer is useful. However, on text transmission the complete reassembly of message is required by the TCP before transmission to the application layer. In such case, the large buffer size is useful that also minimizes the buffer blocking problem of TCP. This paper provides a detailed study on the impact of TCP buffer size on smart-phone applications. A simple scenario is proposed in NS2 simulator for the experimentation.**

Keywords*—TCP; sender buffer; receiver buffer; stream control transmission protocol (SCTP); error detection and correction

I. Introduction

The OSI model in the Internet provides a step by step characterization of the computer and telecommunication systems. Transport layer in the OSI model is one the main layers. It provides congestion control, error control, flow control, a stronger checksum and many other features. In a summarized way, the transport layer works for successful delivery of a process from a sender to a receiver. All of these features are provided the TCP protocol [1]. The two other famous protocols of the transport layer are the User Datagram Protocol (UDP) [2] and SCTP [3] (see also [4]). The UDP is mainly beneficial for applications such as video streaming. It is a less complicated protocol due to its header format but not preferable for applications where the reliability is mandatory. The SCTP is a new protocols and it is still in development phase. The key features in the design of the transport layer are the following:

- **Out-of-order delivery** for faster data transmission to the application layer. In this mode, SCTP the receiver does not wait for complete message it simply forwards the data as soon as it is received. This feature is also available in the UDP. However, it is not available in TCP. One of the main application of out-of-order data is the online video or audio streaming. However, in both the sequenced-data delivery and out-of-order data delivery in online streaming may loss few of the segments. But the overhead of sequencing overhead in out-of-order data is less.
- **Connection-orientation** feature is available in the TCP and SCTP but not in UDP. By this features both the sender and receiver initiate a connection establishment procedure before the data transmission. In UDP all the data units travel independently and forwarded by different routers.
- **Connection formation** is initiated by the SCTP and TCP before the data transmission. The connection formation procedures requires verification of sender and receiver, which also improves the security of the protocols. TCP and SCTP uses 3-way and 4-way handshake procedures for connection formation. However, there is no service of connection formation in UDP.
- **Connection termination** is also completed by the TCP and SCTP after the successful transmission of data from sender to receiver. By this method the sender and receiver agree to close the session. There is no connection termination service provided by the UDP.
- **Reliability** by means of acknowledgment to the sender. This feature is available in TCP and SCTP but not in UDP. One the main factor that affects the buffer size is reliability. For example, a sender keeps a copy of the transmitted segment in the sender buffer until the acknowledgment is received or the time to acknowledgment expires.
- **Flow control** to maintain the data transmission rate of the sender. By flow control the protocol reduces the chances of network congestion and other others errors such as data overflow. With the flow control, TCP tries to maintain a synchronization between the sender and the receiver. For example, the synchronization is needed when the sender is very fast compared the receiver.

Many of the features of the transport layer protocols are presented in Fig. 1. Despite all the good features of SCTP, the TCP is currently fully operational over the Internet. In

TCP, the segments that are in queue for transmission and the segments that are revived are stored in a memory called the TCP sender and receiver buffer. Buffer size play major role in the performance of TCP. If the buffer size is too small the TCP would be unable to complete the message by combining the segments at the receiver. It also leads to no buffer space for the parallel TCP flows of other applications. Such type of buffer condition is called the receiver buffer blocking. Similarly, on the sender side, if the buffer size is small it affects the transmission rate due the less number of segments within the sender buffer. With the increasing number applications such as video streaming, messengers, online chats, VoIP, collaborative scientific projects, wireless sensors and monitoring. It is difficult to decide the buffer size requirement. Because some of the applications require reliability and some of them do not. Further, the development of smart-phone apps is also increasing. It is difficult to determine which type of data processing will be carried by the apps at the time of development, because of the real time data processing and software updates. In order to help the developers of smart-phone apps, the consideration of buffer size for the protocol is necessary. Additionally, the researchers are working on the parallel data transmission by using more than one NIC cards. By parallel transmission throughput increases by the factor depending on the number of NIC cards. For the parallel data transmission a new version of TCP is under development. It is called the Multipath TCP (MPTCP) [5]. This research work aims to provide the experimentation of the TCP protocol with various buffer size. The proposed scenario is composed of multihop network. The background traffic is also added in order to make the scenario more like as real life networks where the bandwidth is occupied by the several number of users. The experimentation results are also useful for evaluation of MPTCP. The simulation is carried in NS2 with varying size of the sender and receiver buffer.

The rest of the paper contains the related work in Section II. The experimental setup and configuration details in Section III. The analysis on the basis of the results is presented in Section IV. The conclusions are summarized in Section V.

II. Related Work

The choice of buffer size affects the performance of TCP. For example, if the receiver buffer size equal to the 50 segments. If the there are two processes transmitted by the sender. Each of the process contains 30 segments. In simultaneous transmission of both the processes the receiver will be occupied by the 50 segments. 25 segments from each of the process. Both the processes are received incomplete. The receiver will be waiting for the remaining segments and none of the process will be delivered to the application layer. This kind of situation is called the receiver buffer blocking. Many researchers reported the problem of receiver buffer blocking while using TCP [7], [8], [9], [18], [19]. The researchers also suggested the use of retransmission policies for the transmission missing data of one process. However, such retransmission polices are beneficial for the parallel transmission of data by using more than one link. For one link between a sender and receiver the role of retransmission policy slightly improves performance.

Feature	**SCTP**	**TCP**	**UDP**
Connection-oriented	Yes	Yes	No
Full-duplex	Yes	Yes	Yes
Reliable Data Transfer	Yes	Yes	No
Partial reliable data transfer	Optional	Yes	No
Ordered data delivery	Yes	Yes	No
Unordered data delivery	Yes	No	Yes
Flow control	Yes	Yes	No
Congestion control	Yes	Yes	No
ECN capable	Yes	Yes	No
SACK	Yes	Optional	No
Preservation of Message Boundaries	Yes	No	Yes
Path MTU Discovery	Yes	Yes	No
Application PDU Fragmentation	Yes	Yes	No
Application PDU Bundling	Yes	Yes	No
Multistreaming	Yes	No	No
Multihoming	Yes	No	No
Protection against SYN flooding attacks	Yes	No	N/A
Allows half-closed connections	No	Yes	N/A
Reachability Check	Yes	Yes	No
Pseudo-header for Checksum	No	Yes	Yes
Time Wait State	for vtags	for 4-tuples	N/A

Fig. 1. Summary of the services provided by the transport layer protocols [6].

The buffer splitting techniques were proposed by the researchers in [10] and [11]. They proposed two kinds of splitting. First, that equally divide the buffer space in the number of destinations or paths. In real life data from different paths to receiver take different time. So, on the slow path (path of smaller bandwidth or longer propagation delay) data transmission may affect the data that is already in the receiver buffer. On the other side the faster paths occupy more buffer space and may reach to the buffer overflow. Second, the technique, which divides the buffer into parts for the different processes according to outstanding data. The outstanding data is the data that already transmitted by the sender but not yet acknowledged. The work in [12] suggested the use of available buffer space in acknowledgment segment, because this value represents the exact free space of the buffer. Normally, TCP uses the advertised buffer space in the acknowledgment segment. The relationship between the buffer size and the round trip time (RTT) is investigated by Want *et al.* [13]. According the their findings the relationship is linear.

The work on RTT and the other path performance characteristics such as bandwidth is investigated by the researchers in [14]. The technique of buffer splitting at the sender and receiver is employed in order to reduce the buffer blocking problem. The splitting is performed on the basis of the RTT. The destinations with longer RTT value (slow paths) are allowed to use the small portion of the buffer size. Whereas, the destinations with shorter RTT value (fast paths) are configured to use the large portion of the buffer space both at the sender and receiver. The similar work to improve the performance by minimizing the buffer blocking is also presented in [15], where the technique of transmission scheduling are proposed. Scheduling of different data flows is based on a priority value,

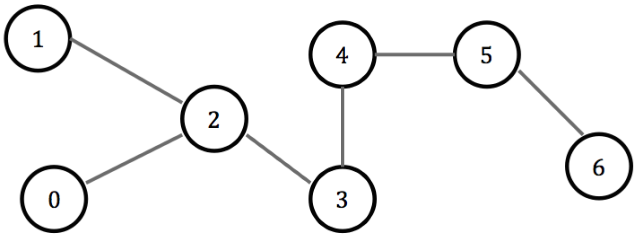


Fig. 2. Network Topology 1: Multihop network.

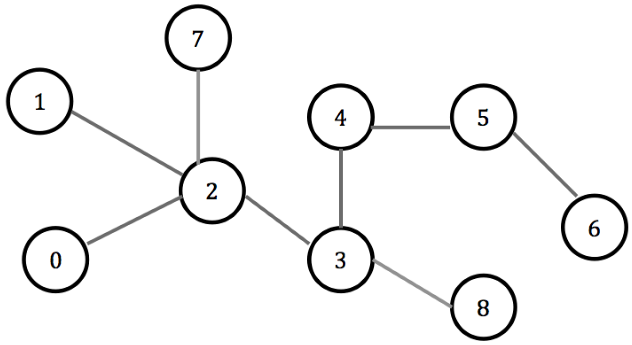


Fig. 3. Network Topology 2: Multihop network with background traffic.

which is calculated by using the outstanding data of each flow. The researchers in [16] suggest that the design of a routing protocol is very important.

III. TCP Implementation and Configuration

The NS2 [17] is used for the implementation and evaluation of the TCP. For the installation of NS2, the Ubuntu 14.04 OS is installed in the virtualbox. The multihop network is proposed for the experimentation. In all of the experiments the throughput is measured by assuming node 0 as source and node 6 as destination. Each of the simulation is repeated 10 times and the results are collect on the basis of average values. The implementation code of TCP is already available in the NS2, however its configuration is required according to the proposed topologies given below. The key parameters of the simulation are presented in Table I.

A. Topology 1

In this topology seven nodes are configured and attached as shown in Fig. 2. Node 0 and 1 are configured with TCP agents as source nodes. Node 6 is the destination node and it is configured with TCP agent as sink. In this topology the main connection that is monitored for throughput is 0–6, whereas 1–6 is just adding an additional TCP flow.

B. Topology 2

In the second topology, nine nodes are used. Nodes 7 and 8 are attached with the intermediate nodes 2 and 3. The main purpose these additional nodes is to provide the background traffic to the TCP. Topology 2 is shown in Fig. 3. The remaining parameters and their configuration are left on the default values present in NS2.

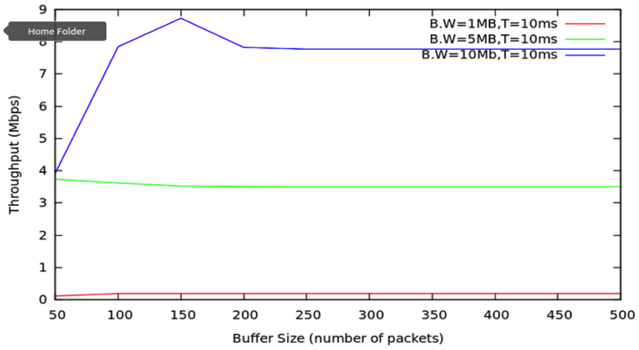


Fig. 4. Experiment 1a.

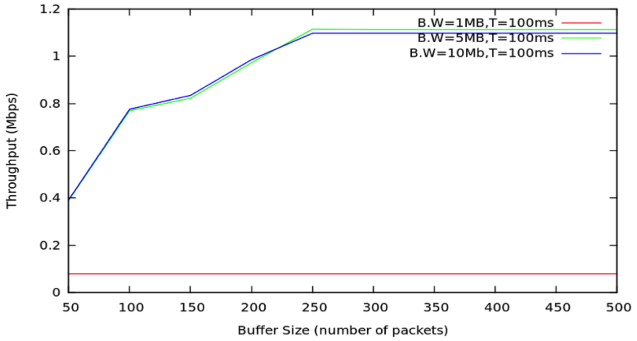


Fig. 5. Experiment 1b.

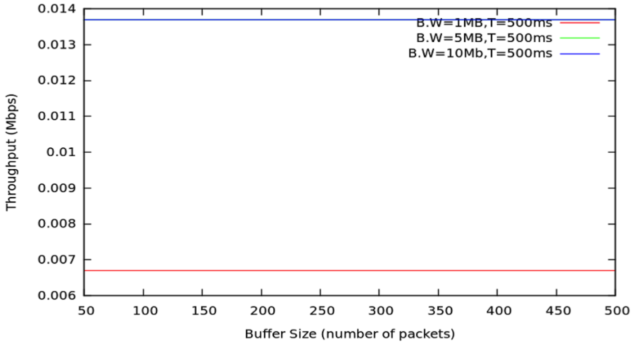


Fig. 6. Experiment 1c.

IV. Results and Discussions

A. Experiment–1

In this experiment, three kinds of simulations are performed each time while changing the values of the propagation delay. The value of sending buffer changes from 50 to 500 number of segments. It is observed that when the delay is small as in Fig. 4, the TCP throughput is directly proportional to the value of bandwidth, i.e. 10Mbps. When the delay increase as in Fig. 5 and 6, the medium is not fully occupied. Hence the throughput at 10Mbps and 5Mbps is also same. However, it is greater than the throughput at 1Mbps. It is also clear that the increase in the buffer size does not significantly affect the data transmission rate. The buffer size of 200–250 is enough to reach the maximum throughput.

TABLE I. Parameters and Values

Experiment	Nodes	Bandwidth(Mbps)	Delay(ms)	No. of Simulations	Network
1. Changes in the sender buffer					Topology–1
1a	0 to 6	1, 5, 10	10	10	
1b	0 to 6	1, 5, 10	100	10	
1c	0 to 6	1, 5, 10	500	10	
2. Background traffic					Topology–2
2a	0 to 6	1, 5, 10	10	10	
2b	0 to 6	1, 5, 10	100	10	
2c	0 to 6	1, 5, 10	500	10	
3. Large Buffer Size	0 to 6	1, 5, 10	10	10	
4. Changes in the Receiver buffer					Topology–2
4a	0 to 6	1, 5, 10	10	10	
4b	0 to 6	1, 5, 10	10	10	
5. Equal Buffer Size	0 to 6	1, 5, 10	10	10	Topology–2

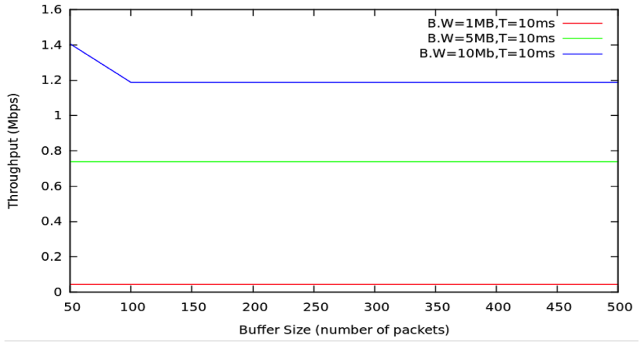


Fig. 7. Experiment 2a.

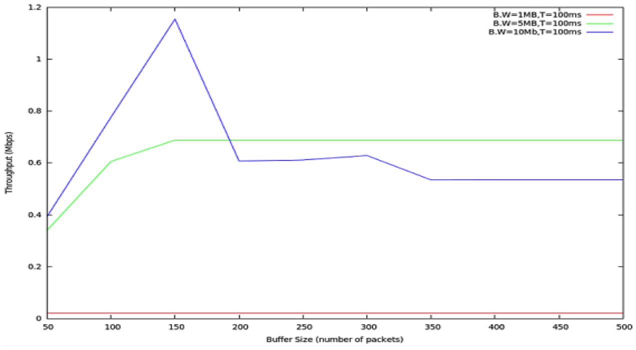


Fig. 8. Experiment 2b.

B. Experiment–2

The topology 2 is used in this experiment. A TCP flow is defined from node 3–8 and its affect on the flow of nodes 0–6 is observed. The same bandwidth and delay values of Experiment–1 are applied. The results trends are very similar to Experiment–1. The results are presented in Fig. 7, 8 and 9. When the buffer is smaller than 200 there is steady improvement in the throughput. However, with a larger buffer of more than 200 packets, the maximum throughput has reached.

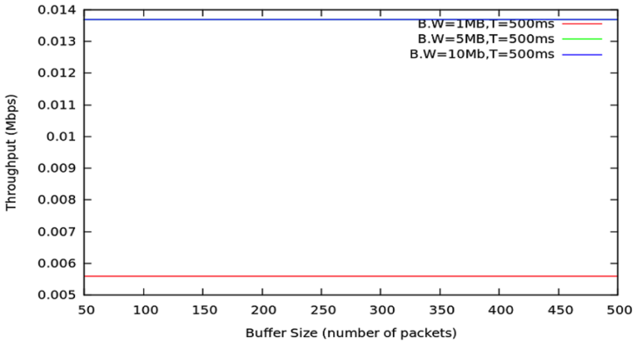


Fig. 9. Experiment 2c.

C. Experiment–3

In this experiment, the simulation of Experiment–2 is extended for a very large buffer size. The buffer size used is from 1000 packets to 10000 packets. The delay is configured to 10ms, however the experiment is repeated with different values of bandwidth, i.e. 1Mbps, 5Mbps and 10Mbps. This expriments proves that the large size of buffer is not useful in the proposed scenario, the throughput remains the same. In Fig. 10, the throughput at 1Mbps, 5Mbps and 10Mbps is 50Kbps, 0.7Mbps and 1.2Mbps.

D. Experiment–4

In this experiment, Topology–2 is used. However the changes are performed in the receiver buffer instead of the sender buffer. According the investigations, when the delay is smaller, the sender transmits data as long as it is available. Due to which there are less occasions of packet loss in the 0–6 TCP flow. So, the output remains at the same value as shown in Fig. 11 and 12, where different values of bandwidth are used. In the case of longer delay of 100ms, the small buffer does not reaches the larger throughput. But as the buffer size increases the the throughput also increases. After the buffer size of 250 packets, the additional space in the buffer space does not increases the throughput.

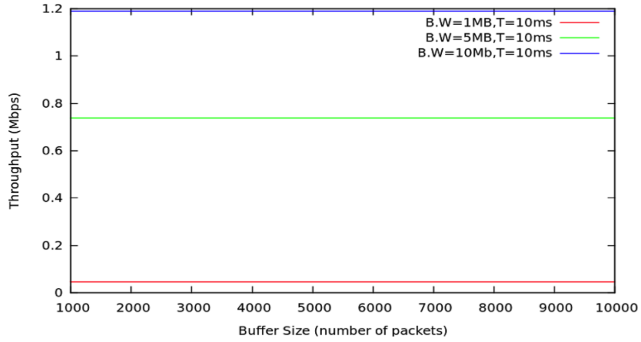


Fig. 10. Experiment 3.

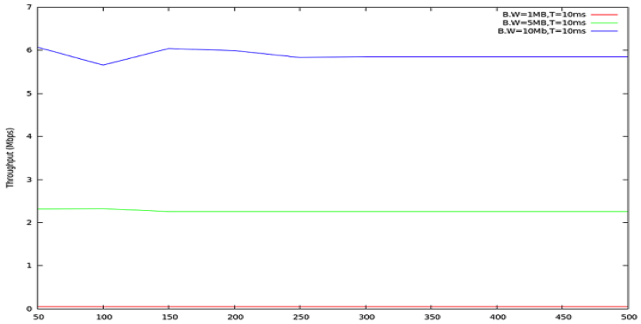


Fig. 11. Experiment 4a.

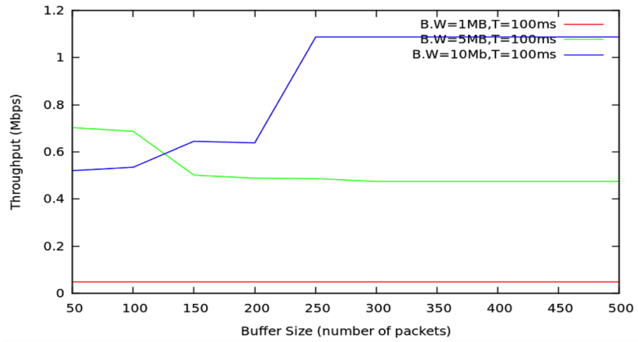


Fig. 12. Experiment 4b.

E. Experiment–5

The last experimentation is carried out by changing the buffer size both at the sender and receiver. In this experiment, the buffer size increases from 50–to–500 packets. It is observed that, when the buffer is small the throughout is less. When the buffer size is large it increase to the throughput but up 100 packets size. After the size of 100, the increase in the buffer space does not increase the throughput. The results of Experiment–5 are shown in Fig. 13.

In all of the experiments after a certain buffer space the performance of TCP remain the same in terms of throughput. According the Experiment–5, the significant amount of the buffer space is 100 packets. That is equal to the bandwidth delay product. In this experiment, the bandwidth delay product is 10Mbps * 10ms = 100K. The suggestion of the buffer size that is twice the bandwidth product is also presented in [6].

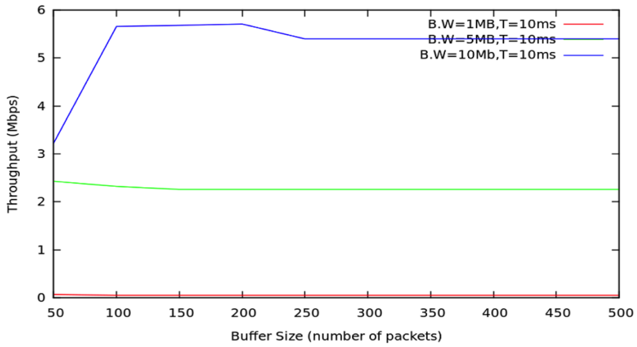


Fig. 13. Experiment 5.

V. Conclusion

The application development using Android or other platforms is increasing. The applications such as video/audio streaming, online collaboration, VoIP, messengers are the need of time. Some of them require sequenced delivery like collaborative writing projects, whereas some of them like online video streaming the sequenced delivery is not the priority. In video streaming the best and fast delivery is important. Many protocols are also available to deal with the sequenced and out-of-order delivery of data such as UDP, TCP and SCTP. TCP is one of the most widely used protocol over the Internet. Depending on the type of application the requirement of the buffer space at the sender and receiver is different. If not considered properly the buffer size the problem such as buffer blocking and buffer overflow may occur. This work provides the details of the experimentation of TCP with different buffer size options. According the results of the simulation over a multihop scenario, the too large buffer size does not increases the throughput. On the other hand the smaller buffer also degrades the performance of TCP. The finds suggest that the buffer size of twice the bandwidth delay product is suitable for the TCP flows. In future, the work may be extended on the upcoming version of TCP called the MPTCP.

References

[1] Allman, Mark, Vern Paxson, and Ethan Blanton. TCP congestion control. No. RFC 5681. 2009.

[2] Postel, Jon. User datagram protocol. No. RFC 768. 1980.

[3] Borella, Michael S. ”System and method for control of packet data serving node selection in a mobile internet protocol network.” U.S. Patent No. 7,346,684. 18 Mar. 2008.

[4] Halepoto, Imtiaz Ali, Adnan Ahmed Arain, and Umair Hussain. ”Evaluation of multimedia streams in internet applications.” Proceedings of the 8th International Conference on Information Systems and Technologies. ACM, 2018.

[5] Ford, Alan, *et al*. Architectural guidelines for multipath TCP development. No. RFC 6182. 2011.

[6] Halepoto, Imtiaz Ali. ”Scheduling and flow control in CMT-SCTP.” HKU Theses Online (HKUTO) (2014).

[7] Janardhan R Iyengar, Paul D Amer, and Randall Stewart. Performance implications of a bounded receive buer in concurrent multipath transfer. Computer Communications, 30(4):818829, 2007.

[8] Janardhan R Iyengar, Paul D Amer, and Randall Stewart. Receive buer blocking in concurrent multipath transfer. In Global Telecommunications Conference, 2005. GLOBECOM05. IEEE, volume 1, pages 6pp. IEEE, 2005.

[9] Halepoto, I. A., Memon, M. S., Phulpoto, N. H., Rajput, U., Junejo, M. Y. On the use of Multipath Transmission using SCTP. IJCSNS, 18(4), 58. Chicago, 2018.

[10] Hakim Adhari, Thomas Dreibholz, Martin Becke, Erwin P Rathgeb, and Michael Tuxen. Evaluation of concurrent multipath transfer over dissimilar paths. In Advanced Information Networking and Applications (WAINA), 2011 IEEE Workshops of International Conference on, pages 708714. IEEE, 2011.

[11] Thomas Dreibholz, Martin Becke, Erwin P Rathgeb, and M Tuxen. On the use of concurrent multipath transfer over asymmetric paths. In Global Telecommunications Conference (GLOBECOM 2010), 2010 IEEE, pages 16. IEEE, 2010.

[12] Rungeler, Irene, Michael Txen, and Erwin P. Rathgeb. ”Congestion and flow control in the context of the message-oriented protocol SCTP.” International Conference on Research in Networking. Springer, Berlin, Heidelberg, 2009.

[13] Yang, Wang, Hewu Li, and Jianping Wu. ”Pam: Precise receive buffer assignment method in transport protocol for concurrent multipath transfer.” Communications and Mobile Computing (CMC), 2010 International Conference on. Vol. 1. IEEE, 2010.

[14] Halepoto, Imtiaz A., Francis CM Lau, and Zhixiong Niu. ”Management of buffer space for the concurrent multipath transfer over dissimilar paths.” Digital Information, Networking, and Wireless Communications (DINWC), 2015 Third International Conference on. IEEE, 2015.

[15] Halepoto, Imtiaz A., Francis CM Lau, and Zhixiong Niu. ”Scheduling over dissimilar paths using CMT-SCTP.” Ubiquitous and Future Networks (ICUFN), 2015 Seventh International Conference on. IEEE, 2015.

[16] Bhangwar, Noor H., Imtiaz A. Halepoto, Intesab H. Sadhayo, Suhail Khokhar, and Asif A. Laghari. ”On Routing Protocols for High Performance.” Studies in Informatics and Control 26, no. 4: 441-448, 2017.

[17] Greis, Marc. ”Marc Greis tutorial for the ucb/lbnl/vint network simulator ns.” (2004).

[18] Halepoto, I.A, Sadhayo, I.H, Memon M.S, Manzoor A, Bhatti, S. Analysis of Retransmission Policies for Parallel Data Transmission. Engineering, Technology & Applied Science Research (ETASR), 8(3), 2018.

[19] Imtiaz A. Halepoto, Adnan Manzoor, Nazar H. Phulpoto, Sohail A. Memon and Muzamil Hussain, Mobility Management Using the IP Protocol International Journal of Advanced Computer Science and Applications (IJACSA), 9(5), 2018. http://dx.doi.org/10.14569/IJACSA.2018.090562

Non-Linear Energy Harvesting Dual-hop DF Relaying System Over $\eta-\mu$ Fading Channels

Ayaz Hussain[†], Nazar Hussain Phulpoto[††], Ubaidullah Rajput[†††], Fizza Abbas[†††], and Zahoor Ahmed Baloch[†]
[†]Electrical Engineering Department, Balochistan University of Engineering & Technology, Khuzdar, Pakistan
[††]Department of Information Technology, Quaid-e-Awam UEST Nawabshah, Pakistan
[†††]Department of Computer Systems Engineering, Quaid-e-Awam UEST, Nawabshah, Pakistan

***Abstract*—In this work, we analyze a wireless energy harvesting decode-and-forward (DF) relaying network with beamforming that is based on a practical non-linear energy harvesting model over η–μ fading channels. We consider a dual-hop relaying system having multiple antennas at the source and destination only. The single-antenna energy constrained relay assists the source to communicate with the destination. At the relay node, we assume a non-linear energy harvesting receiver which limits the harvested power level with a saturation threshold. By considering a power-splitting based relaying (PSR) protocol and a non-linear energy harvesting receiver, we analyze the system performance in terms of the outage probability and throughput for various antennas combinations and for various values of the fading parameters, η and μ. The $\eta-\mu$ fading model has a few particular cases, viz., Rayleigh, Nakagami-m, and Hoyt. These results are general and can be reduced for different fading scenarios as well as for linear energy harvesting relaying.**

***Keywords*—Energy harvesting relay; non-linear energy harvester; $\eta-\mu$ fading; power-splitting-based relaying; throughput**

I. Introduction

Wireless energy harvesting is a method by which energy is harvested from radio frequency (RF) signals. It is a gorgeous solution to increase the wireless network lifetime [1] and got much interest in the last decade [1]–[13]. The RF signals can accommodate energy and information simultaneously, hence energy can be harvested from the RF signals and can be stored for rechargeable devices to operate [1].

Integrating a dual-hop relaying system with wireless energy harvesting techniques can enhance the coverage capacity and network life-time [2], [3]. In a dual-hop wireless energy harvesting relaying network, the relay is an energy constrained device which harvests the energy through RF signals [4]. One of the well-known relaying methods is a decode-and-forward (DF) relaying, in which relay decodes and sends the information to the destination [2]–[4]. Several authors have investigated the wireless energy harvesting in a DF relaying system [2]–[5], [7]–[13] (and references therein).

In the works [2]–[4] (and their references), for the dual-hop energy harvesting DF relaying networks, a linear energy harvesting receiver was assumed. But, in reality, an energy harvesting circuit is non-linear owing to the non-linearity of the electronic components, i.e., inductors, capacitors, and diodes [7]. Therefore, a non-linear energy harvesting receiver is not a practical node and restrict the level of the harvested power. The performance of DF energy harvesting relaying systems with a non-linear energy harvesting receiver was studied over conventional fading channels [7]–[13] and general $\kappa-\mu$ shadowed fading channels [7].

Despite the importance of a non-linear energy harvesting receiver in a dual-hop energy harvesting relaying system, impact of the non-linear mode of energy harvesting receiver over $\eta-\mu$ fading channels is not investigated so far.

Therefore, in this work, the effects of the non-linearity of the energy harvesting receiver and transmit/receive beamforming on the system performance of the considered system in $\eta-\mu$ fading environments is studied. A PSR protocol [4] is assumed and the performance is evaluated for η–μ fading channels. First, in a delay-limited transmission mode, the analytical results for the performance metrics (i.e., outage probability and throughput) are described. Then, the performance in the $\eta-\mu$ fading scenarios is evaluated for different conditions. The $\eta-\mu$ fading model is a generalized fading model and has a few particular cases, viz. Rayleigh, Nakagami-m, and Hoyt [14]. Therefore, the general $\eta-\mu$ fading scenario can be figured to some symmetric and asymmetric cases, namely, the Rayleigh/Rayleigh, Hoyt/Hoyt, Nakagami-m/Nakagami-m, and mixture of these fading links.

We sectioned our paper as: Section 2 explains the system and channel models; in Section 3, performance metrics are described; the distinctive cases are debated in Section 4; the simulated results are presented in Section 5; Section 6 summarizes the paper.

II. System and Channel Models

A. System model

Consider a half-duplex dual-hop DF relaying system with energy harvesting where a source having multiple-antennas (N_1) is transmitting the information to the destination having multiple antennas (N_2) via a single-antenna relay. The relay is an energy constrained device which harvests the energy from RF signals of the source. The source and destination have no direct link, therefore, relay uses the harvested power to forward the source information to the destination. A PSR protocol [4] is assumed at the relay and transmission block T is splitted into two equal parts. During $1/T$, the relay divides the received RF signals into two portions with power splitting ratio ρ, where $\rho \in (0, 1)$, one for signal detection and the second for energy harvesting. In the second time slot $T/2$, employing the harvested power, the relay forwards the information signals to the destination. The destination unify all signals with MRC (maximum-ratio transmission) technique. Let $\mathbf{h}_1$ and $\mathbf{h}_2$ are the $N_1 \times 1$ and $1 \times N_1$ channel vectors of the source-relay and relay-destination, respectively.

The received signal at the relay node can be written as [7]

$$y_{_R} = \sqrt{P_s}\mathbf{h}_1^{\dagger}\mathbf{w}_1 x + n_{a,r} \tag{1}$$

where $\mathbf{w}_1 = \frac{\mathbf{h}_1}{||\mathbf{h}_1||}$ [7], and P_s and x are the source transmit power and the normalized source information signal, respectively. Further, at the relay antenna, $n_{a,r} \sim \mathcal{CN}(0, \sigma_{a,r}^2)$ is the additive white Gaussian noise (AWGN).

The received signal at the relay $y_{_R}$ is divided into two portions: Energy harvester utilizes the one of the portion $\sqrt{\rho}y_{_R}$ and the information receiver utilizes the other remaining portion $\sqrt{1-\rho}y_{_R}$. During $1/T$, energy is collected by the energy harvester with an energy conversion efficiency α as [7]

$$E_h = \alpha\rho P_s||\mathbf{h}_1||^2\left(\frac{T}{2}\right). \tag{2}$$

A non-linear energy harvesting receiver is considered that emits a constant transmit power αP_{th} when the input power is beyond a saturation threshold power P_{th} [7]. Therefore, the relay transmit power P_r can be [7]:

$$P_r = \frac{E_H}{T/2} = \alpha\rho \min\left(P_s||\mathbf{h}_1||^2,\ P_{th}\right) \tag{3}$$

$$= \begin{cases} \alpha\rho P_s||\mathbf{h}_1||^2, & P_s||\mathbf{h}_1||^2 \leq P_{th}, \\ \alpha\rho P_{th}, & P_s||\mathbf{h}_1||^2 > P_{th}. \end{cases} \tag{4}$$

It is clear from (3) that if $P_s||\mathbf{h}_1||^2 \leq P_{th}$ then the energy harvesting receiver operates in a linear mode otherwise it operates as a non-linear device.

At the information processing receiver, the second portion of the signal $\sqrt{(1-\rho)}y_{_R}$ is given as

$$\sqrt{(1-\rho)}y_{_R} = \sqrt{(1-\rho)P_s}\mathbf{h}_1^{\dagger}\mathbf{w}_1 x + \sqrt{(1-\rho)}n_{a,r} + n_{c,r}. \tag{5}$$

Using (5), the signal-to-noise ratio (SNR) at the relay terminal can be:

$$\gamma_{_R} = \frac{(1-\rho)P_s||\mathbf{h}_1||^2}{(1-\rho)\sigma_{a,r}^2 + \sigma_{c,r}^2} \tag{6}$$

where $n_{c,r} \sim \mathcal{CN}(0, \sigma_{c,r}^2)$ symbolizes the AWGN at the relay owing to RF-to-baseband transformation. Employing the harvested power, P_r, the relay sends the information to the destination is written by

$$\mathbf{y}_{_D} = \sqrt{P_r}\mathbf{h}_2\mathbf{w}_2 x_r + \mathbf{n}_{a,d} + \mathbf{n}_{c,d} \tag{7}$$

where $\mathbf{w}_2 = \frac{\mathbf{h}_2}{||\mathbf{h}_2||}$ [7], x_r denotes the transmitted signal from the relay, and $\mathbf{n}_{a,d} \sim \mathcal{CN}(0, \sigma_{a,d}^2\mathbf{I}_{N_2})$ and $\mathbf{n}_{c,d} \sim \mathcal{CN}(0, \sigma_{c,d}^2\mathbf{I}_{N_2})$ represent the AWGNs at the destination antennas and RF-to-baseband transformation at the destination, respectively.

Then the SNR at the destination $\gamma_{_D}$ from (3) and (7) is obtained as

$$\gamma_{_D} = \frac{P_r||\mathbf{h}_2||^2}{\sigma_{a,d}^2 + \sigma_{c,d}^2} = \begin{cases} \frac{\alpha\rho P_s||\mathbf{h}_1||^2||\mathbf{h}_2||^2}{\sigma_{a,d}^2+\sigma_{c,d}^2}, & P_s||\mathbf{h}_1||^2 \leq P_{th}, \\ \frac{\alpha\rho P_{th}||\mathbf{h}_2||^2}{(\sigma_{a,d}^2+\sigma_{c,d}^2)}, & P_s||\mathbf{h}_1||^2 > P_{th}. \end{cases} \tag{8}$$

TABLE I. SPECIAL CASES OF THE $\eta-\mu$ FADING DISTRIBUTION [14]

Fading distribution	η	μ
Nakagami-m	$\eta \to 1$	$\mu = m/2$
Nakagami-q or (Hoyt)	$\eta \to q^2$	$\mu = 0.25$
Rayleigh	$\eta \to 0$	$\mu = 0.5$

B. The $\eta-\mu$ Channel model

The $\eta-\mu$ fading model is a general fading model and has a few particular cases, viz., Rayleigh, Nakagami-m, and Hoyt [14]. If any of the dual-hop link encounters $\eta-\mu$ fading, then the probability density function (PDF) of the instantaneous SNR γ_ℓ $(\ell = 1, 2)$ can be expressed as [14, eq. (3)]

$$f_{\gamma_\ell}(\gamma) = \frac{2\sqrt{\pi}h_\ell^{N_\ell\mu_\ell}}{\Gamma(N_\ell\mu_\ell)H_\ell^{N_\ell\mu_\ell-0.5}}\left(\frac{\mu_\ell}{\overline{\gamma}_\ell}\right)^{N_\ell\mu_\ell+0.5}\gamma^{N_\ell\mu_\ell-0.5} \times \exp\left(\frac{2\mu_\ell h_\ell}{\overline{\gamma}_\ell}\gamma\right) I_{N_\ell\mu_\ell-0.5}\left(2\frac{\mu_\ell H_\ell}{\overline{\gamma}_\ell}\gamma\right) \tag{9}$$

where $h_\ell = (2 + \eta_\ell^{-1} + \eta_\ell)/4$, $H_\ell = (\eta_\ell^{-1} - \eta_\ell)/4$, $\Gamma(\cdot)$ indicates the Gamma function, and η_ℓ and μ_ℓ are the fading parameters. Additionally, γ_ℓ and $I_{\mathrm{v}}(\cdot)$ denote the average SNR of the ℓ-the link and modified Bessel function of the first kind of v-th order, respectively [14]. Particular cases of the $\eta-\mu$ fading model are condensed in Table I where m and q, respectively, symbolize the fading parameters of the Nakagami-m and Hoyt fading distributions.

III. PERFORMANCE ANALYSIS

A. Outage probability analysis

A dual-hop energy harvesting DF relaying system can be in outage when any link (i.e., source-relay or relay-destination) goes in outage. Mathematically, for the considered system, the outage probability can be written by [7]

$$P_{out} = 1 - p_r[\min(\gamma_R,\ \gamma_D) > \gamma_{th}] \tag{10}$$

where γ_R and γ_D are given by (6) and (8), respectively, and $p_r[\,\cdot\,]$ indicates probability.

B. Throughput analysis

The achievable throughput of the considered dual-hop energy harvesting DF relaying system in a delay-limited transmission mode is given by [7]

$$\tau = \frac{1 - P_{out}}{2}U. \tag{11}$$

Using (11), the optimal power-splitting ratio ρ^* and the optimal throughput τ^* can easily be obtained numerically with the help of Matlab or Mathematica.

IV. SPECIAL CASES

The $\eta-\mu$ fading model has a few particular cases, viz., Rayleigh, Nakagami-m, and Hoyt. Therefore, the throughput and the outage probability expressions for the distinctive cases (i.e., symmetric and asymmetric fading scenarios) can be acquired from (10) and (11) with special parameters. They are condensed in Table II with special parameters.

TABLE II. SPECIAL CASES FROM THE OBTAINED RESULTS FOR $\eta-\mu$ FADING CHANNELS

First hop/second hop	η_1	μ_1	η_2	μ_2
$\eta-\mu$/Hoyt	η_1	μ_1	q_2^2	0.5
$\eta-\mu$/Nakagami-m	η_1	μ_1	1	$m_2/2$
$\eta-\mu$/Rayleigh	η_1	μ_1	1	0.5
Hoyt/Hoyt	q_1^2	0.5	q_2^2	0.5
Hoyt/$\eta-\mu$	q_1^2	0.5	η_2	μ_2
Hoyt/Nakagami-m	q_1^2	0.5	1	$m_2/2$
Hoyt/Rayeligh	q_1^2	0.5	1	0.5
Nakagami-m/Nakagami-m	1	$m_1/2$	1	$m_2/2$
Nakagami-m/$\eta-\mu$	1	$m_1/2$	η_2	μ_2
Nakagami-m/Hoyt	1	$m_1/2$	q_2^2	0.5
Nakagami-m/Rayleigh	1	$m_1/2$	1	0.5
Rayleigh/Rayleigh	1	0.5	1	0.5
Rayleigh/$\eta-\mu$	1	0.5	η_2	μ_2
Rayleigh/Hoyt	1	0.5	q_2^2	0.5
Rayleigh/Nakagami-m	1	0.5	1	$m_2/2$

TABLE III. SIMULATION PARAMETERS

	Parameter	Value		Parameter	Value
1	α	0.9	6	$\sigma_{a,d}^2=\sigma_{c,d}^2$	0.04W
2	U	3	8	$\lambda_1=\lambda_2$	1W
3	P_s	5W	5	$\sigma_{a,r}^2=\sigma_{c,r}^2$	0.04W

V. NUMERICAL RESULTS AND DISCUSSION

In this section, through Monte-Carlo simulation, the simulation results are drawn to evaluate the performance of the non-linear energy harvester-capable DF relaying system when both links experience η–μ fading. To evaluate the performance of the considered system, we have numerous choices, for instance, optimal throughput, optimal outage probability, and optimal power-splitting ratio for numerous variable parameters, viz., fading parameters η and μ, energy transformation efficiency, noise variances, and antenna arrangements. In circumstances different from the considered, some set of parameters are placed as presented in Table III.

Fig. 1 reveals the outage probability against the power-splitting ratio, ρ. As expected, growing number of antennas, subsequently enhance the system performance. The outage probability lessens as power-splitting ratio, ρ, enlarges from 0 to ρ^* (i.e., an optimal-value of the power-splitting ratio when maximum throughput is obtained), and the outage probability enlarges as the ρ enlarges from its optimal-value to one.

Fig. 2 shows the throughput with respect to the power-splitting ratio for various saturation threshold power levels. As can be observed, the throughput increases as the level of the saturation threshold power P_{th} increases. Because enlarging the saturation threshold power level, lessens the possibility of saturation of the energy harvesting receiver, the energy harvesting receiver require additional power to harvest the energy.

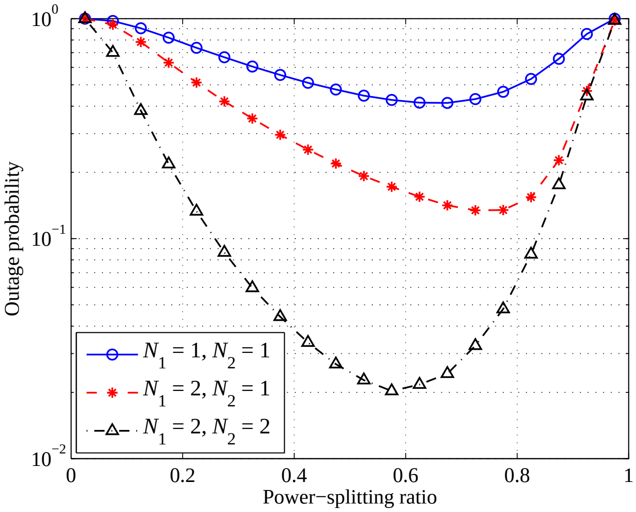


Fig. 1. Outage probability against power-splitting ratio for numerous antenna arrangements when $P_{th}=2$, $\eta_1=\eta_2=0.5$, and $\mu_1=\mu_2=1$.

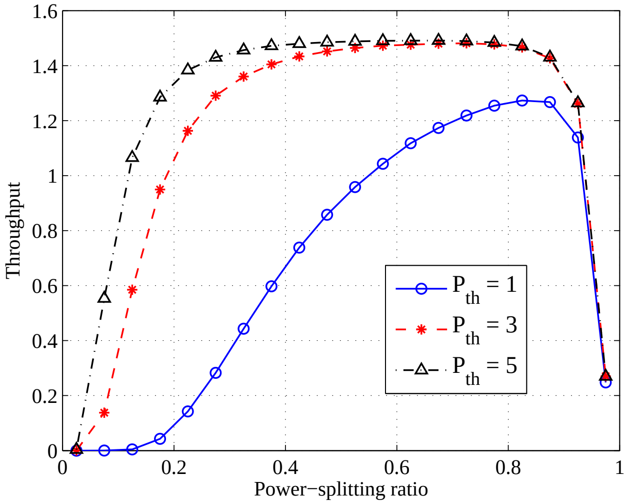


Fig. 2. Throughput as a function of the power-splitting ratio for various values of P_{th} when $N_1=N_2=2$, $\mu_1=\mu_2=1$, and $\eta_1=\eta_2=0.5$.

In Fig. 3, we showed the outage performance versus the power-splitting ratio for different values of the fading parameter μ (μ_1 and μ_2). From Fig. 3, we observe that the fading parameter μ has a notable influence on the system performance. It is also observe that the outage performance is lower at the smaller values of the power-splitting ratio and as well as at the higher values of the power-splitting ratio but outage performance is higher at the medium values the power-splitting ratio.

Fig. 4 presents the outage performance when we vary values of η (η_1 and η_2) and we set $N_1=N_2=1$, $\mu_1=\mu_2=1$, and $P_{th}=5$. We observe that the parameter η has a significant relation with the system performance. It is seen that the system achieves better performance when $\eta_1>\eta_2$ and the overall performance also increases with increasing the values of the fading parameter η.

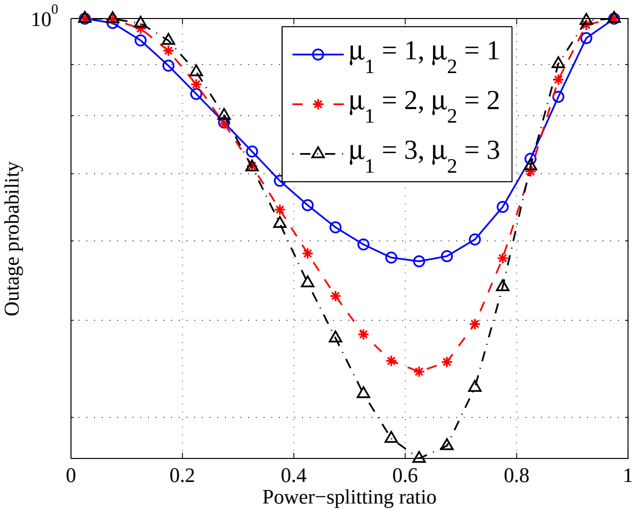


Fig. 3. Throughput versus power splitting ratio for various values of the parameter μ (i.e., μ_1 or μ_2) when $P_{th} = 2$, $N_1 = N_2 = 2$, and $\eta_1 = \eta_2 = 0.5$.

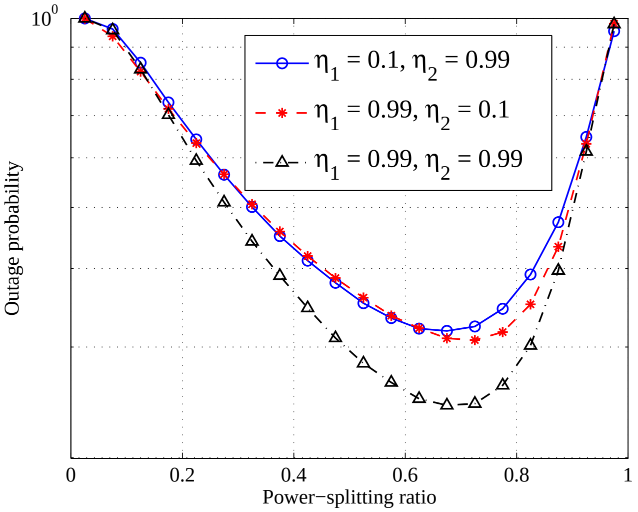


Fig. 4. Outage probability for various values of the fading parameter η (i.e., η_1 or η_2) when $N_1 = N_2 = 1$, $\mu_1 = \mu_2 = 1$, and $P_{th} = 5$.

VI. Conclusion

In this work, a wireless energy harvesting DF relaying network having multiple-antennas at the destination and source terminals that is based on a practical non-linear model is evaluated over η–μ fading channels. A PSR protocol and a non-linear energy harvester were considered at the relay node. The system performance were evaluated regarding the throughput and outage probability for numerous set of parameters, such as number of antennas and parameters, η and μ. The η−μ fading model has a few particular cases, viz., Rayleigh, Nakagami-m, and Hoyt. These results are general, therefore, new results were deduced for different fading conditions. The affected system performance by the insignificant saturation threshold power were minimized with the larger number of antennas.

Acknowledgment

This research is funded by R&D, Balochistan University of Engineering and Technology, Khuzdar, Pakistan.

References

[1] X. Lu, P. Wang, D. Niyato, and *et al.,* “Wireless networks with RF energy harvesting: A contemporary survey”*, IEEE Commun. Surv. Tuttorials., vol. 17, no. 2, pp. 757–789, 2015.*

[2] A. Hussain, Z. Ahemd, I. Ali, and S.H., Kim, *“Energy harvesting relaying network in a delay-tolerant transmission mode over $\kappa-\mu$ shadowed fading channels”*, International Journal of Computer Science and Network Security, vol. 18, no. 3, pp. 119–125, 2018.

[3] A. Hussain, Z. Ahmed, U. Rajpoot, and *et al.,* “Energy harvesting relaying network with hardware impairments in $\eta-\mu$ fading environment”, *International Journal of Computer Science and Network Security, vol. 18, no. 4, 2018.*

[4] A.A, Nasir, X. Zhou, S. Durrani, and R.A, Kennedy, *“Throughput and ergodic capacity of wireless energy harvesting based DF relaying network”*, proc. IEEE ICC 2014, Sydney, Australia, June 10–14, 2014.

[5] O.S. Badarneh, F.S. Almehmadi, I.S. Ansari, and X. Yang, *“Wireless energy harvesting in cooperative decode-and-forward relaying networks over mixed generalized $\eta-\mu$ and $\kappa-\mu$ fading channels”*, Transation on Emerging Telecommunication Technology, DOI: 10.1002/ett.3262, pp.1–18, 2017.

[6] A. Hussain, S.-H. Kim, and S.-H, Chang, *“Dual-hop variable-gain AF relaying with beamforming over κ−μ shadowed fading channels”*, proc. IEEE GLOBECOM 2016, Washington DC, USA, December 6–8, 2016.

[7] A. Hussain, S.-H. Kim, and S.-H. Chang, *“Non-linear energy harvesting relaying with beamforming and hardware impairments in κ−μ shadowed fading environment”*, Transations on Emerging Telelecommunicaiton Technology, https://doi.org/10.1002/ett.3303, 2018.

[8] Y. Dong, M.J. Hossain, and J. Cheng, *“Performance of wireless powered amplify and forward relaying over Nakagami-m fading channels with nonlinear energy harvester”*, IEEE Commun Lett., vol.20, no.4, pp.672–675, 2016.

[9] T.M. Hoang, T.T. Duy, and V.N.Q. Bao, *“On the performance of non-linear wirelessly powered partial relay selection networks over Rayleigh fading channels”*, proc. 3rd NICS 2016, Danag, Vietnam, September 14–16, 2016,

[10] J. Zhang, and G. Pan *“Outage analysis of wireless-powered relaying MIMO systems with non-linear energy harvesters and imperfect CSI”*, IEEE ACCESS, vol.4, no., pp.7046–7053, 2016.

[11] A. Cvetkovic, V. Blagojevic, and P. Ivaniš, *“Performance analysis of nonlinear energy-harvesting DF relay system in interference-limited Nakagami-m fading environment”*, ETRI Journal, vol.39, no., pp.803–812, 2017.

[12] J. Zhang, G. Pan and Y. Xie *“Secrecy outage performance for wireless-powered relaying systems with nonlinear energy harvesters”*, Front. Inform. Technol. Electron. Engg., vol.18, no.2, pp.246–252, 2017.

[13] K. Wang, Y. Li, Y. Ye, and H. Zhang *“Dynamic power splitting schemes for non-Linear EH relaying networks: perfect and imperfect CSI”*, Proc. VTC2017-Fall, Toronto, Canada September 24–27, 2017,

[14] A. Hussain, S.-H. Kim, and S.-H. Chang, *“On the performance of dual-hop variable-gain AF relaying with beamforming over $\eta-\mu$ fading channels”*, IEICE Transactions on Communications, vol.E100.B, no.4, pp.619–626, 2017.

Web Scraper Revealing Trends of Target Products and New Insights in Online Shopping Websites

Habib Ullah(1), Zahid Ullah(2), Shahid Maqsood(3), and Abdul Hafeez(4)
Department of Computer Science and IT, University of Engineering and Technology, Jalozai Campus, Pakistan(1,4)
Department of Electrical Engineering, CECOS University of IT and Emerging Sciences, Peshawar, Pakistan(2)
Department of Industrial Engineering, University of Engineering and Technology, Peshawar, Pakistan(3)

***Abstract*—Trillions of posts from Facebook, tweets in Twitter, photos on Instagram and e-mails on exchange servers are overwhelming the Internet with big data. This necessitates the development of such tools that can detect the frequent updates and select the required information instantly. This research work aims to implement scraper software that is capable of collecting the updated information from the target products hosted in fabulous online e-commerce websites. The software is implemented using Scrapy and Django frameworks. The software is configured and evaluated across different e-commerce websites. Individual website generates a greater amount of data about the products that need to be scraped. The proposed software provides the ability to search a target product in a single consolidated place instead of searching across various websites, such as amazon.com, alibaba.com and daraz.pk. Furthermore, the scheduling mechanism enables the scraper to execute at a required frequency within a specified time frame.**

Keywords*—Django QuerySet (DQS); e-commerce; hamming distance algorithm (HDA); Levenshtein distance algorithm (LDA); scraper; scheduling mechanism

I. Introduction

E-commerce is a mechanism of selling/purchasing products/services over the Internet. It is like a virtual product store where products are available and customers can browse and add products/services to the shopping carts. The customers are required to complete the transaction requirements by filling the transaction form with the required information, such as complete address, number of products, and the credit card number. Following the successful completion of the transaction requirement, an e-mail notification is sent to the customer.

The commercial use of Internet has been increasing exponentially day by day. In the modern era, shopping over Internet is becoming a common trend [1]–[5]. The pervasiveness of e-commerce has enabled an increasing number of transactions over the internet and thus, framing a compelling case for more and more business to turn online. Behavior by teenagers and older adults using smart-phone-based online shopping is becoming significant [1]. Recent work shows that emails, social media and smart-phone based advertising have been an established medium between the customers and businesses. In order to create and maintain efficient communication between customers and business, e-marketing techniques has been developed [6].

Data scraping pertains to the process of extracting data from online files through computer scripts. Such extracted data exists in the form of tables and lists. The interface between the script and the Internet for extracting data is basically a set of commands, i.e., an application programming interface (API). These APIs can be trained and used to extract data for search results across a group of websites. Automating web searches and extracting data from multiple pages for search results, merely requires users to input search items rather than navigating and searching websites individually.

The use and impact of web scraping examples are enormous. Tracking of pricing activities between different competitors can be accomplished via web scraping on a single or group of websites with minimal human intervention. Similarly, web scraping has enabled efficient searching of multiple websites and an increased transparency in research (scholar.google.com). Web scraping is pretty common in academic databases, such as Scopus, web of science and Inspec. Information that is not readily portable, such as author list and the corresponding author information can be extracted efficiently using web scraping. Subsequently, automating web searches is beneficial in other scenarios where web search is time consuming.

From server's perspective that is hosting a website can endure a remarkable strain in case of scraping a bulk of pages from a single or scraping huge volume of pages across multiple websites in a short span of time. In spite of that, acquiring few thousands of search results through scraping can hardly have deleterious startle on the server's performance. In nutshell, web scraping provides a resource-efficient search and transparency with minimum additional efforts.

Individual buyers or small organization can benefit from open-source and free web scraper available over the Internet. Additional developments will make the web scraper even better and easier to use and a well-trained API will benefit the prevailing networks.

However, the customers have limited knowledge about the trends of the target products between different e-commerce websites, such as daraz.pk, alibaba.com, and olx.com. These websites often have different rates for the same product. Finding the best price for a given product thus becomes a daunting task due to a variety of shopping websites. Customers have to search different online websites manually in order to find an optimal price for a target product. Therefore, a specific tool is required that can show the trends of a particular product in online markets and e-commerce websites.

We propose a scarping algorithm for detecting marketing trends in online shopping websites. Specifically our contribution is the pioneer work on using web scraping for extracting best price of the target products from multiple websites rather than a single website. This work entails the products from the top online websites. When the user wants to buy a product, he/she can search the product in one consolidated website and the search results are pulled up from fabulous marketing websites in just one consolidated place. Instead of using datasets from Amazon and Google, etc. our scraping method can be adapted for a variety of other shopping websites.

Rest of the paper is organized as follows: Section II entails the related work. Section III captures modeling of the proposed system. Section IV discusses the results and Section V concludes the paper.

II. Related Work

Numerous scrapers have been written in various programming languages and frameworks are being used for retrieving web data, such as beautiful soup, scrapy, Java, and Ruby. Beautiful soup is used to extract banner ads from different websites [7]. The problem of keyword suggestion was implemented according to the keywords entered by the user using clustering bipartite advertiser-keyword graphs [8]. The clustering submarkets were recovered with the help of advertisers, depicting a usual bidding behavior and the sets of keywords with an affinity to the same market place.

The data collected from booking commercial and large apartments does not reflect the latest market activity and thus, hides the recent knowledge of rental markets in the US. Scraper has been designed and developed to bridge this gap [9]. Scraper has been developed to distill most important news from large amount of news data [10]. Methods have been developed to quantify and predict the feedback from customers on a given product. This can further help marketing and investors to refine their decision making for addressing customer requirements precisely [11]. Web bots have been used for modelling traffic patterns generated by different internet bots [12]. Neural Network has been used to detect buying or non-buying sessions from user sessions that involves only human intervention instead of those carried out by internet bots [13]

To avoid complexity, a simplified version of a scarper has also been implemented [14]. A framework was developed for scraping and retrieving the trendiness of YouTube content and viewers statistics−their watching time and shares [15]. Nonetheless, the YouTube APIs do not allow third parties to easily scrap such information. A framework naming YOUStatAnalyzer enables researchers to create their own data sets based on a variety of search criteria [15]. The framework has also the capability to analyze the created data sets for extracting useful features and distinguishing statistics.

The scheduling of jobs/tasks on processor is the most important and challenging task. Time slicing deals with the switching of context within the processor. However, space slicing specifies the ways for how to map processes onto the processor [16]. In order to achieve an optimal scheduling for processes, a general mathematical framework, resource task networks, was formulated [17]. In another scheme, scheduling of batch jobs based on first-come-first-serve was discussed on large parallel processor [18]. Using gang scheduling, initiated only by embarrassingly parallel jobs, helps preventing severe fragmentation. Furthermore, operating system support was provided in order to provide robust parallelism in addition to hardware-level parallelism [19], [20].

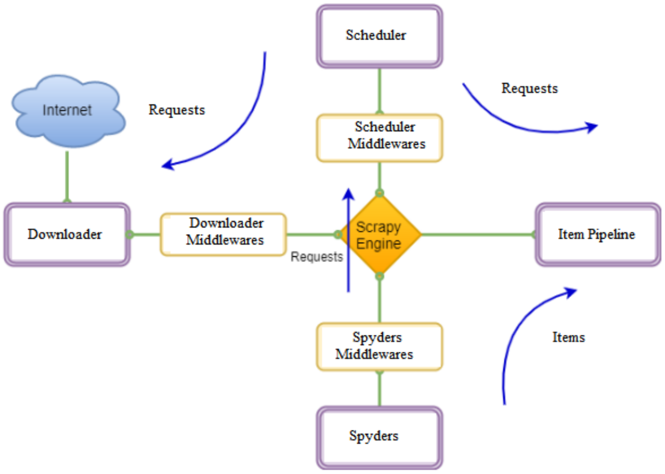


Fig. 1. The proposed system model. The proposed system contains five basic units, which are scheduler, downloader, scrapy engine, item pipeline, and spypder.

Interestingly, a gamut of searching algorithms are more efficient than straight forward searching techniques such as Hamming, Needleman Wunsch, Smith Waterman, Knuth Morris Pratt, Boyer Moore, and RabinKarp. A comparative study of different types of string matching algorithms, observation on their time and space complexities, and corresponding efficiency and performance has been tested with different biological sequences [21].

The entire samples from a finite number of keywords in a given string of text were computed by a simple and efficient algorithm [22]. An algorithm has been designed by constructing a finite state pattern-matching machine from the keywords, which was further used to process the text string into a single pass. The time taken by constructing the pattern matching machine was proportionate to the total sum of spans of the keywords. The speed up achieved by the algorithm was used to accelerate the search in the library bibliography by a factor of $5 - 10$.

III. Modeling of the Proposed System

The proposed system model, given in Fig. 1, has five basic components and are discussed here. In order to build a rapid prototype, the system has been implemented in Python. The implementation modules, details and their relationships are delineated in Fig. 1. As shown, the system is composed of the following four modules: 1) downloader; 2) scheduler; 3) item pipeline; and 4) spyders.

1) **Scheduler**: This component is responsible for scheduling all the requests and responses in Scrapy. This further queues up all the requests that are received from the engine and passes these to the downloader.
2) **Downloader**: The job of this module is to download all the required pages that are passed by the Scrapy engine and to send the downloaded pages back to the Scrapy engine through download middleware.

3) **Scrapy engine**: Scrapy engines are used to scrap large scale data. The heart of the system is the Scrapy engine. The goal is to control all the processes in Scrapy and the entire requests and responses passed through it from one component to another.
4) **Item pipeline**: The functionality of this module is to filter the data. It validates the data and checks to see whether the data is scrapped and also, clean data.
5) **Spyder**: An abbreviation for Scientific Python Development Environment. Spyder integrates important libraries, such as NumPy, Matplotlib, and SciPy and is an open-source tool for scientific programming. It is a class of python that defines how to extract the required data and the target page to be crawled [23]. It generates a request that will be sent to downloader through Scrapy engine. Items are scraped and stored whenever the desired response is received to the spyder.

The subsequent sub sections explains web scraping, scheduling of scrapers, and the search mechanism used for target products.

A. Web Scraping

Web scraping is also called web data extraction. It is a process, which is used to extract large amount of data from websites and to store the extracted data into the local storage in different formats. Web scraping is used for different purposes such as research, analysis of market and comparison of price, collection of opinion of public in business, jobs advertisements, and collection of contact detail of required business.

The data of websites are only shown in the web browser. If we need to check all the data of any website, we cannot do this without going to every page and cannot copy the data for the personal use owing to the longer time it takes to be copied. Web scraping is technique that provides to copy the data from websites in a reasonable amount of time instead of copying manually. It automates the manual copying process using a web crawler and bot.

The web scraping software is connected to the website through hypertext transfer protocol (HTTP). It fetches the page and extracts data from that page and swaps among the multiple pages of websites according the requirement to extract data. When the data is extracted, it will then be exported into different format such as CSV and JSON according to the needs.

B. Scheduling of Scrapers

It is a method in which specified, arranged work or processes are assigned to the resources to complete it. Virtual elements of the work such as threads and processes are scheduled to hardware resources like processors and expansion cards. The goals of the scheduler is to keep the entire resources busy and share them effectively in order to maximize the CPU usage and quickly switch processes onto CPU for time sharing to get a desired output.

Algorithm 1 Hamming Distance Algorithm

```
Input: $S_1$ and $S_2$
Output: dist
if $S_1 \neq S_2$ then
    Raise value error
    statements...
end if
dist = sum($S_1$(x,y != $S_2$(x,y) for $S_1$(x,y) and $S_2$(x,y) in zip($S_1$, $S_2$))
return dist
```

Algorithm 2 Levenshtein Distance Algorithm

```
Input: $S_1$ and $S_2$
Output: Levenshtein distance (LD)
if len($S_1$) $\neq$ 0 then
    return len($S_2$)
end if
if len($S_2$) $\neq$ 0 then
    return len($S_1$)
end if
if len($S_2$)- 1 = 0 then
    LD = 0
else
    LD = 1
end if
$A_1$ = LD($S_1$, len($S_2$) $-$ 1, $S_2$, len($S_1$)) + 1
$A_2$ = LD($S_1$, len($S_2$), $S_2$, len($S_1$) $-$ 1) + 1
$A_3$ = LD($S_1$, len($S_2$) $-$ 1, $S_2$, len($S_1$) $-$ 1) + LD
Minimum($A_1$,$A_2$,$A_3$)
```

Operating system entails a variety of schedulers – long-, medium-, and short-term schedulers. For the purpose to schedule running scrapers, a scraper is using Django-celery. Celery is an asynchronous task queue and supports distributed message passing. Task queues are used in order to distribute the workload among the given processors.

Input to the celery daemon is the task orders and then the execution of corresponding tasks is performed sequentially in order to complete the entire job, which ensures that none of the tasks is lost; even if the system is over burdened. In our proposed work, celery works as a job replacement tool that can be controlled by Django admin interface.

C. Search Mechanism for Target Products in Websites

The goal of string matching algorithm is to find one or more than one patterns within a larger string or text. In this research work, we have implemented two types of searching mechanisms – Hamming distance and Levenshtein distance techniques.

1) Hamming distance: The Hamming distance counts the difference at positions between any two strings (S_1 and S_2) of an identical length. In other words, this means the least number of interchanges required to convert a string S_1 into S_2, as illustrated in Algorithm 1.

2) Levenshtein distance: There are three operations that are used to transform one string into another in order to find the similarities between strings [24]–[27]. The Levenshtein distance between any two strings *C* and *D* of length |C| and |D|, respectively can be formulated by $\text{lev}_{C,D}$ (|C|,|D|) as given

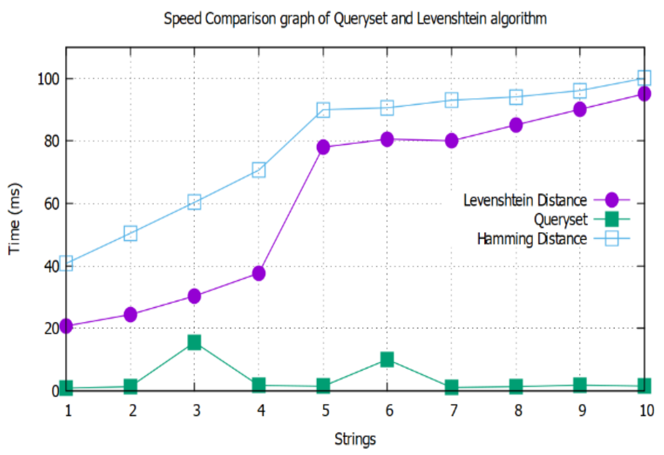


Fig. 2. The speed comparison of Django QuerySet (DQS), Levenshtein distance algorithm (LDA) and Hamming distance algorithm (HDA). DQS takes lesser time compared to LDA and HDA in string comparison.

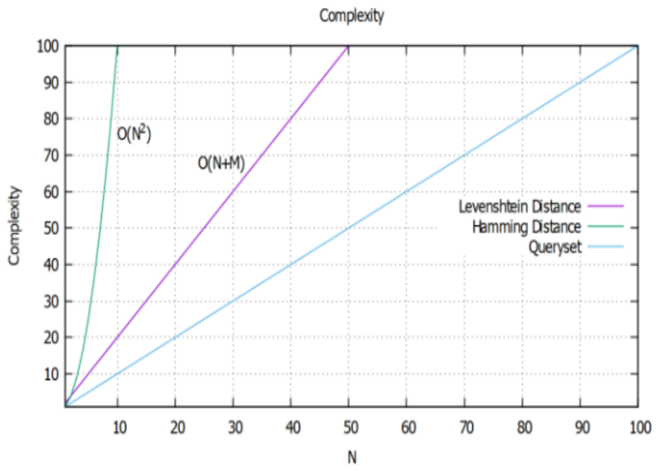


Fig. 3. Complexity curve of Levenshtein distance, Hamming distance and Queryset. As clear from the graph that HDA is more complex compared to LDA while the LDA complexity is much higher compared to DQS.

in (1).

$$lev_{(C,D)}(i,j) = \begin{cases} max(i,j) & \text{if } min(i,j) = 0 \\ min \begin{cases} lev_{(C,D)}(i-1,j)+1 \\ lev_{(C,D)}(i,j-1+1) \\ lev_{(C,D)}(i-1,j-1)+1_{(C_i \neq D_j)} \end{cases} & \text{otherwise} \end{cases} \tag{1}$$

Where ($C_i \neq D_j$) is the indicator function that string *C* and *D* are entirely different and hence, equal to 0. Furthermore, C_i = D_i means equal to 1, where $lev_{(C,D)}(i,j)$ corresponds to the span between the initial character of of *C* and *D*.

It should be noted that the first element in the minimum corresponds to the deletion from *C* to *D*, while the second pertains to the insertion and the third is related to either similarity/dissimilarity of the corresponding symbols. Algorithm 2 shows the Levenshtein distance algorithm.

IV. Experimental Setup and Discussion on Results

A. Experimental setup

The experimental setup entails Toshiba laptop with a main memory of 8 GB and a processing capability of 2.5 GHz with a storage capacity of 1 TB. The operating system used was Microsoft Windows 10. The scrapers were written in Python using Django framework.

We used Scrapy framework for the web scraping to write crawler. The crawler is configured and tested to scrap data from different websites. We scraped 15000 products in approximately 11 minutes. Finally, we compared three types of searching algorithms− Django Queryset, Hamming distance, and Levenshtein distance [28]. We compared these techniques based on the speed, complexity, and the throughput.

B. Discussion on Performance

The speed comparison of Django QuerySet (DQS), Levenshtein distance algorithm (LDA) and Hamming distance algorithm (HDA) is shown in Fig. 2. DQS consumes lesser time than LDA and HDA in string matching process. Each query is executed ten times and then averaged. The average time taken to execute a query has a greater value for LDA and HDA than DQS. The reason for preferring LDA over HDA and DQS is the large number of comparisons.

The LDA filters out the most accurate results for searching algorithm. As the string size increases, the time taken to process also increases in both HDA and LDA. The reason is that both algorithms perform complex operations of insertion and deletion during matching process against the database the database. Increasing string size is proportional to the number of operations.

The complexity of Levenshtein distance algorithm (LDA), Django QuerySet (DQS), and Hamming distance algorithm (HDA) is captured in Fig. 3. It is obvious that HDA is more complex than LDA, while the LDA complexity is greater than DQS. The complexities of LDA, HDA and DQS is O(N+M), O(N^2) and O(N), respectively.

It means that the complexity of LDA is higher than that of the DQS. LDA performs complex operations of insertion, deletion, and substitution; however, DQS only checks whether the database values contain the search phrase. HDA measures the least number of errors required to alter one string into another.

Fig. 4 shows the throughput achieved by LDA, HDA, and DQS. The throughput of DQS is greater than both LDA and HDA; however, LDA is preferred owing to its accuracy (approximately 70 %) and versatility. Experiments have shown that the accuracy of HDA is about 81 %, which is better than LDA; however, LDA is still preferred because the complexity of O(N^2) is higher compared to O(N+M).

Moreover, the major problem with HDA is that the length of both of the strings must be the same; otherwise, the algorithm will not work. The LDA performs more operations on strings than HDA and DQS; it filters most accurate result for our matching algorithm than DQS. For a query set of size

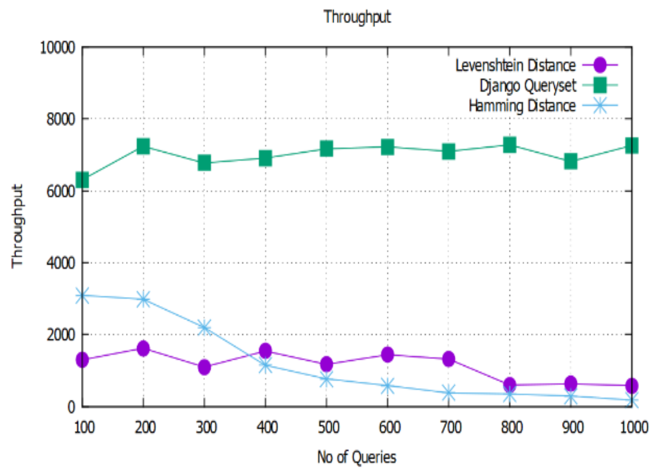


Fig. 4. Throughput comparison among Levenshtein distance algorithm (LDA), Hamming distance algorithm (HDA), and Queryset (DQS). The throughput of DQS is higher compared to both LDA and HDA; however, LDA is preferred because of its accuracy (approximately 70 %) and versatility.

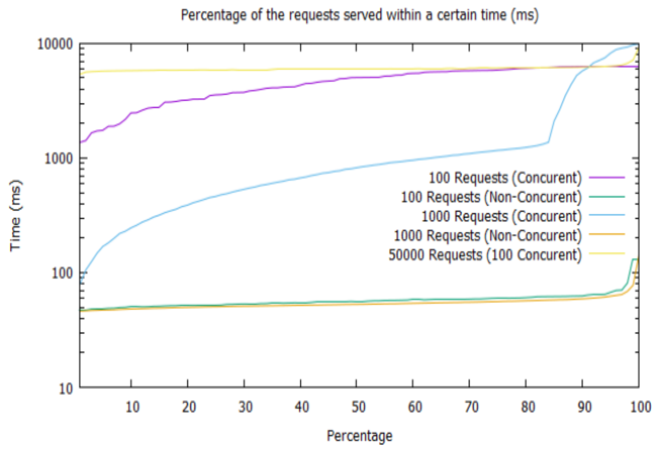


Fig. 5. Time taken to serve an increasing number of requests. A large number of users requests are taken such as 100, 1000, and 5000 against both concurrent and non-concurrent requests.

200, all of the three algorithms produce higher value due to the fact that system was busy performing other tasks as well.

The server response against user requests is delineated in Fig. 5. In our proposed system, two types of user requests were evaluated: concurrent and non-concurrent requests. The response of the web server hosted on a local server was evaluated against user requests.

In order to measure the percentage of the requests served with a certain time frame, the parameter of interest was concurrency. It is observed that the requests having concurrency were served quickly. Howbeit, significant system resources will be allocated in case large number of requests is arrived at once.

Fig. 6 shows the results of simulation for a span of one-minute using the Web Server Stress Tool in order to measure the data transferred, the system memory used, and the CPU load. The network traffic is the amount of data transferred over the network at a given point of time.

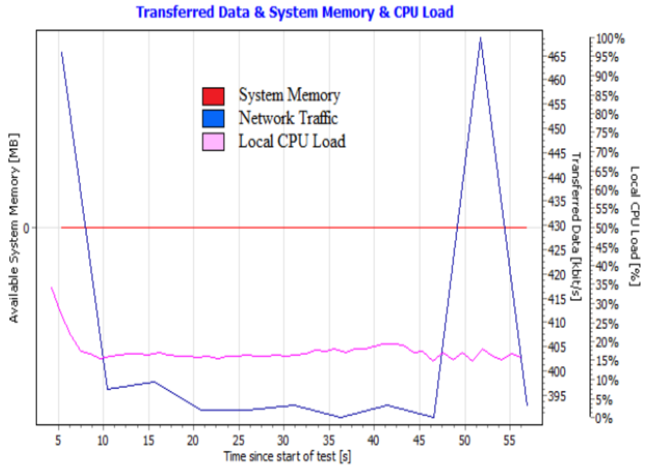


Fig. 6. Results of simulation for a span of one-minute using the Web Server Stress Tool for transferred data, system memory, and CPU load.

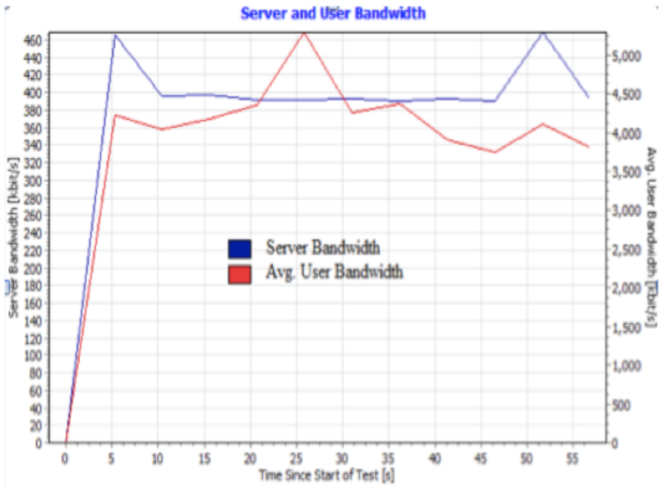


Fig. 7. Server and user bandwidth. Whenever the requests are arrived at the system, there is an increase in bandwidth requirements of both the server and user, which means the server bandwidth is 460 Kb/s and the user bandwidth is 370 Kb/s.

The peaks in network traffic was observed at time stamp 5 sec and at 52 sec. That is, the traffic reached up to 456 Kbps, mainly because of higher number of requests. Comparatively, the network traffic was lower during time span of 10 sec 45 sec. The reason is lower number of requests generated against the server.

The benchmark results of server and user bandwidth in Kbps are shown in Fig. 7. In the beginning as there is zero number of requests generated against the server, therefore, the average bandwidth of user and server is 0. However, with an increasing number of requests received by the system, the bandwidth of the server as well as user goes higher. The peak bandwidth achieved by the server and user requests is 400 Kbps and 370 Kbps, respectively.

V. Conclusions and Future work

In this work, we proposed a filtering web pricing system that exploits web scraping techniques in order to extract trends and suggest best price of a target product from top of the line commercial websites such as amazon.com, alibaba.com, and daraz.pk. The designed framework incorporates Scrapy framework for web crawling and scraping. Celery is used to schedule scraper in order to analyze the crucial pages in the target websites and distill the required information against a given product.

For the sake of string matching between the users typed search and the online products, Levenshtein, Hamming, and QuerySet are used. The results show an improved accuracy and an accelerated response for retrieving search results while using Levenshtein distance. Albeit, throughput of QuerySet is much higher than Levenshtein and Hamming method. To the best of our knowledge, this is the first attempt to filter knowledge about best pricing of a product from top of the line websites.

As of future work, we aim to enable the proposed framework to suggest relevant and non-relevant items based on a factor k. Furthermore, future research directions include integration of the proposed work into social media, such as Google and facebook to suggest best prices about the products based on the user preferences. Ultimately, the goal is to enable users to search for the best price from top of the line website, whether it may be finding best and cheap hotels, or finding the cheapest airfare while traveling or finding the best deal for jewelry at wedding ceremonies and the list continues to increase.

Acknowledgment

The authors would like to offer thanks to all the anonymous reviewers for their nice and valuable suggestions, which increased quality of the work.

References

[1] S.-M. Kuoppamäki, S. Taipale, and T.-A. Wilska, “The use of mobile technology for online shopping and entertainment among older adults in finland,” *Telematics and Informatics*, vol. 34, no. 4, pp. 110–117, 2017.

[2] J. Ogilvie, K. Lindsey, K. Reynolds, and W. M. Northington, “Examining reactive customer engagement strategies in online shopping cart abandonment: A regulatory fit perspective,” in *Rediscovering the Essentiality of Marketing*. Springer, 2016, pp. 755–756.

[3] W. Hong, K. Y. Tam, and C. K. B. Yim, “E-service environment: Impacts of web interface characteristics on consumers online shopping behavior,” in *E-Service: New Directions in Theory and Practice*. Routledge, 2016, pp. 120–140.

[4] M. D. Griffiths and J. Parke, “The social impact of internet gambling,” *Social Science Computer Review*, vol. 20, no. 3, pp. 312–320, 2002.

[5] M. Griffiths, “Internet addiction: does it really exist?” 1998.

[6] H. Kresh, A. Laible, M. Lam, and M. Raisinghani, “Online advertising: Creating a relationship between businesses and consumers,” in *Global Business Value Innovations*. Springer, 2018, pp. 47–61.

[7] E. Vargiu and M. Urru, “Exploiting web scraping in a collaborative filtering-based approach to web advertising,” *Artificial Intelligence Research*, vol. 2, no. 1, p. 44, 2012.

[8] J. J. Carrasco, D. C. Fain, K. J. Lang, and L. Zhukov, “Clustering of bipartite advertiser-keyword graph,” 2003.

[9] G. Boeing and P. Waddell, “New insights into rental housing markets across the united states: web scraping and analyzing craigslist rental listings,” *Journal of Planning Education and Research*, vol. 37, no. 4, pp. 457–476, 2017.

[10] F. Hamborg, N. Meuschke, and B. Gipp, “Matrix-based news aggregation: exploring different news perspectives,” in *Digital Libraries (JCDL), 2017 ACM/IEEE Joint Conference on*. IEEE, 2017, pp. 1–10.

[11] S. Wang, “From social media to innovation and marketing intelligence: A simulation to forecast online review and rating performance,” *Journal of Digital & Social Media Marketing*, vol. 4, no. 3, pp. 251–262, 2016.

[12] G. Suchacka and D. Wotzka, “Modeling a session-based botsarrival process at a web server,” in *Proceedings of the 31st European Conference on Modelling and Simulation (ECMS17)*, 2017, pp. 605–612.

[13] G. Suchacka and S. Stemplewski, “Application of neural network to predict purchases in online store,” in *Information Systems Architecture and Technology: Proceedings of 37th International Conference on Information Systems Architecture and Technology–ISAT 2016–Part IV*. Springer, 2017, pp. 221–231.

[14] R. B. Penman, T. Baldwin, and D. Martinez, “Web scraping made simple with sitescraper,” 2009.

[15] M. Zeni, D. Miorandi, and F. De Pellegrini, “Youstatanalyzer: a tool for analysing the dynamics of youtube content popularity,” in *Proceedings of the 7th International Conference on Performance Evaluation Methodologies and Tools*. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 2013, pp. 286–289.

[16] R. M. Bryant, H.-Y. Chang, and B. S. Rosenburg, “Operating system support for parallel programming on rp3,” *IBM Journal of Research and Development*, vol. 35, no. 5.6, pp. 617–634, 1991.

[17] G. Schilling and C. Pantelides, “A simple continuous-time process scheduling formulation and a novel solution algorithm,” *Computers & Chemical Engineering*, vol. 20, pp. S1221–S1226, 1996.

[18] U. Schwiegeishohn and R. Yahyapour, “Improving first-come-first-serve job scheduling by gang scheduling,” in *Workshop on Job Scheduling Strategies for Parallel Processing*. Springer, 1998, pp. 180–198.

[19] T. Jones, S. Dawson, R. Neely, W. Tuel, L. Brenner, J. Fier, R. Blackmore, P. Caffrey, B. Maskell, P. Tomlinson *et al.*, “Improving the scalability of parallel jobs by adding parallel awareness to the operating system,” in *Proceedings of the 2003 ACM/IEEE conference on Supercomputing*. ACM, 2003, p. 10.

[20] A. Gupta, A. Tucker, and S. Urushibara, “The impact of operating system scheduling policies and synchronization methods of performance of parallel applications,” in *ACM SIGMETRICS Performance Evaluation Review*, vol. 19, no. 1. ACM, 1991, pp. 120–132.

[21] P. Pandiselvam, T. Marimuthu, and R. Lawrance, “A comparative study on string matching algorithms of biological sequences,” in *International Conference on Intelligent Computing*, 2014, pp. 1–5.

[22] A. V. Aho and M. J. Corasick, “Efficient string matching: an aid to bibliographic search,” *Communications of the ACM*, vol. 18, no. 6, pp. 333–340, 1975.

[23] M. Schrenk, *Webbots, spiders, and screen scrapers: A guide to developing Internet agents with PHP/CURL*. No Starch Press, 2012.

[24] V. I. Levenshtein, “Binary codes capable of correcting deletions, insertions, and reversals,” in *Soviet physics doklady*, vol. 10, no. 8, 1966, pp. 707–710.

[25] L. Yujian and L. Bo, “A normalized levenshtein distance metric,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 29, no. 6, pp. 1091–1095, 2007.

[26] B. Gujral, H. Khayrallah, and P. Koehn, “Translation of unknown words in low resource languages,” 2016.

[27] L. C. Guillen, J. Domenico, K. Camargo, R. Pinheiro, and C. Coeli, “Match quality of a linkage strategy based on the combined use of a statistical linkage key and the levenshtein distance to link birth to death records in brazil.” *International Journal for Population Data Science*, vol. 1, no. 1, 2017.

[28] P. E. Black, “Dictionary of algorithms and data structures,” Tech. Rep., 1998.

Marine Engine Room Alarm Monitoring System

Fault Detection and Monitoring Technique via PLCs and SCADA-based System

Isaac Tawiah, Usman Ashraf, Yinglei Song, Aleena Akhtar
School of Electronics and Information Science,
Jiangsu University of Science and Technology,
Zhenjiang, 212003, China

***Abstract*—Alarms affect operations in most part of the ship. Their impact on modern Engine Control Room operations is no less significant. The state of an alarm system serves as an indication of the extent to which the ship's operations are under management control. Thus, the design of efficient and reliable alarm monitoring system is vital for safe and sound operations. Although several design techniques have been proposed, all the proposed design methods employ sophisticated and expensive approaches in resolving alarm issues. In this paper, a cheap, yet reliable and efficient alarm design method for engine room device monitoring is presented. The design method employs PLCs and SCADA-based system and adopts certain basic design requirements of alarm monitoring system presented in literary works. Reasons for such a design method are highlighted, and the programming platforms for the design are given. The strengths and weaknesses of some design methods presented in some published works are reported and solutions to such problems are proposed. The proposed design technique, including fault diagnostic algorithm, have been subjected to real-time online testing at the shipyard, specifically Changjiang Waterway Bureau, China (ship name–Ning Dao 501). The testing results proved that this design technique is reliable, efficient and effective for online engine control room device monitoring.**

Keywords*—*Alarm monitoring system; engine control room; OPC communication; PLCs; SCADA systems

I. Introduction

Reports indicate that alarm monitoring issues are the biggest single issue affecting safety in modern engine control rooms [1], [2], [3], [4]. Reported issues include distractions from nuisance alarms, confused and unclear alarm messages, inappropriate use of alarms to convey non-critical information, long standing alarm lists, and difficulty in alarm interpretations. Moreover, several works have been proposed to mitigate these issues. In [5], Zaghloul presented a practical design of an open architectural ship alarm control and monitoring systems via SCADA system. In Zaghloul's proposal, a collection of standard software (Human Machine Interface) is used to provide a SCADA central host of all the statuses of the field machinery. The shortfall of this design method is that, it only performs monitoring purposes, and ignores basic error detection algorithms. In resolving such design issues, Lee et al. [6], [10], [11] proposed basic design requirements of alarm systems, whereas [7], [8] propose fuzzy logic algorithms to tackle fault detection issues. A SCADA based system for monitoring a central heating and power plant using Siemens equipment and software "Process Control System 7" (PCS7) was proposed in [9]. This design technique employs redundant servers and web-based applications via OPC and web servers to provide remote actions and uninterrupted monitoring. Aside the sophisticated and expensive nature of the method proposed in [9], it is incapable of reporting errors that incurred during device malfunctions. Problems of unreliable communication resulting from complex cabling have also been reported in [12]. In resolving these problems, D'ana [13] discusses a communication based model over GSM network to remotely monitor and control PLC based processes. The design approach is built on industrial PLCs, where each PLC is integrated with processors capable of connecting to a network via GSM modem. Dana's intent was to provide system users with a back-up communication mechanism in case of a network failure. Even though capitally intensive, the idea proved to achieve the same functionality as Profinet but at a higher bandwidth (10/100 Mbps). Wang et al. [14], also presented a communication design method based on CAN bus and RS485. All the above reported solutions employ sophisticated and expensive approaches in resolving the above highlighted issues.

In this paper, the basic design requirements of alarm monitoring system presented in [6] are adopted, and the communication methods of [13] are used to resolve these issues. However, instead of using the more sophisticated technique proposed in [13], the TCP/IP Ethernet based network, via PLC is used for PC communication. In addition, the fault diagnostic algorithm employed in this paper is different from those presented in literature [7]. Most of the methods reviewed in [7] employ two or more sensors to monitor a single parameter (e.g. Temperature). However, while cost is always an important factor, process and personnel safety are also important. Specifying multiple sensors to gain comparative readings, with the aim of obtaining a voting logic leads to added capital cost, added cost of spare sensors to maintain capability while recalibrating units, and perhaps even additional staff personnel to perform calibrations and periodic checks [15]. Thus, with the view of cutting down cost, an error control algorithm is attached to just a single input sensor. From an electrical perspective, it is also appropriate to protect circuits in order to ensure safe operation of the PLCs and field I/O devices. The protection of circuits is often done by fusing each I/O module with a single fuse [16], [17]. In this paper, the technique where each I/O point is fused is used to guarantee that one fault only disrupts the affected point [18].

II. Overview

This paper presents a design method for real-time monitoring of engine room alarm systems, using PLCs and SCADA-base system. Siemens S7-1200 Series PLC (CPU 1214C DC/DC RLy) is used. Whilst this paper is targeting this type of PLCs, the approach used here, can be transferred to other

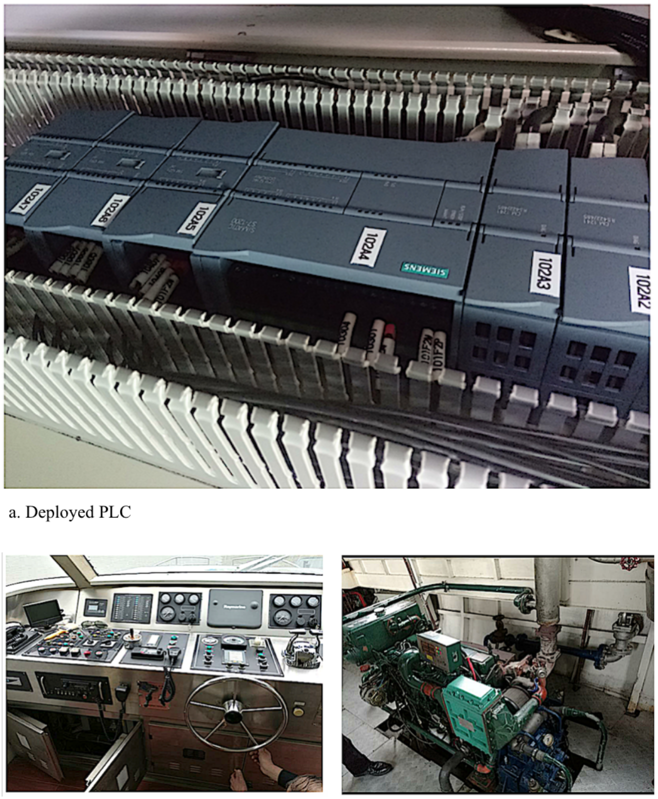
Vendor and System types. Communication between PLC and PC is achieved through Siemens Totally Integrated Automation (TIA) portal software, via Ethernet TCP/IP protocol. This allows simulations to be performed through Ethernet communication with the PC, even before real-time test is carried out in the control room. OLE for Process Control (OPC) is used as a communication server between TIA and the SCADA GUI (NI LabVIEW) software. The data collected is the parameters of the various monitored field machinery. Thirty-Five monitored functions, under five main parameters were considered, including Temperature, Pressure, Voltage/Current, Speed, and Depth of Liquids. Data acquisition was made possible through sensors connected to the input channels of the PLC modules. The designed alarm monitoring system, including the fault diagnostic algorithm was subjected to real-time online testing at the shipyard, specifically Changjiang Waterway Bureau, China (ship name–Ning Dao 501).

Fig. 1a. shows the screenshot of the deployed PLC, showing the signal modules (from slot 102A1 to 102A3), the CPU (slot 102A4), and the signal modules (from slots 102A5 to 102A8). This series of PLC utilizes a maximum of three communication modules for serial communication, and up to eight signal modules for input/output expansion. In this paper, all the three communication modules are utilized, as labeled 102A1 to 102A3 in Fig. 2. Four of the eight available signal modules are used. This constitutes the labels 102A5 to 102A8. Fig. 3 shows the electrical network of the PLC workstation. The workstation constitutes both the CPU and all other modules connected to it. The series of communication modules used is the CM1241 series with port number 1RS485.

The Uninterrupted Power Supply (UPS) used here is a separate standalone, and thus additional pieces of power supplies are needed to convert the correct voltages to power various devices. In the case of the deployed PLC, the maximum allowable input voltage is 24V DC. A voltage of 220V AC of 2.5A, 50/60Hz is stepped-down to 24V DC at 8.8A to power the PLC. As seen from Fig. 4, each power supply connection is connected by a 2A fuse for circuit protection. The industrial PC, the monitor, the field sensors, all have their own operating voltages. A 220V of 1KW AC power source supplies power to other secondary power supplies, both in the engine room and in the ship's cabin across a 2 x 1.5mm and a 2 x 2.5mm wires.

III. Design Methodology

The main idea is to design a reliable and effectively cheap alarm monitoring and fault detection system. The general architectural design employed in this work is summed up in the diagram of Fig. 5. Sensors attached to field devices receive and transmit data to the PLC inputs via communication cables. Unlike the various design methods presented in the literature where two or more sensors are used to access a single input parameter, the proposed technique employs a single sensor for parameter assessments. Data-flow from the PLC to PC proceeds through TCP/IP communication. Within the Siemens TIA portal software, an efficient error control algorithm is developed to provide alarm signals of the malfunctions. This malfunctions and all other signals are made accessible to the operator on a GUI through OPC server station. In summary, the following contributions are made: the technique serves two



a. Deployed PLC

b. Control Room

c. Main Engine

Fig. 1. Screenshots of the deployed PLC workstation, and some Engine room machinery.

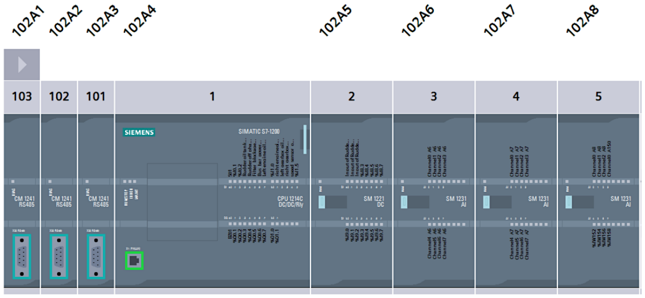


Fig. 2. PLC 1214C workstation.

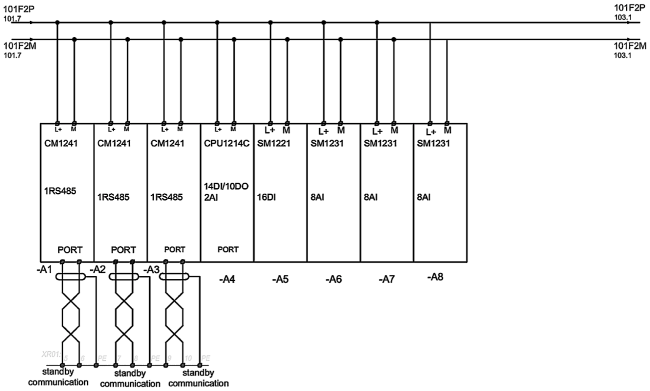


Fig. 3. CPU 1214C workstation wiring diagram.

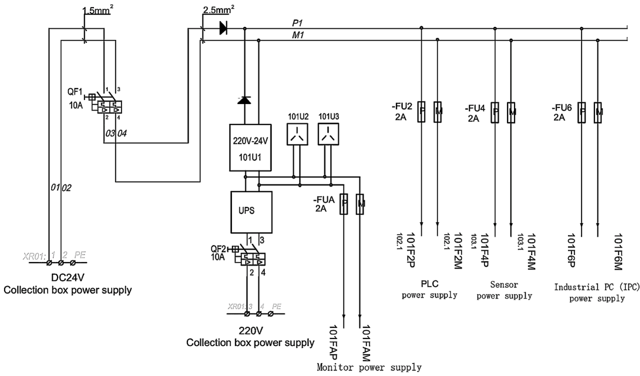


Fig. 4. Schematic of the system power supply.

purposes – both monitoring and fault diagnosis; it is capitally cheap, yet a reliable technique; all the communication proceeds through TCP/IP, through Ethernet cables; and it is able to store alarm record list for future fault identification. The details of the design technique are presented in the subsequent sections.

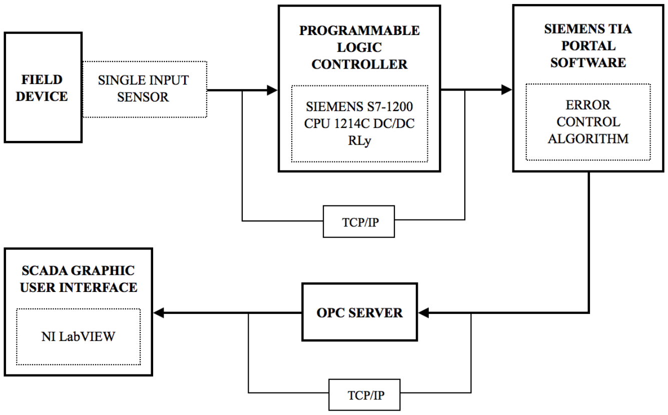

FIELD DEVICE
SINGLE INPUT SENSOR
PROGRAMMABLE LOGIC CONTROLLER
SIEMENS S7-1200 CPU 1214C DC/DC RLy
SIEMENS TIA PORTAL SOFTWARE
ERROR CONTROL ALGORITHM
TCP/IP
SCADA GRAPHIC USER INTERFACE
NI LabVIEW
OPC SERVER
TCP/IP

Fig. 5. General architectural design flowchart.

IV. Monitoring Functions

A. Analog Functions

Analog sensors are connected to the PLC via analog modules without additional amplifiers. The input modules SM 1231 AI from slot 102A6 to 102A8 (see Fig. 2) are used for the input analog sensors. Each of the modules has eight channels counted from 0 to 7, with a selected current or voltage range of either 4mA-20mA or -10V to 10V respectively. Each channel can store integer data ranging from 0 to 32767, however, all integer limits in this work are from 0 to 27648. In total, nineteen analog sensors are monitored (see Fig. 6 to 8), including voltages, currents, temperatures, and pressures. In cases where levels of fluids are to be measured, the formula $p = \rho hg$ is used to convert the measured pressure value to height. As seen from Fig. 6 to 8, each of the channels is protected by a 0.1A fuse. Each of the input is labeled FUxx, where FUxx represents the fuse number. For instance, the analog input "Power station Voltage" is labeled FU01, with a 0.1A fuse protection. This mode of identification provides easier access to faulty sensors, as this identification scheme is used in the alarm logical programming.

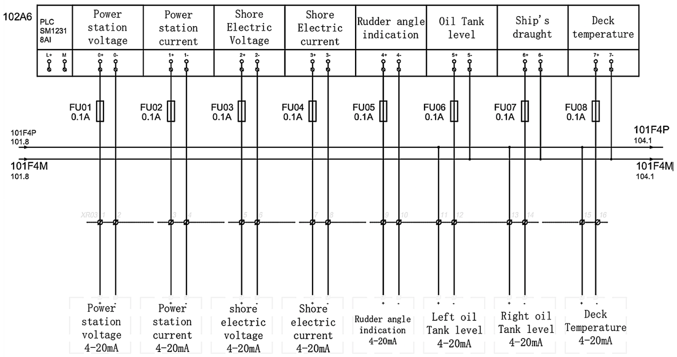


Fig. 6. Analog signals (signal module SM 1231, slot 102A6).

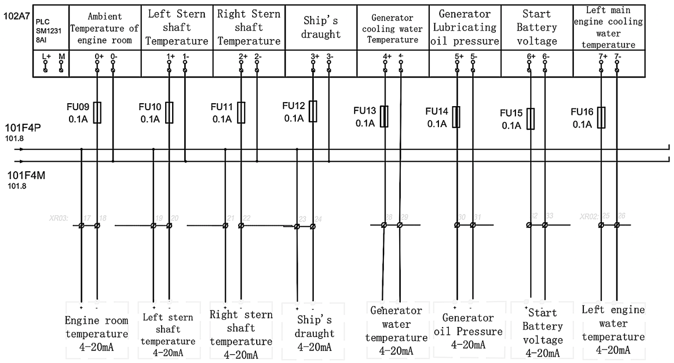

102A7
PLC SM1231 8AI
Ambient Temperature of engine room
Left Stern shaft Temperature
Right Stern shaft Temperature
Ship's draught
Generator cooling water Temperature
Generator Lubricating oil pressure
Start Battery voltage
Left main engine cooling water temperature
L+ M
0+ 0-
1+ 1-
2+ 2-
3+ 3-
4+ 4-
5+ 5-
6+ 6-
7+ 7-
FU09 0.1A
FU10 0.1A
FU11 0.1A
FU12 0.1A
FU13 0.1A
FU14 0.1A
FU15 0.1A
FU16 0.1A
101F4P 101.8
101F4M 101.8
Engine room temperature 4-20mA
Left stern shaft temperature 4-20mA
Right stern shaft temperature 4-20mA
Ship's draught 4-20mA
Generator water temperature 4-20mA
Generator oil Pressure 4-20mA
Start Battery voltage 4-20mA
Left engine water temperature 4-20mA

Fig. 7. Analog signals (signal module SM 1231, slot 102A7).

B. Digital Functions

Digital sensors are connected to the PLC via digital modules. The extensive range of digital modules allows the most suitable signal module to be selected in each case. Sixteen digital inputs are monitored in this work via an SM 1221 DC signal module. This module is located in the first slot of the four signal module, labeled 102A5. Similarly, it houses 16 channels ranging from channel 0 to 15. Fig. 9 and 10 show the electric charts for the input digital sensors.

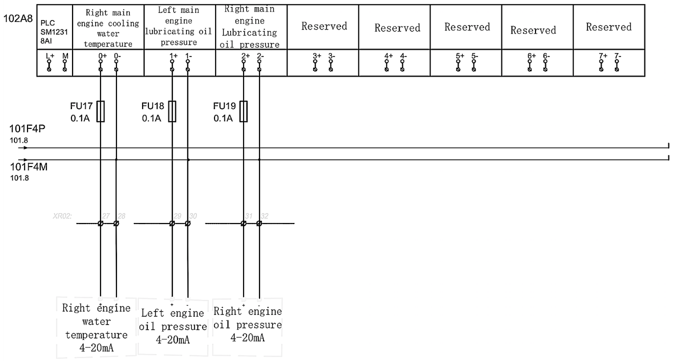

102A8
PLC SM1231 8AI
Right main engine cooling water temperature
Left main engine lubricating oil pressure
Right main engine Lubricating oil pressure
Reserved
Reserved
Reserved
Reserved
Reserved
L+ M
0+ 0-
1+ 1-
2+ 2-
3+ 3-
4+ 4-
5+ 5-
6+ 6-
7+ 7-
FU17 0.1A
FU18 0.1A
FU19 0.1A
101F4P 101.8
101F4M 101.8
Right engine water temperature 4-20mA
Left engine oil pressure 4-20mA
Right engine oil pressure 4-20mA

Fig. 8. Analog signals (signal module SM 1231, slot 102A8).

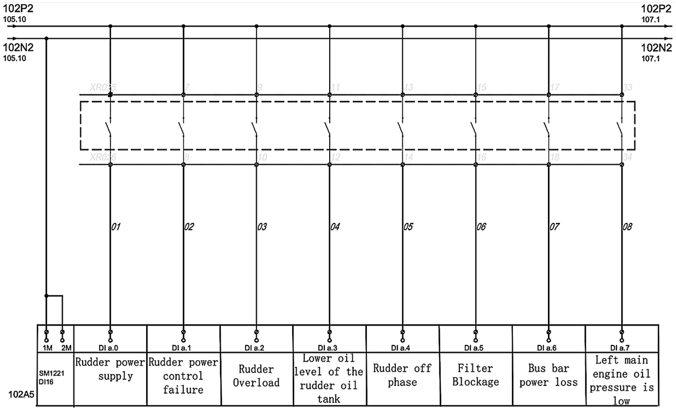


Fig. 9. Digital signals (signal module SM 1221, slot 102A5).

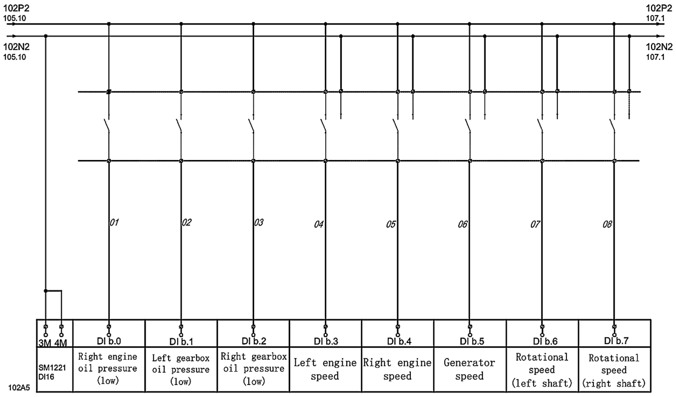


Fig. 10. Digital signals (signal module SM 1221, slot 102A5).

V. Efficient Fault Diagnosis Algorithm

The design of fault diagnosis algorithms in control theory follows three main steps: Fault detection, isolation and recovery. This involves monitoring a system, identifying when a fault has occurred, and pinpointing the type of fault and its location. Generally, two main approaches come in a handy. One, is performing a direct pattern recognition of sensor readings that indicate a fault, and the other is analyzing the discrepancy between the sensor readings and expected values that have been calculated from some model. In the second case, the fault is said to be detected if the discrepancy or residual is above or below a certain threshold. Immediately the fault is detected, it has to be isolated and its location has to be displayed. The fault detection algorithm used in this paper follows the second approach. In other words, an error control logic is attached to all the input sensors, and the responses or outputs of the imposed conditions are displayed on the virtual innovation interface (LabVIEW). The algorithm was designed to be non-intrusive, in recognition that other existing SCADA systems often accommodate fragile timing constraints. The flow chart depicting the detailed blueprint of the logics is given in Fig. 11.

VI. SCADA Design Interface

The SCADA design interface is divided into three major sections. The Indicator section consists of all the alarms, and limit displays. This includes the alarms that require quick response by the operator, and alarms that are frequently used. The Plant section produces a mimic display of the plant status. It also has an Error Report Section, or a segment and distributed architecture that reports failures and faults in real-time processing. The SCADA system is able to process new changes made in the PLC logics to improve signal validation, and alarm suppression. In other words, the display contains at least alarm status, current value, set-point values, and point identification. Input signals from sensors to the system (SCADA GUI) are put into limit checks, necessary engineering unit conversions (e.g. from Kelvin to degree celsius, from atm to Kpa, etc.), and signal validation to calculate the representative values from multiple sensor channels. In order to realize real-time processing, the signal validation is proceeded through simple averaging algorithms. The main purpose of designing alarm and control display is to maximize operator-machine interface visibility. Therefore, the alarm status display is color coded, and indicator buttons are put into different shapes, and are grouped by system and function (that is the analog and digital functions are separated by different Tabs).

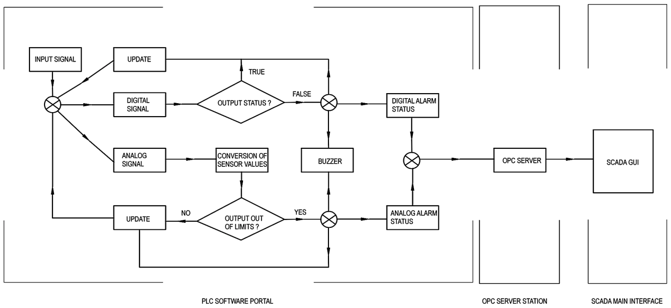


Fig. 11. Logic based alarm design algorithm.

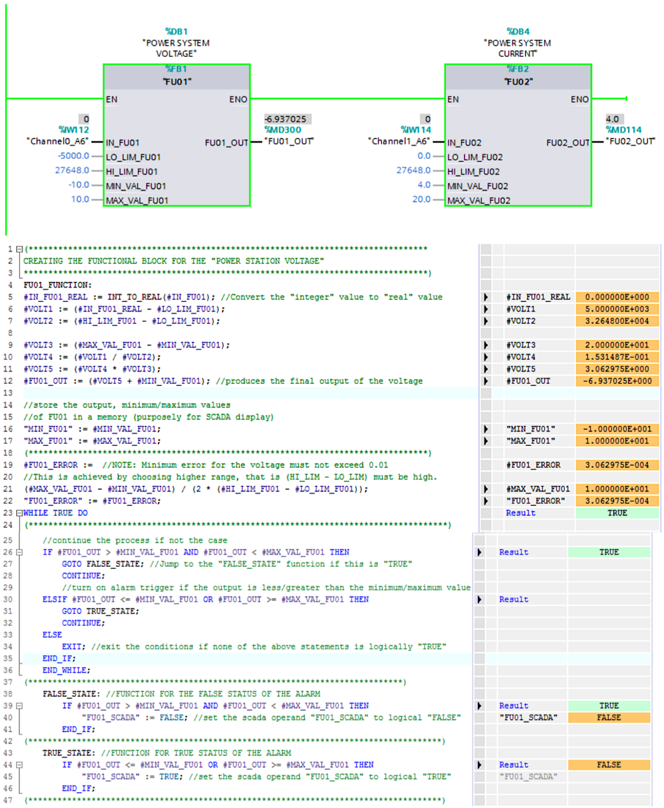
VII. Programming Logics – Analog Signals

A. Monitoring – Voltage and Current

We highlighted in Section IV-A the number of analog signals monitored in this project. In all, nineteen analog signals were being monitored. These signals can be grouped into five categories; as either Voltages, Currents, Pressures, Temperatures, or Levels of fluids. Details of the programming logics of each category is presented and analyzed in this section. The functional block diagram and the logical codes behind the blocks are shown for the monitored voltages/current ("power station") in the codes of Fig. 12. In all cases, the incoming voltage from the voltage sensors to the PLC input is an integer value which needs to be converted into a real value voltage. Internal A/D converter and conversion factors are used to produce a real value output voltage, via predetermined temporary variables (see lines 5 to 12 of the codes). Based on the operating conditions of the field devices powered on by the supplied voltage, operating limits are imposed on the monitored voltages. The PLC channel voltage ranges from -10V to 10V, thus all voltages are within the limits. Similar to the voltage signals, current monitoring proceeds through input current sensors. This input sensor detects the electric current, and then generates a signal proportional to that current. The input DC current appears at the input of the PLC as an integer value which is then converted into a real value output current.

The imposed limits fall within the 4mA-20mA DC operational limits of the input channels of the analog signal module. The availability of functional block diagrams in the TIA portal software provides a genuine opportunity for the operator to monitor both the input and output values of the connected voltage/current sensors. This provides significant amount of information about the status of the monitored voltages/currents. The imposed limits on the outputs similarly serves as checks to avoid device damages. From the block diagram, it can be observed that the inputs are put into checks within “MIN_VAL” and “MAX_VAL“. The tag name FUxx, where FUxx signifies the fuse number, provides easier identification of the monitored device. For instance, FU01_OUT and FU02_OUT indicate the output status of the voltage and current signals, respectively.

The output values and the status of the monitored signals are controlled by a Structured Control Language (SCL) codes behind the functional blocks. Each block is controlled by a different set of codes, with the results observed by the output tags. Within the SCL platform, the operator also gets access to the alarm status of the monitored functions. The maximum values, the minimum values, output and alarm status are all available within the SCL background, as seen from the far right of the voltage codes. Lines 1 to 12 of the codes entails the conversion factors employed in generating the actual output voltage. The parameters hash-tagged VOLTxx are temporary imposed variables within the SCL tags to temporary store the input variables from the voltage sensors. The output is then realized after the input conversion. In order to realize and make accessible of the minimum/maximum voltage values in the SCADA GUI, we store the temporary MIN_VAL_FUxx and MAX_VAL_FUxx in a permanent transferable memory. For instance, MIN_VAL_FU01 and MAX_VAL_FU01 are stored in the permanent real-value memory location MIN_FU01 and MAX_FU01, respectively. Inasmuch as the limits have been imposed to monitor the voltage, an error check criterion is also used to provide additional checks on the voltage surges (from lines 19 to 22). A conditional value of 0.01 is used as a limit check for the surges that may incur during high fluctuation in the voltages. During high voltage spikes, this error check is very significant.



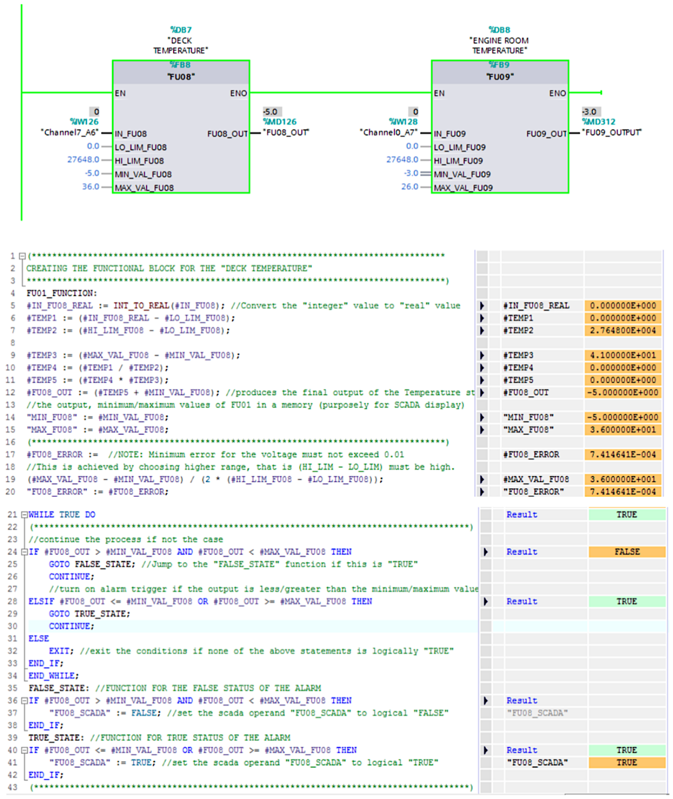
```
1  (*************************************************************************
2  CREATING THE FUNCTIONAL BLOCK FOR THE "POWER STATION VOLTAGE"
3  *************************************************************************)
4  FU01_FUNCTION:
5  #IN_FU01_REAL := INT_TO_REAL(#IN_FU01); //Convert the "integer" value to "real" value
6  #VOLT1 := (#IN_FU01_REAL - #LO_LIM_FU01);
7  #VOLT2 := (#HI_LIM_FU01 - #LO_LIM_FU01);
8
9  #VOLT3 := (#MAX_VAL_FU01 - #MIN_VAL_FU01);
10 #VOLT4 := (#VOLT1 / #VOLT2);
11 #VOLT5 := (#VOLT4 * #VOLT3);
12 #FU01_OUT := (#VOLT5 + #MIN_VAL_FU01); //produces the final output of the voltage
13
14 //store the output, minimum/maximum values
15 //of FU01 in a memory (purposely for SCADA display)
16 "MIN_FU01" := #MIN_VAL_FU01;
17 "MAX_FU01" := #MAX_VAL_FU01;
18 (*************************************************************************)
19 #FU01_ERROR :=  //NOTE: Minimum error for the voltage must not exceed 0.01
20 //This is achieved by choosing higher range, that is (HI_LIM - LO_LIM) must be high.
21 (#MAX_VAL_FU01 - #MIN_VAL_FU01) / (2 * (#HI_LIM_FU01 - #LO_LIM_FU01));
22 "FU01_ERROR" := #FU01_ERROR;
23 WHILE TRUE DO
24 (***************************************************************************)
25     //continue the process if not the case
26     IF #FU01_OUT > #MIN_VAL_FU01 AND #FU01_OUT < #MAX_VAL_FU01 THEN
27         GOTO FALSE_STATE; //Jump to the "FALSE_STATE" function if this is "TRUE"
28         CONTINUE;
29         //turn on alarm trigger if the output is less/greater than the minimum/maximum value
30     ELSIF #FU01_OUT <= #MIN_VAL_FU01 OR #FU01_OUT >= #MAX_VAL_FU01 THEN
31         GOTO TRUE_STATE;
32         CONTINUE;
33     ELSE
34         EXIT; //exit the conditions if none of the above statements is logically "TRUE"
35     END_IF;
36     END_WHILE;
37 (*********************************************************************)
38     FALSE_STATE: //FUNCTION FOR THE FALSE STATUS OF THE ALARM
39         IF #FU01_OUT > #MIN_VAL_FU01 AND #FU01_OUT < #MAX_VAL_FU01 THEN
40             "FU01_SCADA" := FALSE; //set the scada operand "FU01_SCADA" to logical "FALSE"
41         END_IF;
42 (*************************************************************************)
43     TRUE_STATE: //FUNCTION FOR TRUE STATUS OF THE ALARM
44         IF #FU01_OUT <= #MIN_VAL_FU01 OR #FU01_OUT >= #MAX_VAL_FU01 THEN
45             "FU01_SCADA" := TRUE; //set the scada operand "FU01_SCADA" to logical "TRUE"
46         END_IF;
47 (*************************************************************************)
```

```
#IN_FU01_REAL   0.000000E+000
#VOLT1          5.000000E+003
#VOLT2          3.264800E+004

#VOLT3          2.000000E+001
#VOLT4          1.531487E-001
#VOLT5          3.062975E+000
#FU01_OUT      -6.937025E+000


"MIN_FU01"     -1.000000E+001
"MAX_FU01"      1.000000E+001

#FU01_ERROR     3.062975E-004

#MAX_VAL_FU01   1.000000E+001
"FU01_ERROR"    3.062975E-004
Result          TRUE

Result          TRUE


Result







Result          TRUE
"FU01_SCADA"    FALSE



Result          FALSE
"FU01_SCADA"
```

Fig. 12. Functional block diagram and SCL codes for Voltage/Current.

B. Monitoring – Temperature

Temperature is one of the most widely measured parameters in the ship’s engine room. The two most commonly used temperature sensors are the Resistance Temperature Detectors (RTDs) and the Thermocouple. The temperature sensors employed in the engine room, for this work, are the two sensors mentioned above. RTDs have internal resistance that changes with temperature in a predictable, linear way, and hence suitable for small temperature ranges, such as measuring the temperature of water. For wider temperature ranges, up to a couple of thousand degrees, the thermocouple was used. Thermocouple works based on the seebeck effect, where a small voltage is produced across a junction of two dissimilar metals when exposed to temperature gradient. Depending on the temperature characteristics of the field device whose temperature is to be monitored, different limits are imposed on different input sensors. For instance, the limits imposed on the Deck temperature is from -5 to 42 degrees celsius, which is a good range for monitoring environmental temperature. Explanations of the codes for the Deck Temperature is provided in Fig. 13. There are seven monitored temperatures in total, including Deck Temperature, Engine Room, Engine water, and Shaft Temperatures. Just as in the codes for all the other analog functions, lines 4 to 12 produce the real output temperature value, while lines 14 to 15 store the minimum/maximum limits in a permanent working memory. The output and limits are controlled by the lines 21 to 42 of the codes. These conditions are imposed for monitoring purposes. The FUxx_SCADA in all the codes is a bit memory location that indicate either logical TRUE or FALSE depending on the restrictions imposed on the output. For instance, the imposed limits on the deck temperature is between -5 to 36 degrees celsius. Thus FU08_SCADA indicates logical FALSE if the output falls within this range, otherwise TRUE. This is continuously updated as long as the field device and the alarm logics are operational. The speed of this repeated updates and data sequences is controlled by the PLC’s internal clock counters and frequencies. The status of FU08_SCADA triggers the MS01 bit to act in a corresponding way. If MS01 is TRUE, then the signal triggers the BUZZER, BS01 to sound. The flexibility that comes with the TIA portal software makes it easier to monitor the statuses from both within the SCL, FBD programming platform as well as on the Ladder Logic platform. In this paper, all the analog statuses are linked up to a single alarm block (see Section IX). The alarm logics, as well as the functional block diagram used for monitoring the deck temperature is presented in Fig. 13.



```
(***************************************************************************
CREATING THE FUNCTIONAL BLOCK FOR THE "DECK TEMPERATURE"
****************************************************************************)
FU01_FUNCTION:
#IN_FU08_REAL := INT_TO_REAL(#IN_FU08); //Convert the "integer" value to "real" value
#TEMP1 := (#IN_FU08_REAL - #LO_LIM_FU08);
#TEMP2 := (#HI_LIM_FU08 - #LO_LIM_FU08);

#TEMP3 := (#MAX_VAL_FU08 - #MIN_VAL_FU08);
#TEMP4 := (#TEMP1 / #TEMP2);
#TEMP5 := (#TEMP4 * #TEMP3);
#FU08_OUT := (#TEMP5 + #MIN_VAL_FU08); //produces the final output of the Temperature st
//the output, minimum/maximum values of FU01 in a memory (purposely for SCADA display)
"MIN_FU08" := #MIN_VAL_FU08;
"MAX_FU08" := #MAX_VAL_FU08;
(****************************************************************************)
#FU08_ERROR :=  //NOTE: Minimum error for the voltage must not exceed 0.01
//This is achieved by choosing higher range, that is (HI_LIM - LO_LIM) must be high.
(#MAX_VAL_FU08 - #MIN_VAL_FU08) / (2 * (#HI_LIM_FU08 - #LO_LIM_FU08));
"FU08_ERROR" := #FU08_ERROR;

WHILE TRUE DO
 (*****************************************************************************)
 //continue the process if not the case
IF #FU08_OUT > #MIN_VAL_FU08 AND #FU08_OUT < #MAX_VAL_FU08 THEN
    GOTO FALSE_STATE; //Jump to the "FALSE_STATE" function if this is "TRUE"
    CONTINUE;
    //turn on alarm trigger if the output is less/greater than the minimum/maximum value
ELSIF #FU08_OUT <= #MIN_VAL_FU08 OR #FU08_OUT >= #MAX_VAL_FU08 THEN
    GOTO TRUE_STATE;
    CONTINUE;
ELSE
    EXIT; //exit the conditions if none of the above statements is logically "TRUE"
END_IF;
END_WHILE;
FALSE_STATE: //FUNCTION FOR THE FALSE STATUS OF THE ALARM
IF #FU08_OUT > #MIN_VAL_FU08 AND #FU08_OUT < #MAX_VAL_FU08 THEN
    "FU08_SCADA" := FALSE; //set the scada operand "FU08_SCADA" to logical "FALSE"
END_IF;
TRUE_STATE: //FUNCTION FOR TRUE STATUS OF THE ALARM
IF #FU08_OUT <= #MIN_VAL_FU08 OR #FU08_OUT >= #MAX_VAL_FU08 THEN
    "FU08_SCADA" := TRUE; //set the scada operand "FU08_SCADA" to logical "TRUE"
END_IF;
 (*****************************************************************************)
```

```
#IN_FU08_REAL   0.000000E+000
#TEMP1          0.000000E+000
#TEMP2          2.764800E+004

#TEMP3          4.100000E+001
#TEMP4          0.000000E+000
#TEMP5          0.000000E+000
#FU08_OUT      -5.000000E+000

"MIN_FU08"     -5.000000E+000
"MAX_FU08"      3.600000E+001

#FU08_ERROR     7.414641E-004

#MAX_VAL_FU08   3.600000E+001
"FU08_ERROR"    7.414641E-004

Result          TRUE

Result          FALSE

Result          TRUE

Result
"FU08_SCADA"

Result          TRUE
"FU08_SCADA"    TRUE
```

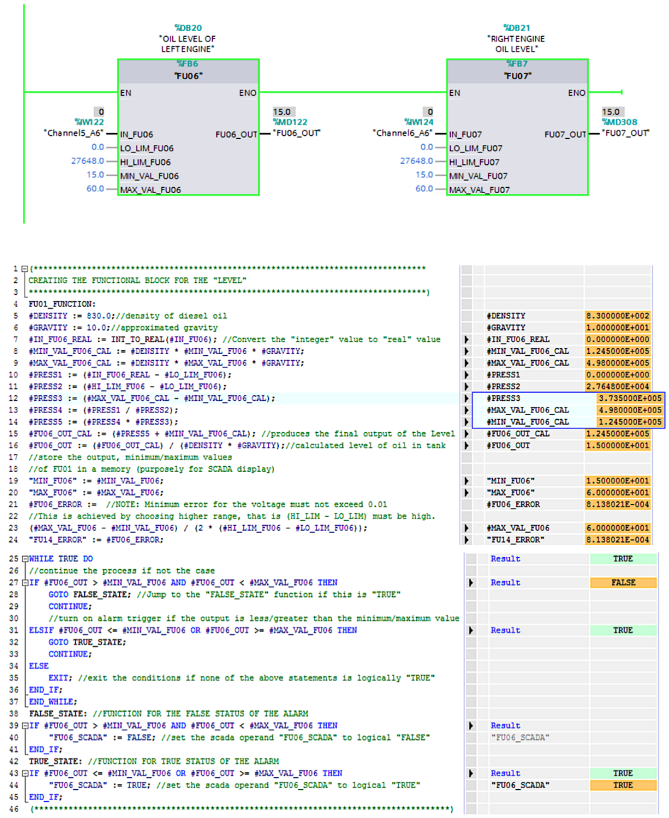
Fig. 13. Functional block diagram and SCL codes for monitoring temperature.

C. Monitoring: Pressure and Level of Fluids

Both pressure and fluid levels proceed through the use of pressure sensors, acting as 4-20mA transducers. The logical codes for the pressure monitoring follows the same logics as that of the temperature. Three pressure sensors are monitored. This includes Engine oil pressures and generator oil pressure. Fluid levels were monitored using pressure sensors, by taking a continues pressure measurement from the bottom of the tank or point where the sensor's diaphragm is placed. Tanks in the engine room were vented to the atmosphere, and so pressure sensors were configured with a vent tube to correct for barometric pressure changes. Levels were calculated from pressure values using the simple relation $p = \rho hg$. Thus the pressure exerted by a column of oil in the tank is dependent solely on the specific gravity of the oil, the level of oil, and the gravitational acceleration. The input pressure sensor values continuously update, and the corresponding level calculated from the given formula above. The values of density, ρ of oil and gravity, g employed in the calculations were $830.0kg/m^3$ and $10.0m/s^2$, respectively. Similar to the case of pressure functions, Engine oil levels, and generator oil level were the monitoring level functions. The codes in Fig. 14 indicate the logics behind the Oil Level measurements. Input limits in the block diagrams are modifiable to the operating conditions.

VIII. Programming Logics: Digital Signals

Sixteen digital functions are monitored in this project. Ladder Logic Programming dominates, except where SCL had to be created to calculate Engine Speed. Digital sensors pass signals to the PLC input, which in turn is recorded in a bit memory location as logical TRUE or FALSE. Output of the digital alarm block receives and acts on the status of the recorded logic. Just as all logics, boolean outputs are labeled FUxx_SCADA, where FUxx represents the Fuse label. Ladder logic for engine speed is presented in Fig. 15.



```
(***************************************************************************
CREATING THE FUNCTIONAL BLOCK FOR THE "LEVEL"
****************************************************************************)
FU01_FUNCTION:
#DENSITY := 830.0;//density of diesel oil
#GRAVITY := 10.0;//approximated gravity
#IN_FU06_REAL := INT_TO_REAL(#IN_FU06); //Convert the "integer" value to "real" value
#MIN_VAL_FU06_CAL := #DENSITY * #MIN_VAL_FU06 * #GRAVITY;
#MAX_VAL_FU06_CAL := #DENSITY * #MAX_VAL_FU06 * #GRAVITY;
#PRESS1 := (#IN_FU06_REAL - #LO_LIM_FU06);
#PRESS2 := (#HI_LIM_FU06 - #LO_LIM_FU06);
#PRESS3 := (#MAX_VAL_FU06_CAL - #MIN_VAL_FU06_CAL);
#PRESS4 := (#PRESS1 / #PRESS2);
#PRESS5 := (#PRESS4 * #PRESS3);
#FU06_OUT_CAL := (#PRESS5 + #MIN_VAL_FU06_CAL); //produces the final output of the Level
#FU06_OUT := (#FU06_OUT_CAL) / (#DENSITY * #GRAVITY);//calculated level of oil in tank
//store the output, minimum/maximum values
//of FU01 in a memory (purposely for SCADA display)
"MIN_FU06" := #MIN_VAL_FU06;
"MAX_FU06" := #MAX_VAL_FU06;
#FU06_ERROR :=  //NOTE: Minimum error for the voltage must not exceed 0.01
//This is achieved by choosing higher range, that is (HI_LIM - LO_LIM) must be high.
(#MAX_VAL_FU06 - #MIN_VAL_FU06) / (2 * (#HI_LIM_FU06 - #LO_LIM_FU06));
"FU14_ERROR" := #FU06_ERROR;

WHILE TRUE DO
//continue the process if not the case
IF #FU06_OUT > #MIN_VAL_FU06 AND #FU06_OUT < #MAX_VAL_FU06 THEN
    GOTO FALSE_STATE; //Jump to the "FALSE_STATE" function if this is "TRUE"
    CONTINUE;
    //turn on alarm trigger if the output is less/greater than the minimum/maximum value
ELSIF #FU06_OUT <= #MIN_VAL_FU06 OR #FU06_OUT >= #MAX_VAL_FU06 THEN
    GOTO TRUE_STATE;
    CONTINUE;
ELSE
    EXIT; //exit the conditions if none of the above statements is logically "TRUE"
END_IF;
END_WHILE;
FALSE_STATE: //FUNCTION FOR THE FALSE STATUS OF THE ALARM
IF #FU06_OUT > #MIN_VAL_FU06 AND #FU06_OUT < #MAX_VAL_FU06 THEN
    "FU06_SCADA" := FALSE; //set the scada operand "FU06_SCADA" to logical "FALSE"
END_IF;
TRUE_STATE: //FUNCTION FOR TRUE STATUS OF THE ALARM
IF #FU06_OUT <= #MIN_VAL_FU06 OR #FU06_OUT >= #MAX_VAL_FU06 THEN
    "FU06_SCADA" := TRUE; //set the scada operand "FU06_SCADA" to logical "TRUE"
END_IF;
 (*****************************************************************************)
```

```
#DENSITY            8.300000E+002
#GRAVITY            1.000000E+001
#IN_FU06_REAL       0.000000E+000
#MIN_VAL_FU06_CAL   1.245000E+005
#MAX_VAL_FU06_CAL   4.980000E+005
#PRESS1             0.000000E+000
#PRESS2             2.764800E+004
#PRESS3             3.735000E+005
#MAX_VAL_FU06_CAL   4.980000E+005
#MIN_VAL_FU06_CAL   1.245000E+005
#FU06_OUT_CAL       1.245000E+005
#FU06_OUT           1.500000E+001


"MIN_FU06"          1.500000E+001
"MAX_FU06"          6.000000E+001
#FU06_ERROR         8.138021E-004

#MAX_VAL_FU06       6.000000E+001
"FU14_ERROR"        8.138021E-004

Result              TRUE

Result              FALSE

Result              TRUE

Result
"FU06_SCADA"

Result              TRUE
"FU06_SCADA"        TRUE
```

Fig. 14. Functional block diagram and SCL codes for monitoring pressure and level of fluids.

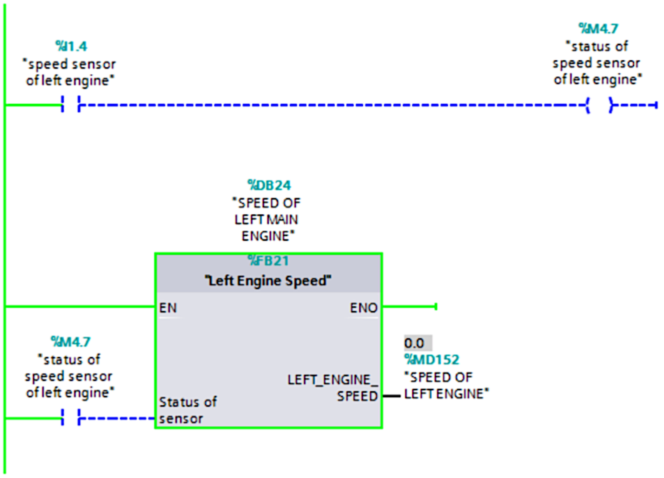

%I1.4
"speed sensor of left engine"
%M4.7
"status of speed sensor of left engine"
%DB24
"SPEED OF LEFT MAIN ENGINE"
%FB21
"Left Engine Speed"
EN
ENO
%M4.7
"status of speed sensor of left engine"
Status of sensor
LEFT_ENGINE_SPEED
0.0
%MD152
"SPEED OF LEFT ENGINE"

Fig. 15. Ladder logic and functional block diagram for the left engine speed.

IX. Programming Logics: Alarm

The alarm architecture constitutes a muffler button and a buzzer, all connected to the input and output respectively, of the CPU in slot 1024A, as seen in Fig. 16. All logical alarm notifications are centered around this network. Alarm notifications are divided into either digital or analog, which are linked to a Functional Block Diagram (FBD). Within the FBD is an SCL logical program that controls the output status of the alarm. Noting from the codes presented in Fig. 17, the alarm notification logic is divided into two parts – "Alarm status for Analog Signal" and "Alarm status for Digital Signal". Line 5 to line 20 constitutes all the bit logics for the analog signals. If the final output "Analog Alarm Status" is logically TRUE, then the notification is sent to the bit location "MS01", which in turn energizes the Buzzer, BS01 to sound. Lines 23 to 34 also show similar situation for the digital signals. The availability of the "STOP" switch, S01 provides manual OFF switch to the Buzzer. The overall digital and analog alarm status can also be observed in the Ladder logic block as seen in the block diagram.

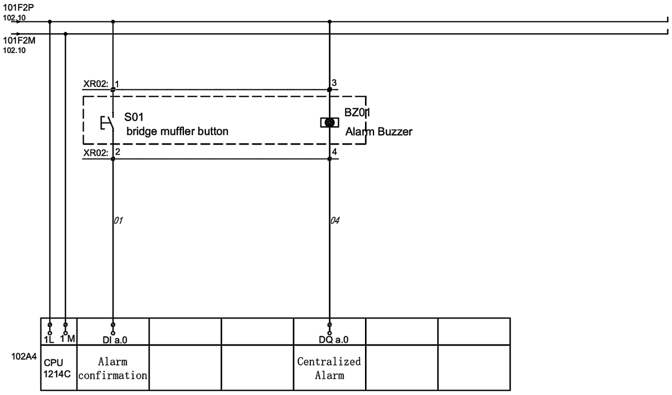
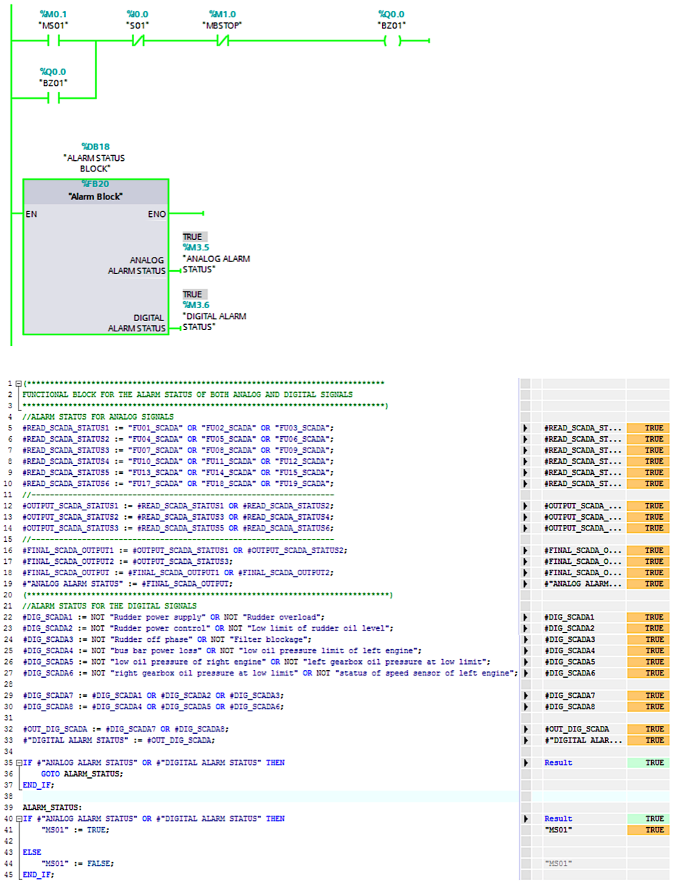


Fig. 16. Alarm network.



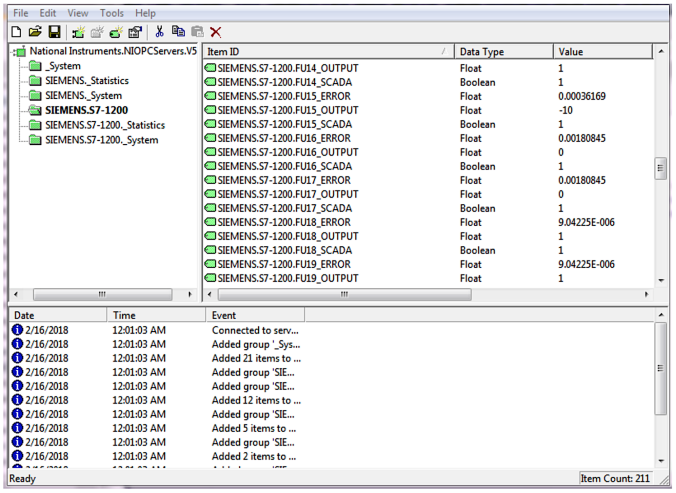
```
 1 (*************************************************************************
 2 FUNCTIONAL BLOCK FOR THE ALARM STATUS OF BOTH ANALOG AND DIGITAL SIGNALS
 3 *************************************************************************)
 4 //ALARM STATUS FOR ANALOG SIGNALS
 5 #READ_SCADA_STATUS1 := "FU01_SCADA" OR "FU02_SCADA" OR "FU03_SCADA";
 6 #READ_SCADA_STATUS2 := "FU04_SCADA" OR "FU05_SCADA" OR "FU06_SCADA";
 7 #READ_SCADA_STATUS3 := "FU07_SCADA" OR "FU08_SCADA" OR "FU09_SCADA";
 8 #READ_SCADA_STATUS4 := "FU10_SCADA" OR "FU11_SCADA" OR "FU12_SCADA";
 9 #READ_SCADA_STATUS5 := "FU13_SCADA" OR "FU14_SCADA" OR "FU15_SCADA";
10 #READ_SCADA_STATUS6 := "FU17_SCADA" OR "FU18_SCADA" OR "FU19_SCADA";
11 //-------------------------------------------------------------------
12 #OUTPUT_SCADA_STATUS1 := #READ_SCADA_STATUS1 OR #READ_SCADA_STATUS2;
13 #OUTPUT_SCADA_STATUS2 := #READ_SCADA_STATUS3 OR #READ_SCADA_STATUS4;
14 #OUTPUT_SCADA_STATUS3 := #READ_SCADA_STATUS5 OR #READ_SCADA_STATUS6;
15 //-------------------------------------------------------------------
16 #FINAL_SCADA_OUTPUT1 := #OUTPUT_SCADA_STATUS1 OR #OUTPUT_SCADA_STATUS2;
17 #FINAL_SCADA_OUTPUT2 := #OUTPUT_SCADA_STATUS3;
18 #FINAL_SCADA_OUTPUT := #FINAL_SCADA_OUTPUT1 OR #FINAL_SCADA_OUTPUT2;
19 #"ANALOG ALARM STATUS" := #FINAL_SCADA_OUTPUT;
20 (*************************************************************************)
21 //ALARM STATUS FOR THE DIGITAL SIGNALS
22 #DIG_SCADA1 := NOT "Rudder power supply" OR NOT "Rudder overload";
23 #DIG_SCADA2 := NOT "Rudder power control" OR NOT "Low limit of rudder oil level";
24 #DIG_SCADA3 := NOT "Rudder off phase" OR NOT "Filter blockage";
25 #DIG_SCADA4 := NOT "bus bar power loss" OR NOT "low oil pressure limit of left engine";
26 #DIG_SCADA5 := NOT "low oil pressure of right engine" OR NOT "left gearbox oil pressure at low limit";
27 #DIG_SCADA6 := NOT "right gearbox oil pressure at low limit" OR NOT "status of speed sensor of left engine";
28
29 #DIG_SCADA7 := #DIG_SCADA1 OR #DIG_SCADA2 OR #DIG_SCADA3;
30 #DIG_SCADA8 := #DIG_SCADA4 OR #DIG_SCADA5 OR #DIG_SCADA6;
31
32 #OUT_DIG_SCADA := #DIG_SCADA7 OR #DIG_SCADA8;
33 #"DIGITAL ALARM STATUS" := #OUT_DIG_SCADA;
34
35 IF #"ANALOG ALARM STATUS" OR #"DIGITAL ALARM STATUS" THEN
36     GOTO ALARM_STATUS;
37 END_IF;
38
39 ALARM_STATUS:
40 IF #"ANALOG ALARM STATUS" OR #"DIGITAL ALARM STATUS" THEN
41     "MS01" := TRUE;
42
43 ELSE
44     "MS01" := FALSE;
45 END_IF;
```

```
#READ_SCADA_ST...      TRUE
#READ_SCADA_ST...      TRUE
#READ_SCADA_ST...      TRUE
#READ_SCADA_ST...      TRUE
#READ_SCADA_ST...      TRUE
#READ_SCADA_ST...      TRUE

#OUTPUT_SCADA_...      TRUE
#OUTPUT_SCADA_...      TRUE
#OUTPUT_SCADA_...      TRUE

#FINAL_SCADA_O...      TRUE
#FINAL_SCADA_O...      TRUE
#FINAL_SCADA_O...      TRUE
#"ANALOG ALARM...      TRUE


#DIG_SCADA1            TRUE
#DIG_SCADA2            TRUE
#DIG_SCADA3            TRUE
#DIG_SCADA4            TRUE
#DIG_SCADA5            TRUE
#DIG_SCADA6            TRUE

#DIG_SCADA7            TRUE
#DIG_SCADA8            TRUE

#OUT_DIG_SCADA         TRUE
#"DIGITAL ALAR...      TRUE

Result                 TRUE




Result                 TRUE
"MS01"                 TRUE


"MS01"
```

Fig. 17. Functional block diagram and SCL codes for the alarm notification.

X. PLC and NI LabVIEW Communication via OPC Server Station

Information exchange can happen in numerous ways between the programmable logic controller and NI LabVIEW. OLE for Process Control (OPC) is used in this paper as a server station to communicate between the PLC and LabVIEW. In order to do this, a device configuration was performed, out of which new channel was created. In regards, a device driver through Siemens TCP/IP Ethernet driver simulator was chosen as the communication network. Optimization configurations were performed, where the write-to-read duty cycle was set to 10 writes for every 1 read. At the same time, configurations for the handling of non-normalized floating point values were set up for the TCP/IP Ethernet driver. The Siemens S7-1200 was then added to the channel, and all tags including analog and digital tags, with their appropriate data types were added to the OPC server. The OPC tags were then connected to the LabVIEW through the creation of I/O server, that updates the connection from LabVIEW to the OPC tags every 100ms. Portion of the OPC tags are shown in Fig. 18.



Item ID	Data Type	Value
SIEMENS.S7-1200.FU14_OUTPUT	Float	1
SIEMENS.S7-1200.FU14_SCADA	Boolean	1
SIEMENS.S7-1200.FU15_ERROR	Float	0.00036169
SIEMENS.S7-1200.FU15_OUTPUT	Float	-10
SIEMENS.S7-1200.FU15_SCADA	Boolean	1
SIEMENS.S7-1200.FU16_ERROR	Float	0.00180845
SIEMENS.S7-1200.FU16_OUTPUT	Float	0
SIEMENS.S7-1200.FU16_SCADA	Boolean	1
SIEMENS.S7-1200.FU17_ERROR	Float	0.00180845
SIEMENS.S7-1200.FU17_OUTPUT	Float	0
SIEMENS.S7-1200.FU17_SCADA	Boolean	1
SIEMENS.S7-1200.FU18_ERROR	Float	9.04225E-006
SIEMENS.S7-1200.FU18_OUTPUT	Float	1
SIEMENS.S7-1200.FU18_SCADA	Boolean	1
SIEMENS.S7-1200.FU19_ERROR	Float	9.04225E-006
SIEMENS.S7-1200.FU19_OUTPUT	Float	1

Date	Time	Event
2/16/2018	12:01:03 AM	Connected to serv...
2/16/2018	12:01:03 AM	Added group '_Sys...
2/16/2018	12:01:03 AM	Added 21 items to ...
2/16/2018	12:01:03 AM	Added group 'SIE...
2/16/2018	12:01:03 AM	Added group 'SIE...
2/16/2018	12:01:03 AM	Added 12 items to ...
2/16/2018	12:01:03 AM	Added group 'SIE...
2/16/2018	12:01:03 AM	Added 5 items to ...
2/16/2018	12:01:03 AM	Added group 'SIE...
2/16/2018	12:01:03 AM	Added 2 items to ...

Fig. 18. OPC server configuration.

XI. SCADA GUI Design

The Graphic User Interface (GUI) was designed in NI LabVIEW. The interface consists of five main sections –

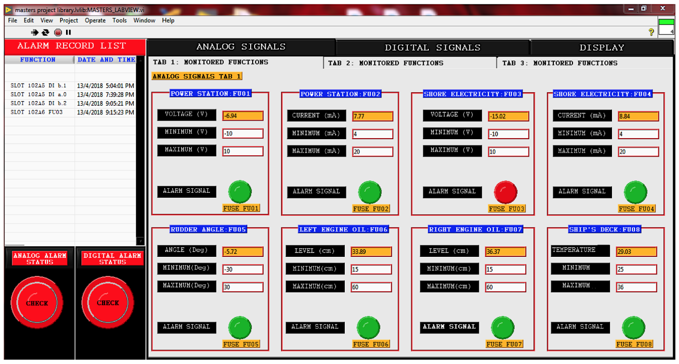
Analog, Digital, Display, Alarm Record List, and Alarm Notification Sections. Analog and Digital Sections were put into subsections (i.e Tab 1, Tab 2, ..., Tab n) to accommodate additional monitored functions. Fig. 19 shows the interface of Tab 1 of the monitored analog signals. For easier identification, functions were labeled by their fuse numbers, and separated from each other in recessed box. FUxx (for instance FU01) in the interface represents the Fuse number connected directly to the input of the monitored PLC channel. All outputs were put into limit checks, and were subjected to constant updates at 100ms. Display formats were set to automatic formatting, with unsigned data types capable of adapting to the source values (OPC Work Station). Signal indicator values were put into necessary precisions, were set to read-only values, and were bound through NI Shared Variable Engine. Each monitored signal was put into necessary scientific unit, and possessed alarm indicator in a form of flash button to prompt the operator of signal values that fell outside limits. Signal indicators were color coded – Red signifies "Fault" whereas Green indicates "No Fault".

Fig. 20 shows the GUI for the monitored digital signals. Digital signal identification starts with slot number and ends with a bit memory location. For instance, logical output "RUDDER POWER SUPPLY" has the label "SLOT 102A5 DI a. 0". SLOT 102A5 symbolizes the location of the signal module on the PLC Workstation, whereas a.0 indicates the bit location. Just as the case for the analog signals, digital signals possess alarm notification in a form of flash button. A red flash indicates a de-energized output, whereas a green flash signifies a working output. In each scenario, "STATUS" responds accordingly – "STATUS" displays "ON" to an energized output and "OFF" to a de-energize output.

The alarm notification section is a combination of the alarm status of all the monitored functions. "ANALOG ALARM STUTUS" responds to the status of the analog signals, whereas "DIGITAL ALARM STATUS" responds to the status of the digital signals. In each case, the flash button turns red when at least one of the signals is in the "red" state. A "green" state indicates "no fault". A written indicator that prompts the operator to "CHECK" the alarm condition can also be observed – this written indicator changes to "GOOD" when all the output statuses are in good condition. The alarm notification section provides easier fault identification. In other words, it prompts the operator of the status of the monitored functions, even before navigating through available tabs.

The "ALARM RECORD LIST" section is one of the very most important sections of the GUI design. This section records the faults and interfaces them according to date and time of the occurrence. In order to produce this, all the monitored functions are stored in a database (Microsoft Access), with their OPC tags attached. Communication is then made between the database and the LabVIEW via the OPC server. The developed program then checks the status of the faulty monitored functions, compares them with the tags in the database, and produces the list according to which fault gets detected first.

Fig. 21 shows the Display Section of the GUI. This section produces a mimic display of the functions that require quick response by the operator. In this section, monitored functions are grouped and separated from each other in a recessed box. Display outputs consist of Temperatures, Pressures, Voltages and Current outputs. Meters are color coded and equally scaled, and labeled for easier identification by the operator. Beneath each meter is an indicator that produces precise values of the measured outputs. In many cases, creating a display section in a SCADA GUI is vital since it provides easier access to real measured values in a more simple way. In other words, it eliminates other indicator buttons. Its purpose is to serve as a platform to observe measured outputs, and not to provide details of alarm statuses. In this paper, all functions in the display section are analog functions.



FUNCTION	DATE AND TIME
SLOT 102A5 DI b.1	13/4/2018 5:04:01 PM
SLOT 102A5 DI a.0	13/4/2018 7:39:28 PM
SLOT 102A5 DI b.2	13/4/2018 9:05:21 PM
SLOT 102A6 FU03	13/4/2018 9:15:23 PM

POWER STATION:FU01
VOLTAGE (V) -6.94
MINIMUM (V) -10
MAXIMUM (V) 10
ALARM SIGNAL
FUSE FU01

POWER STATION:FU02
CURRENT (mA) 7.77
MINIMUM (mA) 4
MAXIMUM (mA) 20
ALARM SIGNAL
FUSE FU02

SHORE ELECTRICITY:FU03
VOLTAGE (V) -15.02
MINIMUM (V) -10
MAXIMUM (V) 10
ALARM SIGNAL
FUSE FU03

SHORE ELECTRICITY:FU04
CURRENT (mA) 8.84
MINIMUM (mA) 4
MAXIMUM (mA) 20
ALARM SIGNAL
FUSE FU04

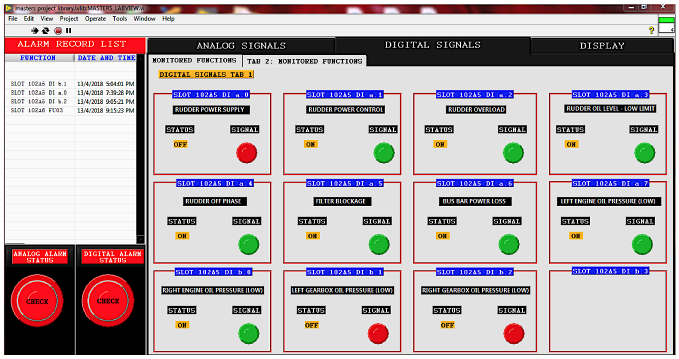
RUDDER ANGLE:FU05
ANGLE (Deg) -5.72
MINIMUM(Deg) -30
MAXIMUM(Deg) 30
ALARM SIGNAL
FUSE FU05

LEFT ENGINE OIL:FU06
LEVEL (cm) 33.89
MINIMUM(cm) 15
MAXIMUM(cm) 60
ALARM SIGNAL
FUSE FU06

RIGHT ENGINE OIL:FU07
LEVEL (cm) 36.37
MINIMUM(cm) 15
MAXIMUM(cm) 60
ALARM SIGNAL
FUSE FU07

SHIP'S DECK:FU08
TEMPERATURE 29.03
MINIMUM 25
MAXIMUM 36
ALARM SIGNAL
FUSE FU08

Fig. 19. Analog signals.



FUNCTION	DATE AND TIME
SLOT 102A5 DI b.1	13/4/2018 5:04:01 PM
SLOT 102A5 DI a.0	13/4/2018 7:39:28 PM
SLOT 102A5 DI b.2	13/4/2018 9:05:21 PM
SLOT 102A6 FU03	13/4/2018 9:15:23 PM

SLOT 102A5 DI a.0
RUDDER POWER SUPPLY
STATUS OFF
SIGNAL

SLOT 102A5 DI a.1
RUDDER POWER CONTROL
STATUS ON
SIGNAL

SLOT 102A5 DI a.2
RUDDER OVERLOAD
STATUS ON
SIGNAL

SLOT 102A5 DI a.3
RUDDER OIL LEVEL - LOW LIMIT
STATUS ON
SIGNAL

SLOT 102A5 DI a.4
RUDDER OFF PHASE
STATUS ON
SIGNAL

SLOT 102A5 DI a.5
FILTER BLOCKAGE
STATUS ON
SIGNAL

SLOT 102A5 DI a.6
BUS BAR POWER LOSS
STATUS ON
SIGNAL

SLOT 102A5 DI a.7
LEFT ENGINE OIL PRESSURE (LOW)
STATUS ON
SIGNAL

SLOT 102A5 DI b.0
RIGHT ENGINE OIL PRESSURE (LOW)
STATUS ON
SIGNAL

SLOT 102A5 DI b.1
LEFT GEARBOX OIL PRESSURE (LOW)
STATUS OFF
SIGNAL

SLOT 102A5 DI b.2
RIGHT GEARBOX OIL PRESSURE (LOW)
STATUS OFF
SIGNAL

SLOT 102A5 DI b.3

Fig. 20. Digital signals.

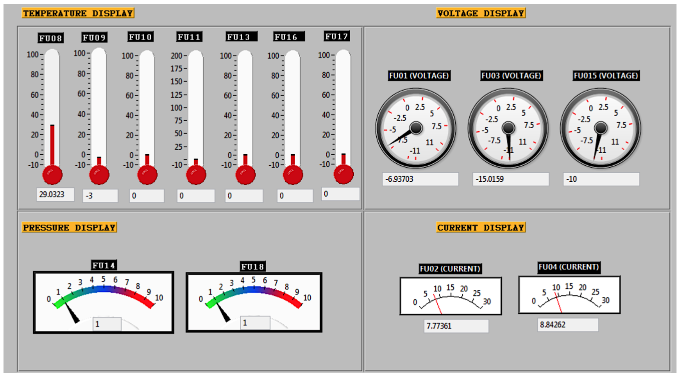


Fig. 21. GUI – Display section.

XII. Conclusion

In this paper, an effective alarm design method for monitoring engine room machinery was presented. The proposed design proceeded through PLCs and SCADA-based systems. Reasons for such a design method were highlighted, and the programming platforms for the design were given. The strengths and weaknesses of some design methods presented in some literary works were also reported, and solutions to such problems were proposed. The designed monitoring system was tested in the shipyard for performance evaluation. The testing results proved that this design technique is reliable and effective for online engine control room device monitoring. However, as effective as it is, one peculiar limitation was encountered. This limitation is peculiar to the voltage measurements. The error control algorithm could not handle voltage spikes. A voltage spike forces the input voltage sensors to record values outside the imposed limits. This in turn produces unusual "ON" and "OFF" flicks of the voltage indicator buttons, thus interrupting the operator of the output voltage statuses. For future work, a direct pattern recognition of sensor readings for fault diagnosis is planned to be implemented. An efficient pattern recognition algorithm will help prevent the unusual ON and OFF behavior of indicator buttons that occur during input voltage spikes. In addition, a Web based monitoring system will be designed to monitor and share data across the internet to other off-site bases.

Acknowledgment

We would like to extend gratitude to Changjiang Waterway Bureau, the School of International Studies of Jiangsu University of Science and Technology, and Ministry of Education of Jiangsu province for providing the necessary financial support.

References

[1] EEMUA, "Alarm System: A guide to design, management and procurement". The Engineering Equipment and Materials Users Association. Publication No. 191. EEMUA 45 Beech Street London EC2Y 8AD (1999).

[2] Keith Stouffer, Joe Falco, Karen Kent, Guide to supervisory control and data acquisition SCADA and industrial control systems security, National Institute of Standards and Technology, 2006. ISBN 20899- 8930.

[3] A. Lakshmi, Velammal Coll., Chennai, B. Sangeetha, A. Naveenkumar, B. Ganesh, Experimental validation of PID based cascade control system through SCADA–PLC–OPC Interface, IEEE Computer Communication and Informatics (ICCCI), (2012) pp. 1–4.

[4] Berge, Jonas. "Fieldbuses for Process Control" : Engineering, Operation, and Maintenance, ISA, 2002.

[5] M. S Zaghloul, "Online Ship Control Systems Using Supervisory Control and Data Acquisition (SCADA)". IJCSA, (2014),DOI: 10.14355/ijcsa.2014.0301.02.

[6] C. Kwon Lee, S. Hur, J.H. Shin, I. S. Koo, and J.K. Park "A basic design of Alarm System for the Future Nuclear Power Plants in Korea". Taejon, Korea. (2012).

[7] D. Sharma, A. Chaubey, A. K Gupta, "Review on Methods of Fault Detection and Protection of Induction Motor". International Journal of Computer Applications, vol 161, (2017).

[8] L. de Miguel, L. Blazquez, "Fuzzy logic-based decision-making for fault diagnosis in a DC motor", Engineering Applications of Artificial Intelligence, vol. 18, no. 4, pp. 423-450, 2005.

[9] M Iacob, G.D Andreescu, and N. Muntean. "SCADA system for a central heating and power plant". 5th International Symposium on Applied Computational Intelligence and Informatics (SACI '09), At Timisoara, Romania. DOI: 10.1109/SACI.2009.5136232.

[10] Bentley Systems, The Fundamentals of SCADA, 2004.

[11] Michael P. Ward, An architectural framework for describing supervisory control and data acquisition SCADA systems, thesis, Naval postgraduate school Monterey, California, September 2004.

[12] B.S. Jones, J.V. Earthy, and D. Gould. "Improving the design and management of alarm systems". A paper presented to the World Maritime Technology Conference, (2006).

[13] S. Da'na, "Development of a monitoring and control platform for PLC-based applications". Journal Computer Standards & Interfaces, Vol. 30, Issue 3, pages 157-166, March 2008

[14] C. Wang, H. R. Xiao, W. G. Pan, Y. Z. Han "Design of Monitoring and Alarm System for the Ship's Engine Room", Advanced Materials Research, Vols. 268-270, pp. 1663-1668, 2011.

[15] T. Kletz, "Learning from Accidents in Industry". Butterworts, London, ISBN 0-408-02696-0, 1988, page 95.

[16] Arthur Wright, P. Gordon Newbery "Electric fuses", 3rd edition, Institution of Electrical Engineers (IET), 2004, ISBN0-86341-379-X, pp. 2–10

[17] S7 7A 125V TD "Rejection Base Plug Fuse", Elliott Electric Supply, (2012)

[18] D. G. Fink, H.W. Beaty, "Standard Handbook for Electrical Engineers" Eleventh Edition, McGraw Hill 1978 ISBN0-07-020974-X pp.10–116 through 10-119.

Performance Analysis of Machine Learning Algorithms for Missing Value Imputation

Nadzurah Zainal Abidin, Amelia Ritahani Ismail*
Department of Computer Science
Kulliyyah of Information and Communication Technology,
International Islamic University Malaysia,
P.O Box 10, 50728 Kuala Lumpur, Malaysia

Nurul A. Emran
Faculty of Information and Communication Technology,
Universiti Teknikal Malaysia Melaka (UTeM),
Hang Tuah Jaya, Durian Tunggal, Melaka, 76100 Malaysia

***Abstract*—Data mining requires a pre-processing task in which the data are prepared, cleaned, integrated, transformed, reduced and discretized for ensuring the quality. Missing values is a universal problem in many research domains that is commonly encountered in the data cleaning process. Missing values usually occur when a value of stored data absent for a variable of an observation. Missing values problem imposes undesirable effect on analysis results, especially when it leads to biased parameter estimates. Data imputation is a common way to deal with missing values where the missing value's substitutes are discovered through statistical or machine learning techniques. Nevertheless, examining the strengths (and limitations) of these techniques is important to aid understanding its characteristics. In this paper, the performance of three machine learning classifiers (K-Nearest Neighbors (KNN), Decision Tree, and Bayesian Networks) are compared in terms of data imputation accuracy. The results shows that among the three classifiers, Bayesian has the most promising performance.**

Keywords*—Data Mining; Imputation; Machine Learning; K-Nearest Neighbors; Decision Tree; Bayesian Networks

I. Introduction

Data mining is a modern approach to solve many complex and real world problems. This fairly self-explanatory term is a well-known and widely used process that evolves with new technologies. In data mining, data pre-processing is the most important step to ensure the quality of data and the results that leads to reliable decisions. According to Vivek, data pre-processing is the process of simple transformation of raw data into understandable format. Data pre-processing major activities include data cleaning, integration, transformation, data reduction and data discretization as shown in figure 1. One critical activity in data pre-processing is dealing with missing data. This process falls under the first stage of pre-processing data, which is data cleaning. This first stage of data pre-processing is concerned about detecting incomplete, inaccurate, inconsistent and corrupt data, and applying techniques to modify or to delete this spurious data [1]. Pyle proposed in his book Data Pre-preparation for Data Mining that major tasks in data cleaning are to impute missing data, remove outliers and resolve inconsistencies. In fact, in data quality, missing values has been recognized as one form of data completeness problem [2].

In certain observation of interest, missing data can be defined as the absence of data value for a variable. Missing data is commonly described as major issue in most scientific research domains that may originate from such mishandling samples, low signal-to-noise ratio, measurement error, non-response or deleted aberrant value [1]. Nevertheless, as claimed, missing data can also introduce the element of uncertainty in analyzing data. Previous researchers have proposed several ways in handling missing values. The simplest technique is to ignore the missing values [3]. This technique is usually adopted when to a missing class label. Nevertheless, the technique is not appropriate and effective in the case where the percentage of missing values differ significantly. The next technique is to manually fill in the missing value, which will only introduce tedious and infeasible results. Somasumdaram and Nedunchezhian claimed that the third technique used in dealing with missing values is using a global constant (such as 'unknown') to fill up the missing values in data sets. Even though this technique use global constant value to substitute the missing value, it treats all data sets as the same. As a results, a considerable amount of distortions will be introduced in the data sets of concerned. In addition, if similar global constant such as 'unknown' is used, the data is still implicitly incomplete, as the value represents a variation of 'NULL' that denotes missing especially in database community. The final technique is data imputation, that relies on observed data sets to predict missing values [4] (Fig. 1).

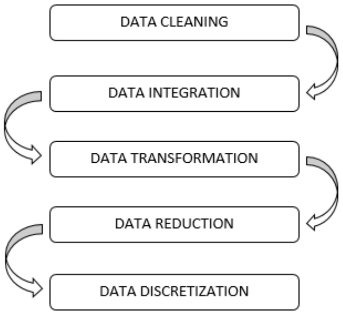


Fig. 1. Data Mining Task (Vivek Agarwal, 2015).

Data imputation is defined as a technique of replacing missing data with substituted values [5]. Selection of imputation method usually determined by the mechanism of how the values are missing. Rubin has described the three missing values mechanism as missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). MCAR is to describe a situation where the missing values is not correlate to a certain value which assumes to be obtained or to an observed responses [5]. MAR data is the situation where the likelihood of missing value instance mostly depends on the known values instead on the real value of the missing data itself [6]. While MNAR describes a situation when the propensity of a missing value in a class instance is to depend on the value of that variable.

In the literature, various data imputation techniques have been introduced, Statistical and Machine Learning techniques have been used in various application contexts of data imputation as we shall see in the next section. Even though the conventional, statistical technique has been adopted for decades, the machine learning-based data imputation techniques are becoming popular in handling missing values especially in large data sets. In the next section, description of statistical and machine learning techniques (classifiers) used for data imputation will be given. Section III covers the evaluation methods for the comparison of three classifiers namely KNN, Decision Trees, and Bayesian Networks. These classifiers will further be measured by evaluating three parameters: Mean Square Error (MSE), Mean Absolute Error (MAE) and Root Square Mean Error (RMSE) in section III. This is followed by Section IV for the results and discussion. Finally, Section V concludes this paper.

A. Literature Review

Data imputation theory is an emerging topic in statistics and machine learning. In this paper, we aimed to explore the characteristics of the techniques.

B. Statistical Approach of Handling Missing Values

1) Listwise Deletion: In imputing missing values, the most traditional theory used is by throwing away data. By this way, we omit records with missing values and continue to analyze the remaining data [4]. This technique is reputably known as listwise deletion, and falls under one of the statistical techniques. Handling missing values with listwise deletion is a default option in most statistical analysis. However, this approach is only pertinent to be used if there is only limited number of missing values, as otherwise it will eventually lead to biased analysis. Another limitation with listwise deletion, it is only relevant when missing values are completely at random (MCAR) which unfortunately rarely happens in reality [5]. Apart from that, one might risk loss of critical information if all missing values are deleted. Ultimately, this approach leads to bias parameters and estimates.

2) Pairwise Deletion: Another known statistical method of handling missing data is pairwise deletion. One researcher [5] claimed that pairwise deletion technique gets rid of information on a particular information data to test if a particular assumption is missing. This statistical testing will be adapted to the observed data if there are missing value elsewhere in the dataset. A disadvantage of pairwise deletion is the tendency to produce a standard of errors that are either underestimated or overestimated [7]. Besides, pairwise deletion is not able to compare analyses as sample dataset different each time.

Marina Soley-Bori mentioned that the two improved approaches that have been proposed to handle missing values are multiple imputation and maximum likelihood [8].

3) Multiple Imputation: In multiple imputation, a new technique of treating missing values is introduced, where it imputes missing values with a set of acceptable values that may contain uncertainty to the original values, instead of replacing a single data to all missing attributes [6].

This approach usually begins with a prediction of the existing data from another variable and then replaced the missing values with the predicted values [6]. A full set of plausible values is the results of the imputed data set. Nevertheless, it has been reported that the downside of this method is different uncertainty values may be yielded for the same data set used for imputation [9].

4) Maximum Likelihood: In Maximum Likelihood is implied, the assumption used is the observed data is from a multivariate normal likelihood function to a linear model. According to researchers [10], the equation of maximum likelihood estimation for incomplete data set are:

$$y \epsilon R^n$$

$$z \epsilon R^1$$

$$(y,z) \epsilon R^n + 1$$

where y is observed data, z is missing data and (y,z) are the complete data.

This technique behaves by estimating the observed data using existing data and estimate missing values with respect to the estimated parameters. The limitation of this approach are it requires specialized software, which may be challenging and time-consuming.

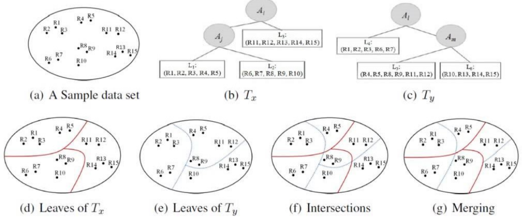
Imputation supposed to produce a complete data set in order to improve its usefulness. However, the statistical techniques described so far still suffer from loss of information. This will eventually lead to invalid conclusions and biased parameters. Therefore, in the next section, alternative way of imputation for missing values using machine learning techniques (or also called as classifiers) will be presented.

C. Machine Learning Approach of Handling Missing Data

Machine learning approach has revolutionized the world with various algorithms to aid data analysis. However, in data imputation, machine learning is in its infancy, and thus offers many research opportunities. In this paper, we focus on four machine learning techniques that have been proposed in data imputation. These techniques are as follows:

1) Decision Tree: Decision tree is another common predictive model used to impute missing values. Decision tree has introduced imputation techniques to the missing values that allows validation of the imputed values against the actual values. This technique begins by splitting the leaves of a tree until running out of questions.

A decision tree has two kinds of nodes. First, this approach tackles imputation by determining each leaf node that has a class label with a majority vote of training examples reached the leaf. Besides, each internal node should represent a question on features that will be branching out according to the answers as Fig. 5 [11] (Fig. 2).


R1
R2
R3
R4
R5
R6
R7
R8
R9
R10
R11
R12
R13
R14
R15
A_i
A_j
L_1: (R1, R2, R3, R4, R5)
L_2: (R6, R7, R8, R9, R10)
L_3: (R11, R12, R13, R14, R15)
A_l
A_m
L_4: (R1, R2, R3, R6, R7)
L_5: (R4, R5, R8, R9, R11, R12)
L_6: (R10, R13, R14, R15)

(a) A Sample data set (b) T_x (c) T_y

(d) Leaves of T_x (e) Leaves of T_y (f) Intersections (g) Merging

Fig. 2. Basic Concept of Decision Tree (Rahman and Islam, 2013).

$$H(D) = -\sum_{i=1}^{k} P(C_i|D) log_k(P(C_i|D)) \tag{1}$$

The equation assumes that all trees are equally split through the dataset.

As claimed, the transparency of decision tree has made it as the most frequent algorithm used in data mining approach [12]. Nevertheless, the researchers explained that the root in decision tree algorithm should illustrate a question with multiple answers. For imputation purposes, each answer should generate a set of questions that help to determine the data and make the final decision based on it. The final result of decision tree should indicate the possibility of all scenario of decision and outcome.

Despite all benefits mentioned, one researcher claimed that main drawback of decision tree is the computational cost such as running time and trees to construct different test samples [13].

2) K-Nearest Neighbor (KNN): K-nearest neighbors (KNN) is the most straightforward algorithm in imputing missing values. Besides, this algorithm has been used to solve many predictive problems.

In order to impute a value of a variable, K-nearest neighbors (KNN) defines a set of nearest neighbor for a sample and substitutes the missing data by calculating the average of non-missing values to its neighbors [6]. Nearest neighbors is measured as the closest values based on the Euclidean distance as follows.

$$D(a,b) = \sqrt{\sum_{i=1}^{n}(b_i - a_i)^2} \tag{2}$$

As KNN imputes missing values based on its neighbor, it may introduce an uncertain analysis in relation to the value of *k*. If *k* is too small for a big dataset, the classifier may be susceptible to over-fitting and sensitive to noise points. On the other hand, if *k* is too large, this may cover all data points that are located far away from its neighbors. The decision will eventually lead to bias as it covers a greater instance space.

As to the matters mentioned in relation to *k*, the best choice of *k* influence t o m ake a b etter d ecision a nd a nalysis. One researcher [14] claimed that the most suitable value for *k* can be obtained through a formula of 1/*k* as shown in Fig. 3 with regards on the size of dataset and percentage of missing values.

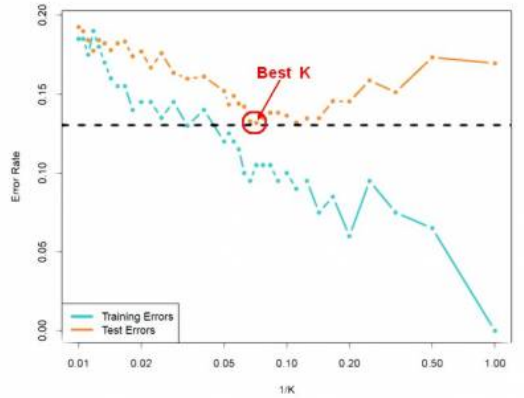


Fig. 3. Best K-Value (Gerardnico, 2017).

KNN is one of the algorithms commonly used because of the simplicity of imputation. However, this imputation technique requires scanning the entire dataset to find the k-nearest neighbors and thus it can be expensive and suffers poor performance especially for a large dataset [15].

3) Bayesian Network: Another machine learning technique used for data imputation is Bayesian networks. Bayesian networks are growing as the model of choice in resolving many problems. Bayesian capture probabilistic relationships between variables in a concise manner by enforcing conditional independence constraints [16]. Using Bayesian networks for imputation offers several advantages: 1) the ability to handle missing data models encodes dependencies among all variables, 2) it preserves the joint probability distribution of the variables which KNN methods do not promise. Unfortunately, Bayesian cannot afford to support a large size of dataset as it requires to learn a network and discretization of all data accurately. This process is usually required unless conditional probability of Bayesian are explicitly modeled and can be parameterized, which frequently with higher computational expense [17].

A particularly elegant way Bayesian handle missing data is as follows (assuming that *xj* has the missing values):

$$P(x_1...x_j...x_d|y) = P(x_1|y)...P(x_j|y)...P(x_d|y) \tag{3}$$

$$\sum_{xj} P(x_1...x_j...x_d|y) = \sum_{xj} P(x_1|y)...P(x_j|y)...P(x_d|y) \tag{4}$$

$$= P(x_1|y)... \sum_{xj} P(x_1|y)...P(x_d|y) \tag{5}$$

$$= P(x_1|y)...1...P(x_d|y) \tag{6}$$

$$\tag{7}$$

The above equation shows that all prediction of missing values will eventually equal to 1. The Bayesian approach relies on the collection of data then calculating the probability that data is significantly related to the information that was extracted.

The key ingredient of Bayesian approach is treating missing data as added unknown quantities to be able to estimate a posterior distribution. A posterior distribution can be defined as the total knowledge of integration between prior distribution and likelihood function to a parameter after been observed [18]. Regardless, the Bayesian approach helps to easily adapt to include partially adapted observed cases as well as incorporate realistic assumptions for the reasons of missingness of datasets.

In the next section, details on how to evaluate the accuracy of the machine learning techniques described in this section will be provided.

II. Experimental Setup

This section attempts to establish the most appropriate classifiers in relation to the percentage of missing values in a dataset.

Fig. 4 shows the flow of experiment conducted. The first step is with acquiring medical dataset from data.gov.uk, Canada Open Data, UCI Machine Learning Repository and World Health Organization (WHO).

Second steps emphasize on calculating the percentage of missing values in all ten medical datasets. The objective of this activities is to analyze the most fitting classifier that suits with various percentage of missing values.

Before the real experiment phase begins, all missing values shall be cleaned to prevent problems caused by missing values when training a model [**?**]. For the purpose of this study, we artificially create missing values from a complete data to validate the imputed missing values against actuals. The validation is measured with MAE, MSE, and RMSE. The third step helps to identify what data need to be analyzed. In this phase also identify a different algorithm for developing the rules and classification techniques to concentrate on the missing information that you need. As claimed by Ian H. Witten, Eibe Frank and Mark A. Hall in their book Data Mining: Practical Machine Learning Tools and Techniques, second and third steps should cover the role of implementing processes and decision making that generate ultimately results.

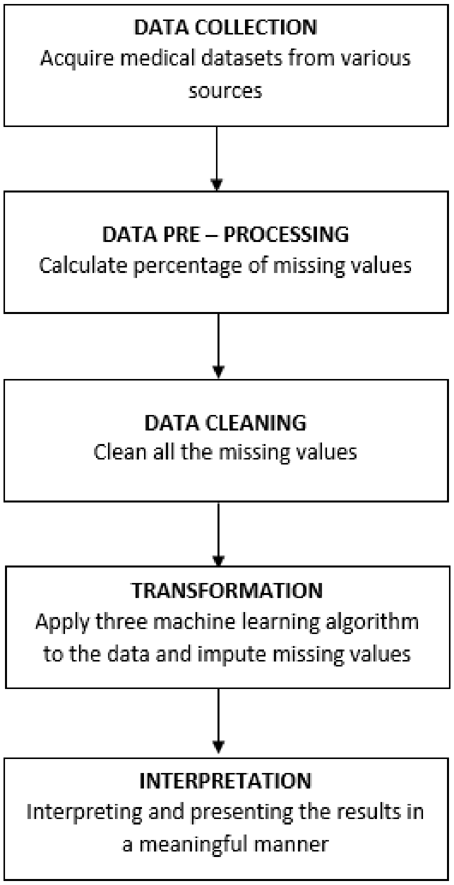

DATA COLLECTION
Acquire medical datasets from various sources
DATA PRE – PROCESSING
Calculate percentage of missing values
DATA CLEANING
Clean all the missing values
TRANSFORMATION
Apply three machine learning algorithm to the data and impute missing values
INTERPRETATION
Interpreting and presenting the results in a meaningful manner

Fig. 4. Experiment Flow.

Next phase covers the identification of relevant values and information, substituting missing data with valid estimations. Besides, this phase should be able to define the appropriate approach to imputing missing values for the medical dataset. The performance of each approach is compared and results presented.

The final step is interpretation step where the results yielded are analyzed. The performance is gathered as an element to validate our hypothesis. In this step, the final results of data imputation is also compiled.

III. Evaluation Criteria

An experiment is conducted to demonstrate the performance of machine learning techniques where ten simulated datasets were acquired and publicly available at: data.gov.uk[1] and Canada Open Data portals[2], UCI Machine Learning Repository[3] and World Health Organization (WHO)[4].

Generally, there are many possible reasons clinical has the most missing values such as patient refusal to answer questions when it related to privacy issues, unable to understand questions given, patient migration, early successes of a treatment,

[1] https://data.gov.uk/

[2] https://open.canada.ca/en

[3] http://archive.ics.uci.edu/ml

[4] http://www.who.int/gho/en/

treatment or instrumental failures, adverse events and death of respondent due to accident or other reasons [16], [19].

All these real life datasets are medical datasets and has missing values due to several reasons mentioned. The percentage of missing value for each dataset are shown in Table I. Table I refers to information regarding the number of records and the amount of missing values (in percentage) are provided along with the data sets.

TABLE I: Summary of Datasets

Dataset	Records	Percentage of Missing Values
Admissions	192	1.56%
Alcohol	39	10.26%
Autism	229	1.4%
Body Mass Index (BMI)	864	1.7%
Drug	458	45.33%
Funerals	60	30%
Infection	1386	19.19%
KPI Health	730	57.22%
Mental Health	108,342	8.07%
Obesity	1458	13.31%

The three machine learning classifiers are evaluated using three criteria: Mean Absolute Error (MAE), Mean Squared Error (MSE) and Root Mean Squared Error (RMSE).

MAE measures the average difference between imputed values and true values as in the following equation:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} \mid y_i - \hat{y}_i \mid \tag{8}$$

While MSE is equal to the sum of variance and squared of the predictions of missing values, defined as:

$$MSE = \frac{1}{n}\sum_{i=1}^{n} (y_i - \tilde{y}_i)^2 \tag{9}$$

RMSE calculates the difference between predicted (imputed) and actuals values. Basically, it represents the sample of differences in standard deviation as follows:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(X_i^a bs - X_i^i mputed)^2}{n}} \tag{10}$$

IV. Analysis of Results

This section presents the result of simulations done on the ten datasets with respect to accuracy and percentage missing values. Based on Table II below, the accuracy of each algorithm were compared using three parameters as mentioned in the previous section. These three parameters: MAE, MSE, and RMSE were estimates by observing the lowest values. All these three parameters are negatively-oriented scores, which concludes the lower results the better.

MAE, MSE, and RMSE are the most useful parameters to evaluate the performance of predicting methods and to measure forecast accuracy. Generally, all these parameters are measured on the error difference between the imputed values and actual values.

TABLE II: Results of Machine Learning Classifiers

Dataset	ML Classifier	MAE	MSE	RMSE
Admissions	KNN	5.823	4.924	7.017
	Decision Tree	4.314	2.289	4.784
	Bayesian Network	2.534	1.919	4.381
Alcohol	KNN	7.560	5.424	7.365
	Decision Tree	139.25	41955.2	204.829
	Bayesian Network	507.25	319865.25	565.566
Autism	KNN	2.1207	6.1548	2.4809
	Decision Tree	2.500	11.500	3.391
	Bayesian Network	0.5	1.0	1.0
Body Mass Index (BMI)	KNN	12.4323	346.4292	18.613
	Decision Tree	15.975	270.788	16.456
	Bayesian Network	12.416	418.579	4.150
Drug	KNN	10.691	172.65	13.140
	Decision Tree	11.925	201.057	14.179
	Bayesian Network	23.0377	642.887	25.355
Funerals	KNN	794.25	916747.2	957.47
	Decision Tree	815.965	1206574.0	1098.442
	Bayesian Network	817.49	1248121.39	1117.2
Infection	KNN	1.124	3.003	1.733
	Decision Tree	4.951	2539.14	50.389
	Bayesian Network	6.3534	136.744	11.694
KPI Health	KNN	9.410	3.603	1.898
	Decision Tree	1.253	2.116	4.599
	Bayesian Network	7.573	5.599	2.366
Mental Health	KNN	6.234	1.725	1.313
	Decision Tree	5.988	1.703	1.305
	Bayesian Network	1.039	3.349	1.830
Obesity	KNN	1.124	3.003	1.733
	Decision Tree	4.951	2539.14	50.389
	Bayesian Network	6.353	136.744	11.694

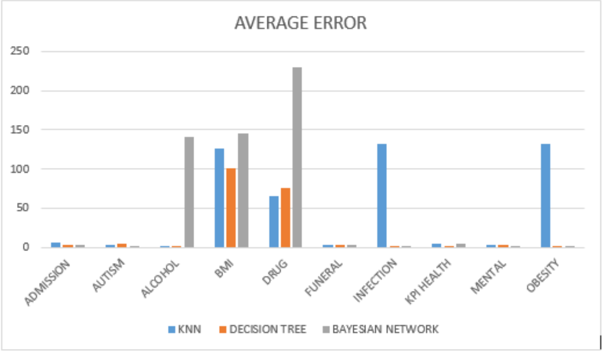


Fig. 5. Average Error for All Datasets.

In accordance with Table II, bayesian has consistently produced the lowest imputation error against all three parameters. This findings in II proves that Bayesian approach is the most appropriate machine learning classifier to impute missing data with regards to smaller sizes of the dataset, less than 20 percent. However, imputation with Bayesian network can be computationally expensive for larger datasets.

Besides, the result drawn from Table II concludes that: the second most standout machine learning classifier is decision tree. Although Bayesian network and decision tree have almost the same results, decision tree is best to apply for larger datasets with higher missing values to imputes.

Nonetheless, KNN also shows the lowest value of error accuracy in some datasets. Surprisingly, the datasets with KNN as the lowest value has a higher percentage of missing values, 30 percent and above. This demonstrates that although KNN consumes time searching through entire datasets, KNN performs better in imputing missing values regardless how big the size of datasets. Nevertheless, the findings also show that KNN imputation method will never extrapolate outside the range of missing value.

To conclude, the experiments have proved that the proposed machine learning classifiers have a better approach of imputing missing values compared to statistical techniques.

V. Conclusion

In data mining, missing values can be a root cause to produce the wrong final analysis. Besides, in many research area, missing data is a universal problem that may influence the biased estimations and wrong conclusions. To overcome the negative impacts of missing values, a process called missing data imputation should be taken before proceeding to the next phase such as data mining. This paper evaluates three machine learning classifiers namely decision tree, KNN, and Bayesian network, to substitutes missing data and compare each accuracy. The result shows that, the Bayesian network has the lowest value for the three parameters which conclude that the best approach to imputing missing values. However, other factors also influence this error estimators such as percentages of missing values and sizes of datasets. Although Bayesian consistently shows the lowest values, the results are only significant for small sizes of the dataset with less than 20 percent missing values.

VI. Future Work

A future work for imputation in medical dataset must emphasis on optimizing the highest accuracy of a machine learning classifier to impute missing values. This optimization helps to boost machine learning performance for out-of-sample trained using the imputed dataset.

Acknowledgment

This research was supported by the IIUM Research Initiative Grants Scheme (RIGS): RIGS16-346-0510.

References

[1] Agarwal Vivek, Research on Data Preprocessing and Categorization Technique for Smartphone Review Analysis, *International Journal of Computer Applications*, 131(4):30–36, 2015.

[2] Emran & Nurul A, *Data completeness measures*, *Pattern Analysis, Intelligent Security and the Internet of Things*, 117–130, 2015.

[3] Tahani Aljuaid, & Sreela Sasi, Proper imputation techniques for missing values in data sets *International Conference on Data Science and Engineering (ICDSE)*, 2016.

[4] R. S. Somasundaram, & R. Nedunchezhian, Evaluation of Three Simple Imputation Methods for Enhancing Preprocessing of Data with Missing Values, *International Journal of Computer Applications*, 21(10):14–19, 2011.

[5] Hyun Kang, The prevention and handling of the missing data, *Korean Journal of Anesthesiology*, 402–406, 2013.

[6] Peter Schmitt, Jonas Mandel & Mickael Guedj, A Comparison of Six Methods for Missing Data Imputation *Journal of Biometrics & Biostatistics*, 6(1), 2015.

[7] Marsh, & H. W., "Nonpositive definite matrices, parameter estimates, goodness of fit, and adjusted sample sizes" in *Pairwise deletion for missing data in structural equation models. Structural Equation Modeling: A Multidisciplinary Journal* 5, 22–36, 1998.

[8] Soley-Bori & Marina, Dealing with missing data: Key assumptions and methods for applied analysis, *Boston University*, 2013.

[9] Susianto, Y and Notodiputro, KA and Kurnia, A and Wijayanto, H, "A Comparative Study of Imputation Methods for Estimation of Missing Values of Per Capita Expenditure in Central Java" in *IOP Conference Series: Earth and Environmental Science*, 58(1), 012017, 2017.

[10] Allassonniere, Stéphanie and Kuhn, Estelle, Convergent Stochastic Expectation Maximization algorithm with efficient sampling in high dimension. Application to deformable template model estimation, *Computational Statistics & Data Analysis*, 91, 4–19, 2015.

[11] Rahman, Md Geaur, Islam and Md Zahidul, Missing value imputation using decision trees and decision forests by splitting and merging records: Two novel techniques, *Knowledge-Based Systems*, 53, 51–65, 2013.

[12] Patidar, Preeti and Tiwari, Anshu, Handling missing value in decision tree algorithm, *International Journal of Computer Applications*, 70(13), 2013.

[13] Gavankar, Sachin and Sawarkar, Sudhirkumar, Decision Tree: Review of Techniques for Missing Values at Training, Testing and Compatibility, *Artificial Intelligence, Modelling and Simulation (AIMS), 2015 3rd International Conference on*, 122–126, 2015.

[14] Gerardnico, Data Mining - K-Nearest Neigbors, *CC Attribution-Noncommercial-Share Alike 4.0 International*, 2017.

[15] Beretta, Lorenzo and Santaniello, Alessandro, Nearest neighbor imputation algorithms: a critical evaluation, *BMC medical informatics and decision making*, 16(3):74, 2016.

[16] Kenward, M. G., The handling of missing data in clinical trials, *Clinical Investigation*, 3(3):241–250, 2013.

[17] Liu, Yuzhe and Gopalakrishnan, Vanathi, An Overview and Evaluation of Recent Machine Learning Imputation Methods Using Cardiac Imaging Data, *Data*, 2(1):8, 2017.

[18] Glickman, Mark E and Van Dyk, David A, Basic bayesian methods, *Topics in Biostatistics*, 319–338, 2017.

[19] Barga, Roger and Fontama, Valentine and Tok, Wee Hyong and Cabrera-Cordon, Luis, Predictive analytics with Microsoft Azure machine learning, 2015.

Introducing a Cybersecurity Mindset into Software Engineering Undergraduate Courses

Ingrid A. Buckley, Janusz Zalewski
Department of Software Engineering
Florida Gulf Coast University
Fort Myers, FL, USA

Peter J. Clarke
School of Computing and Information Sciences
College of Engineering and Computing
Florida International University
Miami, FL, USA

***Abstract*—Cybersecurity is a growing problem globally. Software helps to drive and optimize businesses in every aspect of modern life. Software systems have been under continued attacks by malicious entities, and in some cases, the consequences have been catastrophic. In order to tackle this pervasive problem, emphasis has been placed on educating software developers on how to develop secure systems. The majority of attacks on software systems have been largely due to negligence, lack of education, or incorrect application of cybersecurity defenses. As a result, there is a movement to increase cybersecurity education at all levels: novice, intermediate and expert. At the college level, students can be exposed to cybersecurity skills and principles that will better equip them as they transition into the workforce. A case study is presented which assesses the cybersecurity knowledge of juniors and seniors in a software engineering degree program taught over a one-semester period.**

Keywords—Cybersecurity; security education, software testing; computer security; defect detection, software maintenance

I. Introduction

Software continues to impact all aspects of our lives, including the way we use our phones, computers, home appliances, medical devices, and cars, just to name a few. Cybersecurity has been essential in the development of software due to the continued attacks and exploitation techniques that are performed by malicious entities over the Internet. Due to the ubiquitous nature of software, there is a great demand for skilled software developers.

Cybersecurity is an important element of software development and is an essential process to help prevent or reduce defects and vulnerabilities that can be exploited. Software vulnerabilities and defects have caused significant losses and inconveniences when systems fail or are exploited by hackers across different domains such as health care, financial, government, telecommunications and transportation systems. In general, software developers, testers and programmers are not experts on security. They implement systems that are not equipped to defend against cyber-attacks as they tend to only focus on ensuring that requirements have been adequately implemented. From a business point of view, the cost of cyberattacks are high; they increase maintenance costs, negatively impact customer perception of a product and lead to loss in profits.

However, programmers are now expected to consider threats and vulnerabilities, and to implement applications and programs that cannot be easily attacked or exploited. This is especially true for students who are not yet experienced in software development, or in cybersecurity. This lack of cybersecurity knowledge is a major issue in software development. It has been proven over the years that, software defects account for huge losses [1]-[3] and rework when security is not considered or poorly implemented. At the course level, it is important to motivate students to take a responsible approach to software development by teaching them how to test with the basic goal of evaluating and identifying defects [4].

Due to our reliance on software, there is a great need to educate and equip students with effective cybersecurity skills and knowledge. In this paper, a study is conducted to find an effective approach to expose undergraduate students to security principles. The goal of this exercise is to determine how well students can evaluate control structures by determining the correct output and, identifying defects. The specific objective of this paper is to determine how to increase cybersecurity knowledge of novice software developers which include university juniors and seniors in programming focused courses. The rest of the paper proceeds as follows. Sections 2 presents related work. Section 3 presents the case study and an evaluation of the students' performance. Section 4 discusses future work and Section 5 concludes the paper.

II. Related Work

Due to the urgency to increase cybersecurity awareness, skills and knowledge worldwide, colleges and universities, in particular, have implemented a variety of efforts to teach students about cybersecurity in software development and programming. Chen [5] proposed a teaching tool, called SWEET (Secure Web Development Teaching), for undergraduate and graduate computing courses. SWEET features virtualized web servers and a platform that allows instructors to teach security issues in web application development within undergraduate and graduate courses. This project included a laboratory exercise where students learned how to create a self-signed web server certificate. The goal of this exercise is to guide students on how to create a public and private key pair, a Secure Socket Layer (SSL) certificate and a certificate signing request (CSR). In the security exercise given in this study, students are not developing or creating

This material is based upon work supported by the National Science Foundation under Grant number DUE-1562773.

something new. Instead, they focus on assessing existing code and identifying defects that may be already existing. This provides an alternative way of learning and considering security by assessing existing code.

Similarly, Scheffler [6] designed two projects that use real world scenarios within public key infrastructure and web of trust modeling. They used several secure cryptographic algorithms that were assigned to students for implementation. The objective was to expose and teach students how to implement cryptography concepts in real world applications. Scheffler's work focuses on developing security based application from inception, whereas, the security exercises used in this paper focus on students' evaluation of existing code to uncover defects or defects in its logic.

Peltsverger [7] developed a bottleneck analysis lab with virtual network emulation environment. The lab consists of real work practical exercises using NetKit. The lab is designed to teach students how to set up a virtual network, capture traffic and analyze system performance. The lab exercises reinforced lectures and helped students to better understand computer network security concepts and challenges. Peltsverger's approach is similar to the security exercises utilized in this paper, in that it allowed students to analyze the outcome by reviewing the system performance. In this work, students analyze existing code manually and determine what the correct output should be given a specific input.

Chi et al. [8] implemented modules for teaching secure coding practices to STEM students. The modules were designed to provide fundamental secure programming skills to programmers and application developers. They used static-analysis tools to help with detecting vulnerabilities such as buffer overflows in code. Their aim was to increase security awareness by exposing a variety of students from different STEM disciplines to security principles, techniques and tools. The work by Chi et al. work is similar to the work completed in this study, except that they utilized tools to detect or uncover vulnerabilities in the code. In security exercises in this study, students analyze small code blocks manually to identify defects and determine the correct output given a set input.

Kumaraguru et al. [9] developed a system and game to teach users about phishing to help them make better trust decisions. They developed an email-based anti-phishing system called "PhishGuru", and an online game called "Anti-Phishing Phil", that teaches users how to use cues in uniform resource locators (URLs) to avoid falling for phishing attacks. The results from the PhishGuru studies suggest that the current practice of sending out security notices is ineffective. However, hands-on training can effectively teach people how to avoid phishing attacks. Similarly, the Anti-Phishing Phil exercise demonstrated that participants who played the game performed better at identifying phishing Web sites. Kumaraguru et al. used gamification to educate users about how to avoid phishing attacks. The security exercises in this study are geared towards students who will have to either develop, repurpose or maintain existing software. As a result, the exercise in this study focuses more on assessing existing code to determine defects that can be exploited by a hacker.

III. Security Case Study

In this section a description is given of the security case study completed in two software engineering courses consisting of university juniors and seniors. The primary objectives of this study are to assess (a) the overall cybersecurity knowledge of students, and their (b) ability to identify faults and defects and (c) aptitude to evaluate existing code.

A. Preliminary Work

Buckley [4] proposed a teaching strategy which leverages the use of basic data structures to teach the fundamentals of software testing principles. Software testing is an important phase in implementing secure code. In this approach, students must first understand the fundamental properties and constraints of various data structures and a recursive problem. The idea is to encourage students to fully understand the core properties and constrains of a system; this is analogous to understanding the security requirements of a system. This aspect is imperative in order to write effective test cases to uncover faults and defects. In this project, students are given the exercises to write test cases that ensure that each data structure's properties and constraints are upheld throughout implementation to avoid defects and faults that can be exploited in the future. The initial material which sparked the idea for this project is presented in Table I.

TABLE I. Data Collected from Software Testing Studens in Spring 2016

	Pre/Post-test correct responses											
	Ques. 1	Ques. 2	Ques. 3	Ques. 4	Ques. 5	Ques. 6	Ques. 7	Ques. 8	Ques. 9	Ques. 10	Average	**Std. Dev.**
Pre	63%	67%	86%	53%	55%	16%	84%	53%	61%	31%	**57%**	**0.354**
Post	79%	98%	94%	81%	65%	40%	88%	56%	94%	46%	**74%**	**0.144**

Forty nine (49) students completed the pre-test, while forty eight (48) completed the post-test. Overall, the results of the study showed a significant improvement in the post-test rate (74%) versus (57%) with standard errors of 14.4% and 35.4%, respectively. The average post-test results show a 30% improvement over the average pre-test results. The variation in the proportion of correctly answered questions decreased by 59%; i.e. from 35.4% to 14.4%.

B. Student Background and Aptitude

This case study includes university juniors and seniors who are completing a software engineering degree program. The juniors were enrolled in a data structures and algorithm course which is offered in the spring semester of their junior year. The seniors were enrolled in a software testing course taken in the final semester of their degree program. All the students involved in this study completed programming courses using Java, C and/or C++ in prior semesters. In the data structures and algorithms course, the students are taught different data structures (stacks, queues, binary trees, linked list, etc.) and how to determine the efficiency of algorithms (Big O notation).

In software testing, the students are taught various testing techniques including unit, integration, systems, regression and acceptance testing. They are taught blackbox and whitebox testing techniques, and utilize statement and branch coverage tools. All seniors in this study had already completed data structures and algorithms the previous spring. Additionally, most of the seniors had completed at least one internship experience that involves programming, testing or some other aspect of software development.

C. Cybersecuity Pretest and Posttest

The pretest and posttest were designed to assess students' knowledge of inspecting and evaluating small blocks of code. This exercise challenges students to carefully evaluate code to find faults and defects by determining what the expected output should be. The objective of this knowledge area is to increase software quality by discovering and correcting faults and defects that can be exploited via cyberattacks. Additionally, it illustrates to students how bad programming habits or confusing code can lead to vulnerabilities and defects that are exploitable.

The questions on the pretest and posttest are based on scenarios that provide some hands-on relatable examples that will challenge students to carefully examine basic code that may have defects infused in inconspicuous areas of the code. Each scenario is accompanied by a flow chart to further illustrate the logic as shown in Fig. 1 and 2. The pretest and posttest consist of 11 questions based on two different scenarios. The students were given the pretest at the beginning of the course; they were also given the same test at the end of each respective course. The problem scenarios are summarized below:

- Scenario1 - test a method that takes input as a decimal number and returns a string of "pass" or "fail". Assume that a grade of 70 or higher leads to a "pass", and a grade below 70 leads to "fail"'. All valid grades fall into the range of [0, 100]; otherwise, a grade leads to "fail".
- Scenario2 - test a method that takes input as a decimal number and returns a string of a letter grade based on the grade scale in Table II and Fig. 2.

TABLE II. GRADING SCALE

Grades	Return
90-100	"A"
80-below 90	"B"
70-below 80	"C"
60-below 70	"D"
0-below 60	"F"

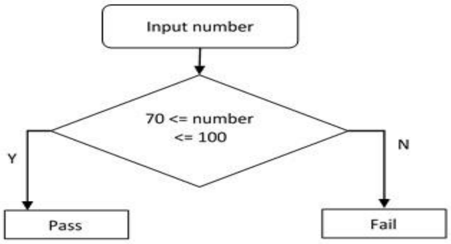


Fig. 1. Scenario 1 flow chart.

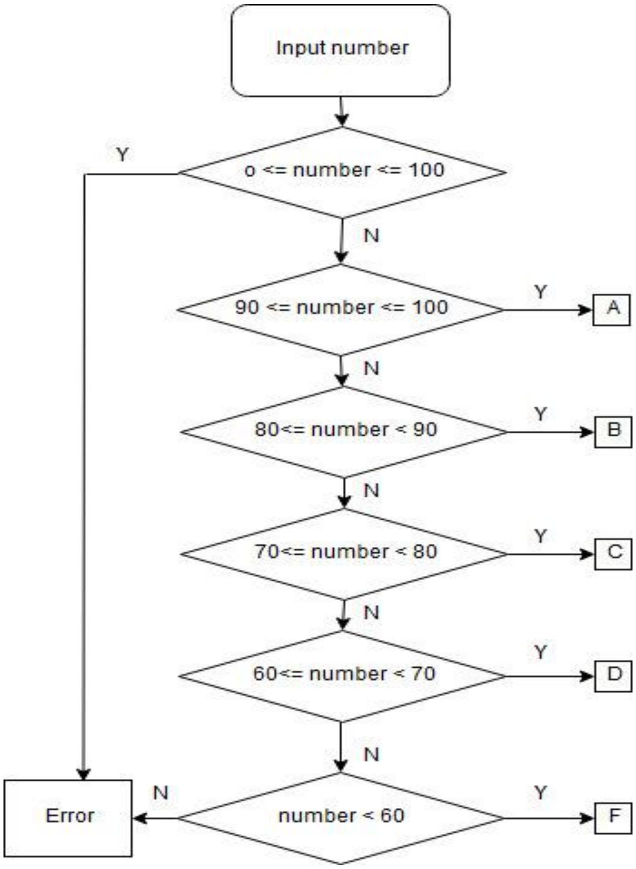


Fig. 2. Scenario 2 flow chart.

D. Evaluation of Results

A total of sixty five (65) students completed the pretest and posttest, which included twenty nine (29) juniors and thirty six (36) seniors. The juniors and seniors were enrolled in programming centric courses, namely data structures and algorithms course and software testing respectively. Overall, the results in Table III illustrate that there is a 21% increase in the mean score of the juniors versus 13% for seniors. There is a 20% increase in the median score of the juniors versus 14% for the seniors. There is a 30% increase in the standard deviation score of the juniors versus a 1.34% decrease for seniors. Both groups obtained the same posttest score which is roughly (8/11), even though the juniors had lower pretest scores.

A more detailed statistical analysis of the results show that there is significant difference between the pretest and posttest for both the juniors and seniors in Table III. Using a paired sample t-test the results for the juniors are $t(26) = 5.51$, $p < 0.01$, which shows significance, and for the seniors the results are $t(35) = 4.12$, $p < 0.01$. Note that there was no significant difference on the pretest between juniors and seniors based on the equality test of variances.

E. Discussion

Overall, both juniors and seniors scored comparably on the posttest. However, juniors achieved a higher percentage of improvement between the pretest and posttest; that is, seniors achieved a lower percentage improvement. Since seniors typically have more development experience and knowledge than juniors, they performed slightly better on the pretest. Despite having two different proficiency levels, both groups showed an improvement in their abilities to detect defects, faults and to determine correct output.

TABLE III. COMPARISON PERFORMANCE BETWEEN JUNIORS AND SENIORS GROUP

	Juniors			Seniors		
	Pretest	Posttest	% Change	Pretest	Posttest	% Change
Mean:	6.68	8.07	20.81	7.02	7.91	12.68
Median:	6.67	8.0	19.94	7.0	8.0	14.29
Std. Dev:	0.88	1.15	30.47	0.97	0.96	-1.52
The maximum score on this exercise is 11.						

Even though the problem scenarios used were basic familiar exercises, the majority of students were not able to answer all of the questions correctly. The exercises were designed to test each student's ability to thoroughly understand the code. Only three (2 seniors and 1 junior) of the sixty five (65) students who completed the exercise answered 90% of the questions correctly on the posttest. Even though the majority of students' scores improved between pretest and posttest, only 4.6% were able to identify the correct output and the majority of the faults.

Threats to Validity: One of the main threats to validity is the different educational levels of the students, it is expected that students in their senior year would have been exposed to the type of problems in the pretest more often than the juniors. This fact is shown in the better performance by the seniors in the pretest. Given that the sample was not randomly selected from the entire student population, it would be difficult to make a generalization based on the students' performance. In addition, it may be a stretch to claim that the sample questions used in the pre and posttest is reflective of the total skill set associated with cybersecurity concepts.

IV. FUTURE WORK

The focus of this study is to encourage students to evaluate existing code with the aim of identifying faults, defects, and assessing their understanding of existing code. This exercise is important, primarily because testing and maintenance are crucial aspects in the software development life cycle; it teaches students to refine their skills on how to approach testing and modification of existing code. Given the results of this preliminary study, another study will be undertaken which considers the aptitude level, grade point average (GPA), programming skill level, knowledge, and experience of each student participating in the study. This additional data provide a benchmark of where students are in their knowledge and skill level. It will also allow for a richer evaluation of their performance, knowledge gain and challenges or obstacles that impact their skill set and knowledge. Additionally, this data will help in identifying what factors and prerequisite knowledge contribute most in preparing or aiding students to better understand existing code with the aim of identify faults and defects.

We also plan to perform additional studies using the Software Engineering and Programming Cyberlearning Environment (SEP-CyLE) [10] that contains cybersecurity learning content. The learning content is in the form of digital learning objects (LOs) and tool tutorials. A learning object (LO) is a module of content that usually requires 2 to 15 minutes for completion, is self-contained, interactive, reusable and can be aggregated [11]. SEP-CyLE also supports embedded learning and engagement strategies that motivate students to interact with SEP-CyLE and access the learning content. The learning and engagement strategies include: collaborative learning, gamification, and social interaction [11].

With the use of SEP-CyLE, a comprehensive assessment of a student's cybersecurity knowledge and expertise can be designed. In that, students will complete a variety of cybersecurity focused learning objects (LOs) and the following data can be collected about a student's learning tendencies such as the (i) time taken to complete a LO, (ii) number of LOs attempted, (iii) number of LOs passed, (iv) number of LOs failed, and (v) total number of virtual points gained. SEP-CyLE has been adopted and used in various studies [11] as an effective supplemental tool and resource that supports students learning and instruction.

V. CONCLUSION

The study presented in this paper concentrated primarily on detection and evaluation, which are fundamental in achieving a secure system. The ability to detect and correct faults and defects is an important skillset that is essential for software developers and testers to acquire. In light of this fact, the exercises were deliberately given to seniors and juniors who were enrolled in software development focused courses. The results showed that juniors achieved a higher percentage improvement between the pretest and posttest, while seniors showed a lower percentage improvement. Both juniors and seniors scored comparably on the posttest and showed improvement in their abilities to detect bugs, faults and determine correct output. Additionally, only 4.6% of students answered 90% of the questions correctly. Although the exercises are simple, the results show that there is value in integrating security knowledge and practical skills in select courses. This exercise shows that a student's knowledge of security can influence the quality of the programs and systems they develop.

ACKNOWLEDGMENT

This work was supported in part by the 2016 Cybersecurity Summer Research and Training for College Faculty, led by Dr. Nasir Memon at New York University, Tandon School of Engineering Computer Science and Engineering Department.

REFERENCES

[1] W. Du, A.P. Mathur, "Categorization of Software Errors that led to Security Breaches", In Proceedings of 21st NIST-NCSC National Information Systems Security Conference, Arlington, Virginia, October 5-8, 1998, pp. 392-407.

[2] R. Telang and S. Wattal, "An Empirical Analysis of the Impact of Software Vulnerability Announcements on Firm Stock Price", In

Proceedings of IEEE Transaction on Software Engineering, Vol.. 33, No. 8, pp. 544-557, August 2007.

[3] Symantec Corporaton, "Internet Security Threat Report", Vol. 21, Mountain View, Calif., April 2016. Last Accessed: June 26, 2018, https://www.symantec.com/content/dam/symantec/docs/reports/istr-21-2016-en.pdf

[4] I. A. Buckley and W. S. Buckley, "Teaching Software Testing using Data Structures", International Journal of Advanced Computer Science and Applications (IJACSA), Vol 8, No. 3; 2017.

[5] L. Chen, L. Tao, X. Li, and C. Lin, "A Tool for Teaching Web Application Security", In Proceedings of the 14th Colloquium for Information Systems Security Education, Baltimore, Maryland, June 7 - 9, 2010.

[6] P. Scheffler, M. Hylkema, A. Temkin, "Putting It All Together: Theory and Practice in Courses on Cryptography", In Proceedings of the 14th Colloquium for Information Systems Security Education, Baltimore, Maryland June 7 - 9, 2010.

[7] S. Peltsverger, C. Zhang, " Bottleneck analysis with NetKit: teaching information security with hands-on labs", In *Proceedings of the 15th Annual Conference on Information technology education* (SIGITE '14). ACM, New York, NY, USA, 45-50, 2014.

[8] H Chi, E. L. Jones, and J Brown, "Teaching Secure Coding Practices to STEM Students" In *Proceedings of the 2013 on InfoSecCD '13: Information Security Curriculum Development Conference* (InfoSecCD '13). ACM, New York, NY, USA.

[9] P. Kumaraguru, S. Sheng, A. Acquisti, L. F. Cranor, and J. Hong, "Teaching Johnny not to fall for phish", *ACM Trans. Internet Technol.* 10, 2, Article 7, June 2010.

[10] R. Chang-lau and P. J. Clarke. Software engineering and programming cyberlearning environment (SEP-CyLE), 2018. Last Accessed: June 28, 2018, https://stem-cyle.cis.fiu.edu/

[11] I. A. Buckley, P. J. Clarke, "An approach to Teaching Software Testing Supported by Two Different Online Content Delivery Methods", In proceedings of 16th LACCEI International Multi-Conference for Engineering, Education, and Technology, "Innovation in Education and Inclusion" Lima, Peru July 18 – 20, 2018 (to appear).

Trust and Security Concerns of Cloud Storage: An Indonesian Technology Acceptance

Nurudin Santoso, Ari Kusyanti
Department of Information Technology
Universitas Brawijaya
Malang, Indonesia

Harin Puspa Ayu Catherina, Yustiyana April Lia Sari
Department of Information System
Universitas Brawijaya
Malang, Indonesia

***Abstract*—Cloud drive is a service that offers data storage on the cloud. As the worldwide rapid growth of cloud drive there are ongoing concerns about trust, privacy and security concerns about how the user's personal information and data are visible to other users or even abused by the cloud drive provider. This study provides empirical evidence about the factors affecting the acceptance of cloud drive users by using seven construct variables which are Trust, Perceived Risk, Perceived Ease of Use, Perceived Usefulness, Security, Behavioural Intention and Subjective Norm. Data were collected from 294 respondents by using online questionnaire. The data analysis method used was Structural Equation Modelling (SEM) analysis. The results of this study show that the factor affecting the intention of using cloud drive are trust, perceived risk and subjective norm.**

Keywords*—Cloud drive; structural equation modeling (SEM); trust; security; risk; behavior intention

I. Introduction

Information technology has developed to adjust to time. Currently, new technologies have been developed to improve productivity, including data storage. Cloud drive is a service to store documents or files for free or paid depending on the amount of storage capacity offered by the provider. Cloud drive allows users to store files on their servers, synchronize files across devices, and share files. Not only that it has inexpensive price but this provider also provides the whole package of office applications. Cloud drive has a data storage capacity of up to a several Giga bytes. This will certainly make it easier for users to access the data they hold.

When users decide to store their files on cloud drive indirectly the data will be owned by the cloud provider. It may pose a risk to users losing their data. Many internet users are not aware of this risk, users will generally only choose comfort and convenience without taking into account data security although in fact they sometimes feel uncomfortable when providing personal data. Apart from the benefits provided by the cloud drive, a serious risk associated with the use of data storage services is also anticipated. As an example in 2014, Google announced that some data were leaked through the URL stored in Google drive [1]. Such case is certainly contrary to the promise of security provided by Google drive. Furthermore in 2015, almost 5 million Gmail accounts were hacked which means that hackers got access to Google drive data of nearly 5 million people. More than that, the hacked database was dumped on various public forums for other people to access users' data. This can result to personal information theft, identity theft, stolen blueprints and much more.

The model in this study adapted from several previous studies, including a study conducted by [2] entitled "Personal Cloud User Acceptance: The Role of Trust and Perceived Risk in the Technology Acceptance Model" which examines the individual user acceptance of cloud computing, the model used has 5 constructs, namely: trust, perceived risk, perceived ease of use, perceived usefulness and behavioral intention. In addition, another variable that was adapted from the research done by [3] is a security that will be used to measure the level of security when using data storage on cloud drive. The purpose of this study was to determine whether the factor perceived risk, trust, perceived ease of use, perceived usefulness, perceived risk, security and subjective norm affect the intention of users to store their data on cloud drive.

II. Model Structure and Hypothesis

This research is confirmatory research based on model and hypothesis by [2] and [3]. The data was analyzed using Structural Equation Modelling (SEM). There are two stages in this SEM analysis: structural model and measurement model. Structural model shows the relationship between latent variables, while measurement model is used to determine the relationship connection between indicator and variables.

A. Definition of Each Construct

1) Trust (TR)

In this study, trust is defined as an individual's willingness to provide their personal information at risk while in a state of uncertainty [4].

2) Perceived Risk (RI)

Risk can be defined as an individual's beliefs about the possibility of gains or losses associated with the acquisition of products or services online [5].

3) Perceived Ease of Use (PEU)

Perceived ease of use defined as the extent to which an individual expects the use of a technology is free of effort [6].

4) Perceived Usefulness (PU)

Perceived usefulness is defined the extent to which an individual believes that using a technology will be able to improve their job performance [6].

5) Security (SC)

Security can be defined as the belief of the individual against the security level of a particular technology [7].

6) Subjective Norm (SN)

Subjective norm defined the extent to which an individual perceives that the other person whom is important to them assure them to use the new technology [8].

7) Behavioral Intention (BI)

Behavioral intention is defined as an individual's willingness to keep using a technology [9].

B. Hypothesis for the Construct

According to [10] it has been empirically validated that the trust of individuals towards a technology will lower their risk perceptions on the technology. From this statement, it can be drawn hypothesis as follows:

1) Trust has a Positive Effect towards Perceived Risk

In research conducted by [3] states that trust is an important factor as a determinant when the user decides to use an application. When an application is a trustworthy application it will be many users who decide to use the application. In addition, when the application is a trustworthy application it will create interest from users of the application to use the application. From this statement, it can be drawn hypothesis as follows:

2) Trust has a Positive Effect towards Behavioral Intention.

The more users trust a technology, the less effort they make to examine the details of the technology. On a trusted technology, users will not spend time and cognitive effort for learning the technology, read the privacy policy, term of use etc., and therefore they will see it as an easy to use technology. Some authors have shown that the influence of trust in the perceived usefulness and perceived ease of use [11]. From this statement, it can be drawn hypothesis as follows:

3) Trust has a Positive Effect towards Perceived Ease of use.

4) Trust has a Positive Effect towards Perceived Usefulness.

Perceived ease of use defined as the degree to which an individuals believes that using a particular technology can be free of effort [6]. When a technology is an easy-to-use technology and does not require much effort in its use, the users will tend to be able to feel the benefits of using the technology that indirectly will also improve the performance of users. Therefore it can be said that perceived ease of use have a positive impact towards perceived usefulness. From this statement, it can be drawn hypothesis as follows:

5) Perceived ease of use has a Positive Effect towards Perceived Usefulness.

In a study conducted by Davis [6] found that there is a relationship between perceived ease of use and behavioral intention. When users can feel the ease when they use a technology, then the user will tend to use the technology [6]. Empirical studies have recently been found that perceived ease of use has positive and significant effect on the intention to use, defined as behaviour intention [11]. From this statement, it can be drawn hypothesis as follows:

6) Perceived ease of use has a Positive Effect towards Behavioral Intention.

According to [12] shows that the relationship between the perceived usefulness and behavior intention in the context of TAM is statistically supported. When users of a technology can feel the benefits when they use the technology and the technology can improve their performance, the user will be inclined to use the technology [6]. It is hypothesized as:

7) Perceived usefulness has a Positive Effect towards Behavioral Intention.

In a study conducted by Van Slyke et al [10] found that there is a positive relationship between perceived risk and behavioral intention. When a user of a technology finds that the technology is risk-free and it can minimize the likelihood of possible risks, the user will tend to use the technology without worrying about future risks. From this statement, it can be drawn hypothesis as follows:

8) Perceived Risk has a Positive Effect towards Behavioral Intention.

Currently, security issues of a technology can be said to be very high [13]. When a technology is secure and can guarantee the security of its users, the user will tend to use the technology without worrying about possible risks. From this statement, it can be drawn hypothesis as follows:

9) Perceived Risk has a Positive Effect towards Behavioral Intention.

Subjective norm defined as the influence of an individual's social networks (e.g. family and friends) to the individual's behavior [14]. When a user gets a lot of influence from people around them to use a technology, it will affect that user to use the technology [14]. From this statement, the hypothesis can be drawn as follows:

10)Subjective Norm has a Positive Effect towards Behavioral Intention.

Based on the explanation of the hypothesis made in this study, the research model that is used in this study can be seen in Fig. 1.

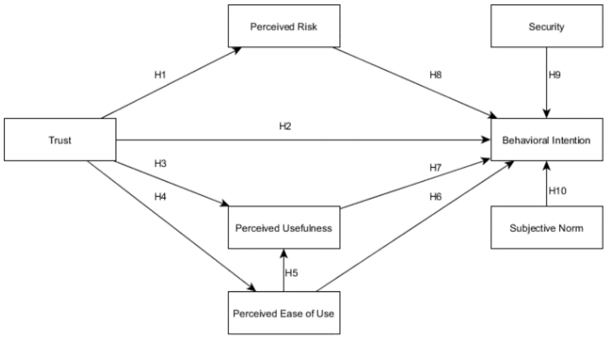


Fig. 1. Research model.

The model in Fig. 1 is used to represent the relationship between latent variables of factors affecting users to use cloud storage. Questionnaires were developed based on the model and used to obtain respondents' data.

III. Data Analysis

Structural Equation Modeling (SEM) is used to analyze data from respondents that have been collected through questionnaires. Structural Equation Modeling (SEM) provides a systematic mechanism for validating relationships between constructs and can be used to determine relationships between constructs in a model and it offers powerful and conscientious analysis technique to test a complex models [15]. The respondents of this study were all peoples whom actively using cloud drive.

A. *Descriptive Analysis*

The characteristic of respondents is shown in Table I.

TABLE I. Characteristic of Respondents

Gender	**Total**	**%**
Male	152	50.67
Female	142	47.33
Total	300	100

B. *Missing Data and Outlier*

Based on the test of the missing data conducted using Little's MCAR, there is no incomplete or missing data in this study. Outlier data can be verified by finding the Mahalanobis distance value with the error rate of 1%. In this research, it is ascertained that the obtained Mahalanobis distance value is 74.919, so that data that are exceeding this value will be removed. In this study data which exceed the value of mahalanobis distance is 30, hence the valid data is 264 out of 294 data in total.

C. *Reliability Analysis*

Reliability test is used to examine the level of consistency of an indicator when measuring its latent variables. Reliability test can be determined by using Cronbach's Alpha. The results of reliability testing can be seen in Table II.

TABLE II. Cronbach Alpha Value

Factor	**Cronbach Alpha**
Criteria	**>0.6**
Trust	0.851
Perceived Risk	0.919
Perceived Usefulness	0.949
Perceived Ease of Use	0.931
Behavioral Intention	0.668
Security	0.800
Subjective Norm	0.751

D. *Sample Adequacy Test*

According to [16] Kaiser-Meyer-Olki (KMO) of sampling adequacy test is used to examine whether the data to be used is sufficient for data analysis. In this study, the value of KMO obtained is 0.831 with a significance value of 0.000 (Sig. <.001) so it can be concluded that the variables in this study is considered great and sufficient to conduct further analysis.

E. *Normality Test*

Normality test is used to examine whether the data has been normally distributed or not [17]. Normality test can be evaluate by using the value of Skewness and Kurtosis. If the value of Skewness and Kurtosis has a value between ± 2 then it can be said that it has been normally distributed [15]. In this study, obtained Skewness and Kurtosis value within range of ± 2 then it can be said that data used in this study is normally distributed.

F. *Levene Test*

Homogeneity test is used to evaluate whether the data is homogeneous or not [18]. Homogeneity test can be examined by using Levene test. The data is considered homogeneous if it has Sig. value of > 0.05. In this study, each latent variable has a Sig. value of > 0.05 so it can be concluded that the data used is homogeneous.

G. *Measurement Model Fit*

Measurement model fit test is used to assess the correlation of the indicator and its latent variables. Measurement model fit test can be evaluated by using the value of goodness of fit indices. The results of the measurement model fit test can be seen in Table III. Based on Table III, it can be seen that all criteria have met the specified criteria.

TABLE III. Goodness of Fit Indices (GoFI) Values

Index	**Criteria**	**Value**	**Info**
Chi-square	>0.05	1238.020	Good
CMIN/DF	1.00 < CMIN/DF < 5.00	3.027	Good
Goodness of Fit Index (GFI)	>0.8	0.806	Good
Root Mean Square Error of Approximation (RMSEA)	<0.09 marginal fit	0.083	Marginal Fit

H. *Structural Model Fit*

Structural fit model test is used to assess the relationship between latent variables in the research model. Structural model fit test can be examined by using path analysis. The results of the structural model fit testing can be seen in Table IV.

TABLE IV. Structural Model Results and SEM Model Hypothesis

Hypothesis	**P-value**	**Result**
	<0.05	
RI ← TR	***	Accepted
BI ← TR	.023	Accepted
PU ← TR	***	Accepted
PEU ← TR	***	Accepted
PU ← PEU	***	Accepted
BI ← PEU	.112	Rejected
BI ← PU	.168	Rejected
BI ← RI	.045	Accepted
BI ← SC	.232	Rejected
BI ← SN	***	Accepted

Based on the results of structural model testing, it can be seen that from 10 hypotheses that have been evaluated, there are 7 accepted hypothesis and 3 rejected hypothesis.

The impact of trust (p=0.023), risk (p=0.045) and subjective norm (p=***) on behavioral intention are significant at p=0.05. Thus, H2, H8 and H10 are accepted. Meanwhile, perceived ease of use, perceived usefulness and security have no significant impact on the behavioral intention, and thus H6, H7 and H9 are rejected.

Furthermore, the impact of trust (p=***) on perceived risk is significant at p=0.05, therefore H1 is accepted. The impact of trust (p=***) on perceived usefulness is significant at p=0.05 hence H3 is accepted. It gives similar result of accepted hypothesis H4 since the impact of trust (p=***) on perceived ease of use is significant at p=0.05. Similarly, the impact of perceived ease of use (p=***) on perceived usefulness is significant at p=0.05 so H5 is accepted.

IV. Research Result

A. Discussion on Hypothesis 1

Hypothesis 1 is accepted which can be concluded that the respondents believe there is no risk posed when storing data on cloud drive. It shows that the Trust (TR) has a significant influence on the Risk (RI).

The results of this study is similar to the results of research conducted by [19] which suggests that users who already have a high sense of confidence in an application will tend not to think about the risks that can occur from the use of the application.

B. Discussion on Hypothesis 2

From the results of Hypothesis 2 testing which is accepted, can be concluded that respondents believe in cloud drive provider. It shows that the trust factor (TR) has a significant influence on the behavioral intention (BI).Respondents assume that cloud drive provider will always be honest with regard to data provided by its users, respondents believe that the data they have provided to cloud drive will not be misused and will always be protected from unauthorized access. This is the reason why respondents use cloud drive.

The results of this study is in accordance with the results of research conducted by [3] which states that when the user already has a trust in the application and assume that the application is trustworthy then the user will tend to continue to use the application.

C. Discussion on Hypothesis 3

Hypothesis 3 is accepted which shows that trust (TR) has a significant effect on Perceive Usefulness (PU). Respondents trust to store their data on cloud drive so that they can gain benefit from the use of cloud drive, such as improve their job performance.

The results of this study are similar to the results of research conducted by [20] which suggests that if users of an application have a high sense of confidence in an application they will tend to feel the various benefits derived from the use of the application.

D. Discussion on Hypothesis 4

Hypothesis 4 is accepted. It shows that trust (TR) has a significant effect on Perceived of Use (PEU). From the results of hypothesis 4, it can be concluded that the respondents believe that by storing data on cloud drive they feel the ease of use in using cloud drive.

The result of this study is similar to the results of research conducted by [20] which suggests that if users of an application have a high sense of confidence in an application they will tend to feel the application is easy to use.

E. Discussion on Hypothesis 5

Hypothesis 5 is accepted which shows that Perceived Ease of Use (PEU) has a significant effect on the factor of Perceived Usefulness (PU). From the results of hypothesis 5, it can be concluded that the respondents find the ease of use of storing data in cloud drive and its use does not require much effort.

The results of this study is supported by the results of research conducted by [20] which suggests that if a user feels that an application they use is easy to use and does not require much effort in using it, then automatically they will also be able to feel the benefits of use of the application.

F. Discussion on Hypothesis 6

Hypothesis 6 is rejected which can be concluded that respondents do not find ease of use in storing data on cloud drive. In other words, it is impractical for user to store their data in the cloud. It shows that the ease of use factor (PEU) has no significant effect on the Behavior Interest (BI).

The results of this study is similar to the results of a study conducted by [21] who argued that when a user can not feel the ease of using an application or feel that the application is difficult to use then the user will tend not to use the application.

G. Discussion on Hypothesis 7

Hypothesis testing 7 is rejected which indicates that the factor of Perceived Usefulness (PU) has no significant influence on the behavioral intention (BI). The result shows that respondents were not able to feel the benefits of storing their data on cloud drive. In other words, respondents believe that by storing their data on cloud drive does not help them to improve the quality of their daily activities.

The results of this study is align with the results of research conducted by [22] who argued that when users feel that the application they use do not benefit the user, they will tend not to use the application.

H. Discussion on Hypothesis 8

Hypothesis 8 is accepted. It shows that the factor of perceived risk (RI) has a significant influence towards Behavior Intention (BI). From the result of Hypothesis 8, it can be concluded that respondents believe that there is no risk when they store their data on cloud drive. They feel safe when using cloud drive, because cloud provider can minimize the risks that are likely to occur and harm users. Therefore, respondents have an intention to use Cloud drive.

The results of this study is supported by the results of research conducted by [23] who argued that when an application has little risk and it can minimize the occurrence of a risk to its users, then users will tend to continue to use the application.

I. Discussion on Hypothesis 9

Hypothesis 9 is rejected which indicates that the security level factor (SC) had no significant effect towards Behavior Intention (BI). Hence, it can be concluded that the respondents did not believe the level of security found in cloud drive so that it affects the intention of the respondent in storing their data on cloud drive. They believe that cloud provider cannot guarantee the security of its users 'data.

The results of this study is in line with the results of research conducted by [24] which suggests that when users feel that the application does not have a high level of security to protect users then the user will tend to choose not to use the application.

J. Discussion on Hypothesis 10

Hypothesis 10 is accepted. It shows that the factor of subjective norm (SN) has a significant influence towards Behavior Intention (BI). The result indicates that influence from others such as their family and friends can affect the respondents' interest to store their data on cloud drive.

The results of this study is similar to the results of research conducted by [25] who suggested that if a friend or family of an app user advises that user to use the app then the user will continue to use the application.

V. Conclusion

Based on the analysis result, the factors affecting cloud drive users in Indonesia to store their data on cloud drive are trust, perceived risk and subjective norm. The findings of this study reveals that cloud drive users in Indonesia do not find any usefulness and ease of use in storing their data in cloud drive. In addition they do not feel secure regarding their data on the cloud. However, they keep on storing their data on the cloud. This is due to the users trust the cloud provider despite there is a risk that endanger their data. They trust the cloud provider could manage the risk. Apart from that, the ultimate reason is that they are influenced by the people around them to store their data in the cloud.

This study is an attempt to investigate the factors affecting the acceptance of cloud drive users in Indonesia by using empirical data that were collected using quantitative research and the questionnaire method. In future studies, to help illustrating the result of empirical data, a qualitative research can be conducted to get more detailed information. Additionally, future research can be undertaken by extending the research model and is expected to examine additional factors of cloud drive acceptance. Furthermore, the extended model can be used in other cultures or countries.

References

[1] Shanahan, E. (2014). Cloud drive Latest To Leak Users' Data. Web Page: https://www.encryptedcloud.com/blog/google-drive-latest-leak-users-data/ [Retrieved on 22 January 2017].

[2] Moqbel, M., Bartelt, V. L., and Cicala, J. E. 2014. Personal cloud user acceptance: The role of trust and perceived risk in the technology acceptance model. Proceedings of Southwest Decision Sciences Institute, At Dallas, TX

[3] Shin, D.H. (2010). The effects of trust, security and privacy in social networking: A security-based approach to understand the pattern of adoption. Vol. 22 No. 5, pp 428-438.

[4] Milne, G.R., & Culnan, M.J. (2004). Strategies for reducing online privacy risks: Why consumers read (or don't read) online privacy notices. Journal of Interactive Marketing, 18(3), 15-29.

[5] Mayer, R.C., Davis, J.H., & Schoorman, F.D. (1995). An integrative model of organizational trust. The Academy of Management Review, 20(3), 709-734.

[6] Davis, F.D., Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. [pdf] University of Minnesota. 1989. Web Page : https://www.researchgate.net/publication/200085965_ Perceived_Usefulness_Perceived_Ease_of_Use_and_User_ Acceptance_of_ Information_Technology [Retrieved on 22 January 2017].

[7] Yenisey, M.M., Ozok, A.A., Salvendy, G., (2005) Perceived security determinants in ecommerce among Turkish University students. Behaviour and Information Technology 24 (4), 259–274.

[8] Venkatesh, V., Morris, M.G. (2000). Why don't men ever stop to ask for directions? Gender, social influence, and their role in technology acceptance and usage behavior. MIS Quarterly 24, 115-139.

[9] Prasarry, Y. V. (2012). Persepsi Mahasiswa Terhadap Penggunaan Internet Berbasis Teknologi Wi-Fi Dengan Pendekatan TAM. Universitas Brawijaya.

[10] Van Slyke, C., Shim, J.T., Johnson, R., & Jiang, J.J. (2006). Concern for information privacy and online consumer purchasing. Journal of the Association for Information Systems, 7(6), 415-444.

[11] Pavlou, P.A. (2003). Consumer acceptance of electronic commerce: Integrating trust and risk with the technology acceptance model. International Journal of Electronic Commerce, 7(3), 101-134.

[12] Lee, Y.C., Kozar, K.A., & Larsen, K.R.T. (2003). The Technology Acceptance Model: Past, present, and Future. Communications of the Association for Information Systems, 12(50), 752-780.

[13] Acquisti, A., Gross, R., 2006. Imagined communities: awareness, information sharing, and privacy on the Facebook. In: Golle, P., Danezis, G. (Eds.), Proceedings of 6th Workshop on Privacy Enhancing Technologies. Robinson College, Cambridge, UK, pp. 36–58.

[14] Lai F, Li D, Hsieh C-T. Fighting identity theft: the coping perspective. Decision Support Systems 2012;52(2):353–63. doi:10.1016/j.dss.2011.09.002.

[15] Chandio, F. H. 2011. Studying Acceptance of Online Banking Information System: A Structural Equation Model. London: Brunel University.

[16] Field, A. (2009). Discovering statistics using spss 3rd ed. [e-book]. Sage Publications. DOI= http://fac.ksu.edu.sa/sites/default/files/ktb_lktrwny_shml_fy_lhs.pdf.

[17] Ghozali, I. (2005). Aplikasi Analisis Multivariate dengan program SPSS, Badan Penerbit Universitas Diponegoro, Semarang.

[18] Levene. 1960. Contributions to Probability and Statistics. Standford University Press. CA.

[19] Gurung, A., X. Luo, and Q. Liao. 2009. Consumer motivation in taking action against spyware: An empirical investigation. Information Management and Computer Security 17 (3): 276–289.

[20] Carlota, L., Chiappa, G., D., Alarcón-del-Amo, M. (2014). The users adoption and usage of social network sites: an empirical investigation in the context of Italy.

[21] Juniwati, 2014. Influence of Perceived Usefulness, Ease of Use, Risk on Attitude and Intention to Shop Online. European Journal of Business and Management.

[22] Fong, K. K. K and Wong, S. K. S. 2015. Factors Influencing the Behavior Intention of Mobile Commerce Service Users: An Exploratory Study in Hong Kong. International Journal of Business and Management; Vol. 10, No. 7

[23] Grewal, D., Iyer, G. R., Gotlieb, J., and Levy, M. 2007. Developing a deeper understanding of post-purchase perceived risk and behavioral intentions in a service setting. Journal of the Acad. Marketing Science (2007) 35:250–258

[24] Topaloğlu, C. 2012. Consumer Motivation and Concern Factors For Online Shopping in Turkey. Asian Academy of Management Journal, Vol. 17, No. 2, 1–19.

[25] Hsu, M. H., Ju, T. L., Yen, C. H., & Chang, C. M. (2007). Knowledge sharing behavior in virtual communities: the relationship between trust, self-efficacy, and outcome expectations. International Journal of Human-Computer Studies, 65, 153-169.

Dimensions of Open Government Data Web Portals: A Case of Asian Countries

Sanad Aarshi[1], Babur Hayat Malik[3], Fariha Habib[4],
Kinza Ashfaq[5], Irm Saleem[6]
Department of CS & IT
The University of Lahore, Gujrat, Pakistan

Usman Tariq[2]
Department of Business Administration
University of Punjab, Pakistan

***Abstract*—Citizen Factors of the open government data are being explored in this study in the selected Asian countries. As per the open data availability countries have been selected on global open data index and well-structured open government data portals of Asian countries. To identify and analyze the differences of selected Asian countries through the principals of open government data which are eight in number, analysis the portal activities and observed the Open government data benefits. In analysis, the datasets of selected countries have been analyzed for the purpose of defining the portal activities. These activities include the Visitants, Suppliers, Applications, Developments, generation of Knowledge and overall resources utilization. Open government data of these countries are examined through web contented analysis, in order to understand the open government data's status. This study also describes different challenges on how adoption, promotion and acceptance of the open government data and portals have been carried out by Asian countries. Moreover, there are some recommendations according to the key problems and status in the open government data initiatives. Also, the study has limitations regarding the number of countries and future directions emphasize the need for Open Government Data analysis in less developed countries also.**

Keywords—Transparency; accountability; portal activities; adoption; principals of open government data which are eight in number; benefits of open government data; recommendations

I. Introduction

In the last era there have been a lot of investments and activities in the public sectors to opening up that information to the public, for political, scientific and commercial purposes. For scientific research this information can be fruitful in numerous different domains e.g. administrative, political, management sciences, economic, and social [1]. Moreover, journalists and citizens can b use it for achieving well and deep understanding of visions into the happenings and government agencies spending's. These results should be in mature, evidence-based and effective political processes. As well, open government data can have an optimistic impact on innovation and the economic growth, as they empower the development of new products, applications and services [2]. The data reusability and transparency are the two major goals. The few are, there is the Europe's Public-Sector Information Directive in the year 2003 , along with the initiatives of the United State President's Obama in the year 2009, Open Government Partnership in 2011, and the Open Data Charter G8 in 2013 [3].

National and international administrations along with the enormous political leaders have acknowledged progressively various benefits of Open Government Data. David Cameron, Prime Minister of England in May 2010, stated that his government intended to use the open data in a tactic to decrease the deficit and get improved assessment for money in the public funds expenditures, and also to obtain economic improvements substantially through by the public data establishment which has allowed companies and nonprofit governments to build inventive websites and applications [4].

There is huge lagging behind in terms of open data adoption by Asian countries at governmental level, in the inclusive data availability and in the usage of open data for the transparency, accountability and collaboration. In many Asian countries lots of datasets is available on their national portals and according to these datasets a very less numbers of applications and services are developed that are identified in this study, yet in addition information utilization level made accessible, as appeared by the altogether conversion rate of different dataset-to-application applied on the Asian countries. Adoptive Open Government Data phenomenon has led us to conclude that although it is important to open the data, which is not enough. However, it is also pivotal to promote the reutilization of the open data by the civilization and the data that is provided on the countries portals are sometimes not open.

This problem has been found hitting the Asian countries leading these countries facing problems regarding policy making at the local level and international level. For the purpose of resolving these issues it is pivotal eliminating the issue in the near future.

Moreover, this research focuses the open government data in Asian countries including Taiwan, Thailand, Hong Kong, Japan, Singapore, Israel and Bangladesh. In the past studies, these Asian countries have not given due consideration with regard to the open government data therefore, this study discusses the matter in the context of Asian countries. Researchers have given emphasis on the open government data to be studied further as the importance of it is boosting with technological advancements.

The objectives of this study include:

- To explore open government data portals in Asian countries.

- To conduct open government data's content analysis and to providing the portal activities of Asian countries.
- To analyze the open government data's challenges faced by Asian countries.
- To provide recommendations for open government data in Asian countries.

To the theoretical field, this research would be a great contribution and to practical field as well. This will enrich the literature regarding open government data in Asian countries on one side and on the other side it would enhance the ability of general government to evaluate the activities of government leading them playing pivotal role in country development through influencing political decisions.

This paper describes the comparative study of open government data emergence in the Asian countries. After this introduction, Section II this paper discus the open government data and the adoption of the open government data and then the adoption in the Asian countries and Section III there is the research methodology of selection of eight countries of Asia and the criteria used for analysis in Section IV there is data analysis phase and different technical related challenges in the Asian countries and Discussion. In Section V there is conclusion and discussion, and in Section VI there is some recommendations to overcome open government data's the challenges in the adoption.

II. Open Government Data

In the recent years, lots of open data activities sprung up around the whole world, with transparency and data reusability as two of the most important aims [5]. This Open Data movements, while recently, that initiated in order to promote the research and to discover obstacles and benefits, requirements and the technical services to promote values creation and the policy issues and implications. According to the researchers "Open government is related all around in order that transparency can be improved and also public affair's accountability" and consequently opportunities can be improved for citizens so that political decisions might have effect [6]. However, open government data's aspect of content analysis has been ignored in this research. Researchers have argued that a potential prerequisite for this is "open data", which denoted to the awareness that government data ought to freely accessible [7]. Based on the literature, open data however, public sector collaboration is not covered with its residents, only the additional information's establishment is meant by the open data [8]. Effectually, governments comprehend the utilization of Open Government Data with the perception of endorsing greater transparency, participatory and collaboration with the other different sectors of society [9]. According to researchers those are the important factors of Open Government Data for citizens [10]. Nonetheless, the principles of open government data have been ignored in this study (Table I).

TABLE I. Factors of Open Government Data

Factors	Description
Transparency	The information that what government is providing and doing is depicted in the transparency along with the accountability promotion. Moreover, Federal government maintains the additional information. That has two key aspects: citizens have the right to access info from public agencies on request, government have the responsibility to publish records of policies [11].
Participation	Engagement of public has improved the efficiency of the government where quality of decisions has also been improved. Administrative departments and organizations should offer citizens amplified chances to contribute in policy-making and to deliver their government with the welfares of their mutual skill and info.
Collaboration	In the governmental work Collaboration vigorously engages citizens. Administrative organizations and departments should utilize inventive methods, systems and tools in order to collaborate among themselves, through all stages of government, and with businesses, non-profit organizations, and private sectors individuals.

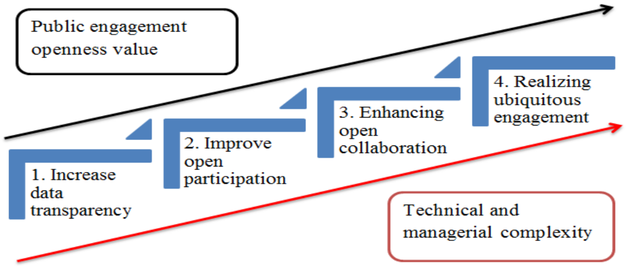

Public engagement openness value
1. Increase data transparency
2. Improve open participation
3. Enhancing open collaboration
4. Realizing ubiquitous engagement
Technical and managerial complexity

Fig. 1. The factors of open government data for citizen.

As appeared in Fig. 1, as the execution stages increases, the public involves with progressively more in government tasks and openness of government work upsurges, in this manner creating more greater values and advantages for both government and people in general. The researchers have debated as the implementation stages upsurges, the procedural and administrative unpredictability of the open government initiatives will also upsurge. However, these studies missed the aspect of challenges and recommendations for the organizations on small scale and for the country on the larger scale. Subsequently, organizations should hope to confront more prominent challenges and hazards in later implementation stages that are challenges are discussed later in Section IV.

A. Adoption of Open Government Data

Movement of the open government data has taken upward trend since last years. The Open government data perception is not just a political idea, but also an innovative government facility that is provided through Information Technology platforms. Based on the views of researchers provided in previous research it has been illustrated open government data as an entity on political level which peculated the right of citizens in order to have governmental information freely accessible via the use of digital government platforms on local as well as federal level [12]. Nevertheless, this definition does not include the elements of transparency and accountability. Open government data can be perpetuated as “the web portal which is official in nature, accessible on local and federal levels and thus aims to make the datasets which are understandable and readable in machine format with the help of internet” [13]. One of the important principals in literature is the machine readable format for open government data’s implementation. To guarantee that published data are completely usable and accessible by the end users, it is crucial that format should be machine readable [14]. These definitions juts focus the factor of data availability and analysis while decision making factor of citizen participation is ignored.

Based on the reports the first open government data portal was launched in USA in May 2009. This platform fortified government, institutions and cities to publish their ancient datasets; hence increase the aptitude of citizens to effortlessly find, share and use the composed data by government [15]. In context to describe and to merge the concept of open government data, a working group of thirty people met in California in 2007 to mature what was called the “Eight Principles of Open Government Data”, which curved to be factors for assessing open government data initiatives [16]. In the light of views presented by previous studies by providing the comprehensive and useful information and guidance, the following principles are highly recommended to provide best practices to avoid the publication of low and poor-quality data and also help in publishing the useful and meaningful data over the globe [17]. Open Government Working Group in Meeting was agreed on following principles, which was held in California in December 2007. Below is the illustration of principles [16]:

1) *Completeness of data is necessary*
2) *Data should be Primary*
3) *Data should be Timely*
4) *Data should be Accessible*
5) *Machine Process able data is necessary*
6) *There should be Non-Discriminatory access*
7) *There should be Non-Proprietary formats of data*
8) *There should be License-free data*

A couple of more facts were introducing by keeping the view of technological and political nature of data were published by the governments. However, the use of these principles for the purpose of eliminating the open data challenges has not been discussed in previous research.

A remarkable breakthrough was occurred during the Presidency of Barak Obama in 2009 when, it was recognized that open government data should be able to boost up the public interest and involvement for the transparency of the system for the enhancement in quality, efficiency and effectiveness of the public system. Reportedly, As a result of that United States introduced the Data.gov in May 2009 was known as the quality step of the OGD movement to improve the delivery of data about the federal matters for the research purpose by keeping view of government’s structure [18]. According to studies this concept was quickly penetrated into the European governments, and as a result of this revolutionary step the United Kingdom provided its online platform for the availability of data to the public [19].

B. *Adoption of Open Government Data in Asia*

Development in open data access to public is spreading all across the globe and many under developing countries are adopting this concept. Several countries have centralized portals but there is a need of getting different to hire staff and provide budget for the secure and successful publishing of data.

There is huge lagging behind in terms of open data adoption by Asian countries at governmental level, in the inclusive data availability and in the usage of open data for the transparency, accountability and collaboration. There are still some countries that don’t have their open data portals, mention below (Table II).

TABLE II. OPEN GOVERNMENT PORTALS IN ASIAN COUNTRIES

No.	Countries	Portals
1	Japan	http://www.data.go.jp/
2	India	https://data.gov.in/
3	Indonesia	https://data.go.id/
4	Thailand	https://data.go.th/
5	Singapore	https://data.gov.sg/
6	Philippines	https://data.gov.ph/
7	Hong Kong	https://data.gov.hk/
8	Malaysia	www.data.gov.my/
9	Korea	https://www.data.go.kr/
10	Pakistan	https://data.org.pk/
11	Iran	http://iranopendata.org/en/
12	Israel	https://data.gov.il/
13	Myanmar (Burma)	Not available
14	Sri Lanka	http://www.data.gov.lk/
15	Cambodia	Not available
16	Taiwan	https://data.cdc.gov.tw/en/
17	Saudi Arabia	www.data.gov.sa/en
18	Bangladesh	http://data.gov.bd/
19	Nepal	http://data.opennepal.net/
20	Afghanistan	Not available
21	United Arab Emirates	https://bayanat.ae/
22	Uzbekistan	https://data.gov.uz/ru
23	Oman	http://www.oman.om/
24	Bhutan	Not available
25	Bahrain	www.data.gov.bh/
26	Brunei	https://www.data.gov.bn/
27	Timor-Leste	http://timor-leste.gov.tl/

To publish elementary information Open government portals can be used, or electronic systems can be used to generate and continuation on specific requests. Additionally, the existing developments on opening government data boost information sharing using the open formats and standards that can be machine readable, so allowing the reuse and the exploitation of data to create public values. From the software perspective, the foremost development towards open government data was the development of portals for the open data providence. Researchers have found that they permit data detection through classify the resources, search and ability to exchange and use of information through well-documented

APIs [20]. But these portals have been reported as confronted several challenges regarding the open data usage and these challenges have not been provided in literature (Table III).

TABLE III. RANKS, SCORE AND OPENNESS ACHIEVED BY ASIAN COUNTRIES IN THE GLOBAL OPEN DATA INDEX OF 2017 (SOURCE [36])

No	Countries	Rank	Score	Openness	Asian region
1	Japan	16	61%	26%	East Asia
2	India	32	47%	13%	South Asia
3	Indonesia	61	25%	0%	Southeast Asia
4	Thailand	51	34%	6%	Southeast Asia
5	Singapore	17	60%	33%	Southeast Asia
6	Philippines	53	30%	0%	Southeast Asia
7	Hong Kong	24	51%	20%	East Asia
8	Malaysia	87	10%	0%	Southeast Asia
9	Korea	*	*	*	East Asia
10	Pakistan	72	19%	0%	South Asia
11	Iran	67	21%	0%	Middle East
12	Israel	41	41%	13%	Middle East
13	Myanmar (Burma)	94	1%	0%	Southeast Asia
14	Sri Lanka	*	*	*	South Asia
15	Cambodia	74	17%	0%	Southeast Asia
16	Taiwan	1	90%	80%	East Asia
17	Saudi Arabia	*	*	*	Middle East
18	Bangladesh	61	25%	6%	South Asia
19	Nepal	69	20%	0%	South Asia
20	Afghanistan	84	12%	0%	South Asia
21	United Arab Emirates	*	*	*	Middle east
22	Uzbekistan	*	*	*	Central Asia
23	Oman	81	14%	0%	Middle East
24	Bhutan	64	23%	0%	South Asia
25	Bahrain	*	*	*	Middle East
26	Brunei	*	*	*	Southeast Asia
27	Timor-Leste	*	*	*	Southeast Asia

Asterisk (*) are those countries that don't have any data or information on the Open Government Data Index, that's the Global ranking system.

Different countries have attained enormous echelons of open government data growth, and that can be perceived from the various ranks of positions attain on Index by respective country on the Open Government Data. Based on literature, that is some annual rankings which are being published by Open Knowledge Foundation from which countries are evaluated around the whole globe comparative to its index of OGD growth [21]. There are some facts that were gathered from some public sources to analyze the availability of government data according to the definition of Open Data. Conversely, the activities of these portals and benefits of the open government data to the Asian countries have not been given the consideration.

III. RESEARCH METHOD

As described earlier the objective of this research is to analyze the open government data portals and content analysis in Asian countries with the description of challenges and recommendations provided.

This portion of paper describes that how the measurements were taken and which areas were targeted, how those areas were adopted and which criteria was use for the analysis.

A. For Analysis Selection of Countries

Three decisions guided the selection and identification of the initiatives in order to analyze in the paper. The 1st decision was to select the countries which have not their proper open government data portals. In the 27 Asian countries 4 countries (Myanmar, Cambodia, Afghanistan and Bhutan) were eliminated.

The 2nd decision was to base the identification of the countries on the list of Global open data index and the ranking published by open knowledge foundation (https://index.okfn.org/place/). In this global ranking system some countries have not their governmental data. In the remaining 23 countries we eliminate 8 more countries (Korea, Sri Lanka, Saudi Arabia, United Arab Emirates, Uzbekistan, Bahrain, Brunei and Timor-Leste). Therefore, 15 countries have been selected which have their open government data on their web portals and also ranked in global open data index. Unfortunately, when the analysis started, it was observed that 7 countries (Indonesia, Philippines, Malaysia, Iran, Pakistan, Nepal and Oman) has some information on their governmental portals and ranked on the global open data index but they are not opened to their public. When we were in the analysis phase the Indian open government data portal was not active so we also discard it.

After all these decisions, finally seven countries were selected through their well-structured open governmental data portals and the ranking on the global open data index and openness. Those seven countries were Taiwan, Singapore, Japan, Hong Kong, Israel, India, Bangladesh and Thailand (Table IV).

TABLE IV. LIST OF COUNTRIES/INITIATIVES ANALYZED

No.	Countries	Portals
1	Taiwan	https://data.cdc.gov.tw/en/
2	Singapore	https://data.gov.sg/
3	Japan	http://www.data.go.jp/
4	Hong Kong	https://data.gov.hk/
5	Israel	https://data.gov.il/
6	Bangladesh	http://data.gov.bd/
7	Thailand	https://data.go.th/

B. Criteria used for Analysis

The following standards are used for detailed examination of facts:

1) Performed Activity Level Analysis on the Open Data Plate Forms

Five separate types of activities were inspected: the activity of open data site visitors, the activity of data providers, the actions taken by developers for building the applications, the activities correlated with knowledge discovery, and the activities related to the complete and efficient resource utilization which was available on the open data sites.

- **Open data visitor's activity** measured the strength of availability and utilization of the portal by common visitors, what provides a gesture of the utility and relatedness of the Open Government Data portal. The value accredited to this standard was collected by inspecting a element's collection for instance site analytics, targeted counter's access, posted comment's number, information quantity provided through networks community networks, and discussion for a participation quality.

- **Portal supplier's activity** measured the strength and quality of the open data plate form "feeding of process". The accreditation of value has been carried out to this standard was collected by considering and analyzing the provider of data in numbers on the portal, along with the action quality for instance the number of published data sets, as well as the readiness and of excellence the data shared. Special observation is needed for the readiness of data inspection, moreover various initiatives are considered because for someone one year data might be outdated and for other three years data means a lot.

- **Applications development related activities** measured how the open data plate form can be helpful for building the new software-based application. These applications should be available on data sites. Number of applications accredited the value to this standard which was available on the data sites as well as by the review of opinion and feedback delivered to these applications.

- **Activity related to the generation of new knowledge**. This standard aimed to assess how the data published on the Open Government Data portal has been utilized to discover latest knowledge and hidden patterns. Inspection elements accredited and gained the value to this standard which expresses the presence of sharing (of initiatives, of documents, of movements, of applications,) among the various users of portal, since we assumption has been carried out that basic gadget to knowledge development is sharing.

- **Overall resource usage available on the portal activities.** This is objected to the standard to inspect the way how multiples users attached with it are handling the data and also to evaluate the publishing standards of data and applications created and developed through the use of open data. For this, the user's activities related to application usage on portals are to be accessed.

2) *Open Government Data benefit's Observable Analysis*

Every portal was also inspected in order to find out how role is being played for acquiring the benefits which are normally obtained from Open Government Data strategies. For the measurements we supposed the following kinds of prosperities:

- **Transparency of government** published dataset's nature exhibits a valuable attempt by the government administrations to accomplish their tasks more clearly.

- **Public participation:** This type of published datasets exhibits an attempt of administration to encourage the involvement and attachment of citizens in social and political life.

- **Entrepreneurship and Innovation** the newly published data sets provide the users an ability to analyze the data and create and discover new meaning of data and build new applications and services.

- **Efficiency of Government** published dataset's types are convenient to promote the betterment of public strategies and the accomplishment of quality gains and efficiency of governmental services.

For analyzing the factors of government transparency, public participation, entrepreneurship, efficiency of government, and innovation have been analyzed through the datasets of 7 Asian countries. For this, 256 data sets of Taiwan, 1275 datasets of Singapore, 21,029 datasets from Japan, 632 from Hong Kong, 556 from Israel, 58 datasets of Bangladesh, and 1095 datasets of Thailand were analyzed.

3) *Each initiative Differentiating factors Analysis*

Apart from the inspection of the standard exhibited in (1), (2) and (3), from the beginning of the content examination on web of the portals, identification was also strained to the attention of any specific exclusive characteristic or feature that might be able to differentiate the initiatives.

IV. Data Analysis

Summary is presented in Table IV concerning the number of datasets, formats of datasets and available services or applications on the open government data portals. This table also describe conversion rate of the application dataset. Indication is given by conversion rate gives an indication of the active data utilization that is shared on the portals of countries in order to produce something. That conversion rate calculation of dataset to application was done by this formula: (number of applications/services developed / number of datasets) × 100).

There are surely huge contrasts among the activities not just in what concerns the suppliers of the data, with nations having in excess of ten thousand datasets accessible to nations that have just minimal in excess of a hundred, yet in addition in the level of utilization of the information made accessible, as appeared by the altogether various conversion rate of dataset-to-application introduced by a few nations (Table V).

The assessment's conformance level with the open government data is compressed in Table VI. For explanations of space, in the table every rule is characterized by a code as indicated by the accompanying correspondence: complete (OGD-P1), primary (OGD-P2), timely (OGD-P3), accessible (OGD-P4), machine processable (OGD-P5), non-discriminatory (OGD-P6), non-proprietary (OGD-P7), and license-free (OGD-P8).

TABLE V. DATA COLLECTED ASSOCIATED TO DATASETS AND APPLICATIONS/SERVICES ARE FOUND ON PORTALS. SEVEN PORTALS OF DIFFERENT COUNTRIES

Countries	**No. of dataset**	**Formats of datasets**	**No. of applications/ service developed and available**	**Dataset to application conversion rates**
Taiwan	*256*	csv, json, xml, pdf, geojson	02	0.78
Singapore	*1,275*	csv, pdf, kml, shp, api	14	1.10
Japan	*21,029*	csv, zip, xlsx, pdf,html and 39 other formats	36	0.17
Hong Kong	*632*	Asc, csv, gif, gml, ics and 14 other formats	27	4.27
Israel	556	xls, xlsx, zip, html, csv and 23 other formats	47	8.45
Bangladesh	58	csv, xlsx, pdf, excel, xlb, data, xls, zip	04	6.90
Thailand	1,095	csv, xls, xlsx, pdf, xml and 8 other formats	02	0.18

TABLE VI. ANALYSIS OF THE EIGHT OPEN GOVERNMENT DATA PRINCIPLES

Countries	**Eight principles of Open Government Data**							
	P1	**P2**	**P3**	**P4**	**P5**	**P6**	**P7**	**P8**
Taiwan	B	A	A	A	B	A	A	A
Singapore	B	B	A	A	B	A	A	A
Japan	B	B	A	A	B	A	A	A
Hong Kong	B	B	B	B	B	A	A	A
Israel	B	B	C	C	B	B	B	A
Bangladesh	C	B	C	C	B	B	B	A
Thailand	B	B	C	B	B	A	A	B

Legend: : A: all, B: most, C: some, D: none

The eight Open Government Data principles are applied in the initiatives sensibly as evinced by the data, showing a comprehension with respect to public institutions on how data ought to be made accessible. The third principle, Timeliness (OGD-P3), that is the aptitude to made data accessible as rapidly as essential to save its actual value, is the rule that accomplished grouping at lower terms, showing dependable specialists may have some exertion in keeping the systematic accessibility of data.

The fifth principle, Machine readability (OGD-P5), is almost good in all the countries, but is lot of other data formats that first has to transform in machine readable format before data any manipulations.

This data likewise demonstrates that, while there aren't affected distinctions among all the Asian countries tend to show unrivaled conformance levels with the open government data principles.

Perceptions results in what portal activities are concerned shown in Table VII. In the case of portal activity the difference is very important, contrarily to what happened with compliance principles of open government data, and there was tenuous difference in between the Asian countries.

TABLE VII. ANALYSIS OF PORTAL ACTIVITIES

Countries	**Portal Activities**				
	Visitants	**Suppliers**	**Applications developments**	**Knowledge generation**	**Overall utilization of resources**
Taiwan	A	A	A	B	A
Singapore	A	B	A	B	B
Japan	A	B	A	A	A
Hong Kong	B	A	B	B	B
Israel	B	A	D	B	C
Bangladesh	D	D	D	C	D
Thailand	C	D	D	C	D

Legend: : A: very good, B: good, C: enough, D: poor

The observation concerned with open government data benefits is enumerated in Table VIII.

TABLE VIII. OBSERVED OPEN GOVERNMENT DATA BENEFITS

Countries	**Observable OGD benefits**			
	Government Transparency	**Public Participation**	**Innovation and entrepreneurship**	**Government efficiency**
Taiwan	B	A	A	A
Singapore	B	A	B	A
Japan	B	A	B	A
Hong Kong	A	C	B	B
Israel	B	B	C	B
Bangladesh	B	D	D	D
Thailand	B	B	C	D

Legend: : A: very good, B: good, C: enough, D: poor

Once more, in the intended table, a distinction in between the seven Asian countries is apparent. This distinction is much disreputable for advantages of "participation of public", "government efficiency", " entrepreneurship and innovation". It is intriguing to take note for the achievable benefits, in which it is required that data is released as well as that the reutilization of open data is possible, i.e., they rely upon the effective utilization of the datasets made accessible on the portals, through the expansion of new applications and services, and accordingly it is essential that society (enterprises, public and numerous different entities) has a culture of transparency and participates in the establishment and data reutilization. Maybe, open data's social appetence may validate the distinctions found amongst Asian countries.

Perceptions consideration as a whole is carried out along with the collection of the data we are move forward to the concluding terms that the primary distinction between the quality of the open government data development at present accomplished in seven selected Asian countries lays not on the amount and sort of datasets made accessible yet primarily on the amount and way those data is salvaged, as delineated in Table IX.

TABLE IX. AVERAGE VALUES OF ASIAN OPEN GOVERNMENT DATA PORTALS

Average values	**Asian countries**
Available dataset's average number	3,557
Average number of developed applications/ services	18.8
applications conversion rate's average number	3.12

As the values appear, the average conversion rate of seven selected Asian countries is calculated and displayed. This leads us to reason that to encourage the open government data singularity, while it is basic to open the data, that isn't enough. It is additionally vital to advance the reuse of the open data by the general public.

A. Challenges of Open Government Data faced by Asian Countries

Open Government Data as a movement and theory initiated in the countries which are developed and after this now the developing Asian countries are adopting it. Moreover, open government data challenges are illustrated in Fig 2. Subsequently, they have done plethora of development in this regard and are currently countries well-ahead from the other regions in terms of both quality and quantity of datasets. But still according to the many factors, OGD initiatives of the many developing countries are still at its early stages, subsequent and numerous challenges at the implementation level [22].

- **Cost for releasing public data:** Releasing public data may incur some potential costs allied with the creation and presentation of open data that need to be deliberated [23]. This may be factual in tasks like data collection, data management and data cleaning that needs assured abilities for the human resources.
- **More complex data more barriers will come:** If the task is more multifaceted which the user desire to achieve, the more barriers will occur therefore forcing the organizations to train their human resources with such technical skills [24].
- **Appropriate infrastructures of OGD:** By opening government data, the government organizations may necessitate preparing appropriate infrastructures such as buying the new server or upgrading network infrastructures [25]. For the successively OGD initiatives, these factors comprising cost that may be the foremost obstacles for government.
- **Lack of metadata and accuracy in open data portals:** Another dispute in OGD application is data quality. Lacks of Meta standards, lack of accuracy, outdated and non-valid data are amongst the problems originate in existing data in the open government data portal [26].
- **Outdated and non-valid data in OGD portals:** Government organizations also face the data privacy issues when some of the datasets comprises personal identities, when merging with different datasets [27]. In precise, government leans towards to publish the data which is easier to collect, unstructured and incomplete.
- **Government organizations also face the data privacy issues**. In determining the adoption of OGD initiatives the level of understanding of what is OGD amongst government organizations correspondingly plays a central role. It is frightened that some government organizations just hopped into the supportive central agency's or top management decisions but not on voluntarily reasons. Because of this problem, organizations may be emancipating data only once and no supplementary movements to publish more datasets [27].
- **Pressure from public to release the data:** Technical barriers are not the only aspect in OGD adoption, on the other side, government may suffer the pressure from community to release more practical data as civil society organizations, civic hackers, citizens, non-government agencies to label a few are receiving more aware of valued datasets e.g. from the health data, crime data, government spending and transport data [28].
- **Multiple and decentralized data sources and incomplete datasets in different sectors:** There are multiple and decentralized data sources. There is not proper integration in the datasets and the incompleteness of datasets in different sectors. Information which is published is very limited and majority municipal level data portal have more data than the nation level data portals.
- **Zombie data that are dump data:** In some cases, there are lots of zombie data that are not "live" but are only databases dump which can be downloaded and present in the intended format. Even if retrieved over an online API, there is not any assurance of updates in source. None of either national portal unambiguous data updates policy or used open data portals. "Existence of information is guaranteed by the system in their databases however data updating is not guaranteed" [29].
- **Government released one-way data to public**: Governments lean towards to release the data which is one-way from the government to the public. Such data can be illustrated by applications, but there is quite limitation in the foremost online facilities and valuable applications, with public returns data to government [29].

- **Published online data through full disclosure policy:** Another aspect is a failure to Deliberate Audience, the envisioned the policy design's users of the data. Datasets can be reviewed in the policy in order to look in the datasets that are desired by public or are of concentration to user values and create the essential documents to be published online through the Full disclosure policy [30].

- **Public is not conscious about the available data on the websites:** Groups of citizens are not conscious that data is available on the websites. They also struggle in engaging with released data of government. If the government is serious with guaranteeing transparency and accountability, should notify and local participants are capacitated on how to navigate and how to access along with the utilization local government data [31].

- **Low internet diffusion:** Open data assists only the info requirements of less than 40% of the populace. In a perspective of low internet diffusion, public are dependent on other ways of information distribution to protect local government data [32].

- **Data is in raw form and not open:** Deliberated the raw material of the 21st century, the data must be refined, located, and extracted in a directive to produce value. Accordingly, in any significant sense the data is not open to the public when published in its raw form. Frequently, a normal citizen is incapable to travel the collection of available datasets due to the deficiency of essential statistical and computational expertise [33].

- **Lack of Data quality:** Data quality is a main aspect on the open government data portals. There is lots of invalid and missing data on different open portals so that's the foremost obstacles for the public and other organizations [34].

- **Attained datasets in automated way:** Attained datasets in an automated way through the application of Information Technology Systems is still a bigger challenge [35].

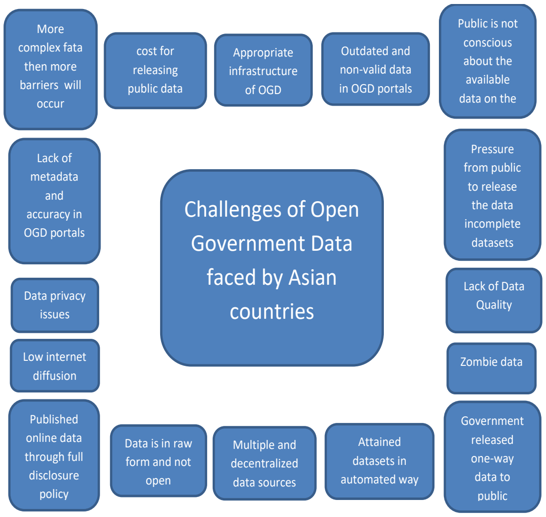

More complex fata then more barriers will occur
cost for releasing public data
Appropriate infrastructure of OGD
Outdated and non-valid data in OGD portals
Public is not conscious about the available data on the
Lack of metadata and accuracy in OGD portals
Challenges of Open Government Data faced by Asian countries
Pressure from public to release the data incomplete datasets
Data privacy issues
Lack of Data Quality
Low internet diffusion
Zombie data
Published online data through full disclosure policy
Data is in raw form and not open
Multiple and decentralized data sources
Attained datasets in automated way
Government released one-way data to public

Fig. 2. Challenges of open government data faced by Asian countries.

As was stated through this paper, technological and political urges showed to be conclusive factors in the growth of Open government data crusade. First, the acknowledgement of Obama in 2009, necessity Administration for application of the notion, which roused the rest world to this necessity and, second, the Internet, the benefits of a single resource for the information sharing. In a few years, significances of this OGD are clear: at a political phase, that delivers better transparency, collaboration and accountability to representatives and particular policies, donating to a more participatory and independent society. Moreover, the material establishes a rich resource that delivers the formation of new services and products whether envisioned for civil usage or to encourage better efficacy in governmental services themselves. Consequently, it can be said that, the open government data notion gives society the aptitude to influence public data and reutilize it for drives that encounter the desires of the parties intricate, thus endorsing the generation of novel knowledge, entrepreneurship and innovation.

The analysis demonstrates that there is still some lots of illiteracy about this notion on civilization side, which is reflected in the feeble participation of social representatives on the reprocess of the released data. Which are the motives for the feebler reprocess of opened data in Asian countries is consequently a fascinating research query to explore.

The reflections and analysis directed propose that the dissimilar appetence level exploit adopt and to accept the notion of open government data, which are demonstrated by some countries, may be owed to the presence of dissimilar ethnicities and cultural attitudes about notions for instance privacy of data. Do cultural problems have an influence on the level of data opened to civilian reprocess? Are there any kin between country size and open government data growth achievement? Or do cultural problems have an influence on the societal appetence actors to reutilize and exploit the values of open data? In future works, that we propose to achieve on this stream of research, more nations should be elaborate and various ways of collecting the data about initiatives, for example, interviews to those accountable for open government data portals, to individuals that publish datasets in open government data portals and to individuals that are using information available on open government data portals to develop the applications and services should be applied.

The study has theoretical as well as practical implications. This study is a major contribution to literature regarding the activities of data portals, principles and the benefits of open government data. Furthermore, the development of the countries has been depicted through the study which will contribute to the literature. Additionally, challenges and recommendations would help researchers find some other research related dimensions. On the other hand, the practical implications include the use of benefits and principles by the seven and other Asian countries. The less developed Asian countries would also benefit from the study.

V. Conclusion

The notion of Open Government Data swiftly developed one of the most pertinent subjects of dispute among governments of many countries around the world.

The results of the study indicate usage of data sets on the larger scale by the countries of Japan, Singapore and Thailand that depict the high scale of open data usage in these countries with the inclusion of technological development. Also, the open datasets are increasing in Hong Kong, Israel and Taiwan as well. Bangladesh is found lagging behind because of underdevelopment and less technological usage. Furthermore, the dataset conversion rate has been described depicting the data sharing on country's portal for boosting public participation in order to produce something of value. The conversion rate is for Japan and Thailand has been viewed as attractive as these countries maintain huge number of datasets on the applications that portrays proper management of open government data. In addition, the open government data principles have been analyzed. These were given the rankings of A, B, C and D where A stands for very good, whereas B for good, C for enough and D for poor. Out of 8, the principles of primary, machine process able, non-proprietary nondiscriminatory and license free are ranked as very good and good for all the countries while principles complete, timely and accessible hare regarded as good and enough. The reason behind the factors of completeness, time and accessibility being god and enough is because of the skills held by people related to the management of open data. Also, the portal activities are very good and good for Taiwan, Japan, Singapore and Hong Kong. On the other hand, these have been regarded as enough and poor for Israel, Bangladesh and Thailand. This is because the developed countries are much advanced in technology development for the purpose of data management and opening the government data. Additionally, the benefits of government transparency, public participation, innovation and entrepreneurship and government efficacy has been regarded as good and very good for Taiwan, Japan, and Singapore. Hong Kong have been regarded as very good, good and enough because of difference in use of open data by different level of governments. In the nutshell, the number of applications developed for the purpose of open government data and data conversion rate is not so high in the Asian government which faces different challenges including cost and complexity. The need for improving the mechanism has been provided through the recommendations.

The study carries limitations as in form of limited number of countries studied. Countries like Japan, Singapore and Thailand are enough advanced to the level that they use and manage the open government data. Also, the study has limitation in methodological form as it lacks the proper use of sampling technique and other forms of web portals used for open government data. The future researchers must include the less developed countries including Pakistan and Iran for the purpose of understanding the open government data in these countries so that use of open government data could be enhanced in such countries as well.

VI. Recommendations

As per the key problems and status in data initiatives of the open government in the sample areas, eight recommendations are proposed for the expansion of opening governmental data in Asian countries which can are illustrated in Fig. 3.

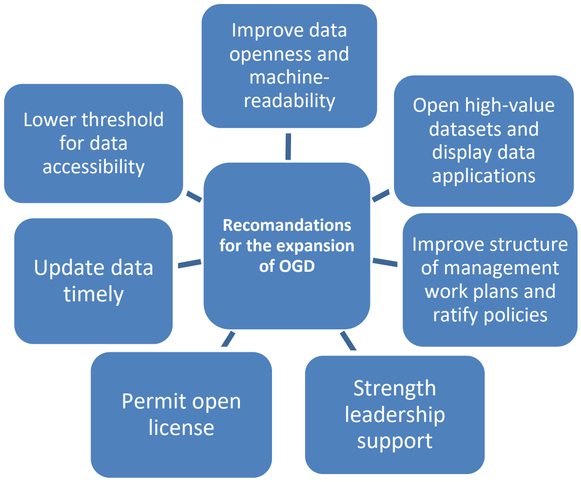

Improve data openness and machine-readability
Lower threshold for data accessibility
Open high-value datasets and display data applications
Recomandations for the expansion of OGD
Update data timely
Improve structure of management work plans and ratify policies
Permit open license
Strength leadership support

Fig. 3. Recommendations for the expansion of open government data in Asian countries.

A. Improve Data Openness and Machine-Readability

Open government data initiatives are recommended, whether still under-development and in future it has to be developed, ought to improve the machine-readable format of open data, and to certify that data are not published in different other formats for instance web pages, pictures, or PDF, etc. Intended prerequisite should not only be the work plans of the initiatives of open government data but local-policy documents should also be personified and, but also be occupied as an important indicator to evaluate all released data in several initiatives. To explicate the goals and meaning of various formats i.e. machine readable, trainings should be provided, familiarize mutual machine-readable formats, and to provide consistent tools for the help to convert non-machine-readable data into machine-readable formats. Additionally, all initiatives ought to establish endorsement process to sternly observe data formats before they are published with mutually manual and automatic approaches, so as to certifying that published data gratify the prerequisite of open formats.

B. Open High-Value Datasets and Display Data Applications

It is recommended that initiatives of open data should be concerned with user's demands and parallelized the data release departments and demanders of data (like reporters, enterprises and individual developers, etc.) there should be round table meetings in order to communicate face to face on data needs that can improve the quantity, quality form and open data worth. In addition, data release sections ought to collect public needs through shared communications via social media and websites, to constantly display developed based applications and open high-value data and on open data.

C. Improve Management Structure and Ratify Policies and Work Plans

Presently, numerous regions are short of plans, policies and actual management architectures targeted at open government data schemes. Firstly, it is recommended to create or elect capable sectors to take work responsibility related open government data and permit full authorities sectors to participate data provided by supplementary business sectors. Temporarily, partition of work and responsibilities of business sectors ought to be elucidated. Secondly, it is endorsed to formulate specifications, work plans, practical policies, and to describe requirements, principles, forms, boundaries and objects of open government data to certify the systematism and normalization of data opening. Lastly, to formulate yearly work plans and make them public is suggested, as a struggle to improve social public’s responsiveness of and contribution in the open government data.

D. Strength Leadership Support

It is distant from being sufficient to leave departments in responsibility only to endorse open government data schemes. Obvious support from domestic managers is no doubt a serious influence in lashing open data initiatives. It is recommended that local administrators ought to excavate their considerate of the drive and implication of open government data also fortify their sustenance for opening governmental data in significant local discussions, public discourses and setting up an ethos on the daily works, of open government data, growing the cognizance of open data, refining the aptitudes and providing supervision and sustenance for sectors in responsibility to the development of open data.

E. Permit Open License

Initiatives of open data ought to workout open license that are valid in Asia. The license ought to be in line with prevailing laws of their countries, but also clearly certify public the privileges to freely attain, use, and redistribute the data. Governments must focus on the obligation and importance of open license and ease data publishers and data users to develop the open license reciprocally.

F. Update Data Timely

Open data activities ought to set up comparing particulars and supervision components to guarantee that data are refreshed timely. For instance, exposure buttons could be determined to data pages to enable general society programmed investigation strategies can be built up in open data stages to naturally convey refreshing alarms to comparing departments or to report data that are neglected to be refreshed timely. Given the cost and weights of task and upkeep, data discharging departments may enable social associations or undertakings to assume the liability of looking after data. Through open private association, dynamic data can be reliably given to the general public.

G. Lower Threshold for Data Accessibility

Open data stages ought to data convenience and collaboration for lower thresholds with data holders, enable the public to obtain data and contribute in the communications without having to register. In addition, the user’s valuation on data requests, datasets and recommendations intend to be timely studied and retorted. Lastly, the practice of social media tools ought to be reinforced to timely broadcast the news and happenings of open data to the public, in directive to entirely promote and broadcast open data crusade and develop public consideration, contribution and sustenance. Furthermore, communications could arise not only among and the public and the government, but consistently between data users themselves to deliberate and interconnect on substances associated to open data and motivate more notions and produce more applications, so as to paradigm a dynamic, maintainable and well functioned ecosystem of open data.

H. Promote Innovative Ideas

Another recommendation is to promote innovative ways to successfully engross with participants to source notions and co-create resolutions and grasp the prospects provided by digital government tools, together with the usage of open government data, to sustenance the accomplishment of the aims of open government initiatives and policies.

References

[1] Alexopoulos, C., Zuiderwijk, A., Charapabidis, Y., Loukis, E., & Janssen, M. (2014, September). Designing a second generation of open data platforms: Integrating open data and social media. In International Conference on Electronic Government (pp. 230-241). Springer, Berlin, Heidelberg.J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[2] Fitzgerald, A., Hooper, N., & Cook, J. S. (2013, August). Implementing open licensing in government open data initiatives: a review of Australian government practice. In Proceedings of the 9th International Symposium on Open Collaboration (p. 39). ACM.

[3] Attard, J., Orlandi, F., Scerri, S., & Auer, S. (2015). A systematic review of open government data initiatives. Government Information Quarterly, 32(4), 399-418.

[4] Gomes, Á., & Soares, D. (2014, October). Open government data initiatives in Europe: northern versus southern countries analysis. In Proceedings of the 8th International Conference on Theory and Practice of Electronic Governance (pp. 342-350). ACM.

[5] Luna-Reyes, L. F., Bertot, J. C., & Mellouli, S. (2014). Open government, open data and digital government. Government Information Quarterly, 1(31), 4-5.

[6] Heckmann, D. (2011, January). Open government-Retooling democracy for the 21st century. In System Sciences (HICSS), 2011 44th Hawaii International Conference on (pp. 1-11). IEEE.

[7] Lathrop, D., & Ruma, L. (2010). Open government: Collaboration, transparency, and participation in practice. " O'Reilly Media, Inc.".

[8] Wijnhoven, F., Ehrenhard, M., & Kuhn, J. (2015). Open government objectives and participation motivations. Government information quarterly, 32(1), 30-42.

[9] Albano, C. S. (2013, June). Open government data: a value chain model proposal. In Proceedings of the 14th Annual International Conference on Digital Government Research (pp. 285-286). ACM.

[10] McDermott, P. (2010). Building open government. Government Information Quarterly, 27(4), 401-413.

[11] González-Zapata, F., & Heeks, R. (2016, May). The influence of the transparency agenda on open government data in Chile. In E-Democracy and Open Government (CeDEM), Conference for (pp. 156-163). IEEE.

[12] Kassen, M. (2013). A promising phenomenon of open data: A case study of the Chicago open data project. Government Information Quarterly, 30(4), 508-513.

[13] Burwell, S. M., VanRoekel, S., Park, T., & Mancini, D. J. (2013). Open Data Policy—Managing Information as an Asset. Office of Management and Budget, Executive Office of the President.

[14] Wang, H. J., & Lo, J. (2016). Adoption of open government data among government agencies. Government Information Quarterly, 33(1), 80-88.

[15] Mitrovic, Z. (2015). Building open data capacity through e-skills acquisition. Open Government Partnership Open Data Working Group. Accessed from: http://www. opendataresearch. org/dl/symposium2015/odrs2015-paper3. pdf.

[16] Corrêa, A. S., Corrêa, P. L. P., & da Silva, F. S. C. (2015, May). A collaborative-oriented middleware for structuring information to open government data. In Proceedings of the 16th Annual International Conference on Digital Government Research (pp. 43-50). ACM.

[17] DESA, U. (2013). Guidelines on open government data for citizen engagement.http://workspace.unpan.org/sites/Internet/Documents/Guidenlines%20on%20OGDCE%20May17%202013.pdf

[18] U.S. General Services Administration. 2014. DATA.GOV. http://www.data.gov/.

[19] UK Cabinet Office. 2014. DATA.GOV.UK: Opening Up Government. http://data.gov.uk/.

[20] Corrêa, A. S., Corrêa, P. L. P., & da Silva, F. S. C. (2015, May). A collaborative-oriented middleware for structuring information to open government data. In Proceedings of the 16th Annual International Conference on Digital Government Research (pp. 43-50). ACM.

[21] Gomes, Á., & Soares, D. (2014, October). Open government data initiatives in Europe: northern versus southern countries analysis. In Proceedings of the 8th International Conference on Theory and Practice of Electronic Governance (pp. 342-350). ACM.

[22] Hossain, M., & Chan, C. (2016). Open data adoption in Australian government agencies: an exploratory study. arXiv preprint arXiv:1606.02500.

[23] Ubaldi, B. (2013). Open government data: Towards empirical analysis of open government data initiatives. OECD Working Papers on Public Governance, (22), 0_1.

[24] Janssen, M., Charalabidis, Y., & Zuiderwijk, A. (2012). Benefits, adoption barriers and myths of open data and open government. Information systems management, 29(4), 258-268.

[25] Kucera, J., & Chlapek, D. (2014). Benefits and risks of open government data. Journal of Systems Integration, 5(1), 30.

[26] Nurakmal, M., Hanum, F., & Hamid, S. (2017). Post-adoption of Open Government Data Initiatives in Public Sectors.

[27] Conradie, P., & Choenni, S. (2014). On the barriers for local government releasing open data. Government Information Quarterly, 31, S10-S17.

[28] Janssen, M., Charalabidis, Y., & Zuiderwijk, A. (2012). Benefits, adoption barriers and myths of open data and open government. Information systems management, 29(4), 258-268.

[29] dos Santos Brito, K., da Silva Costa, M. A., Garcia, V. C., & de Lemos Meira, S. R. (2014, June). Brazilian government open data: implementation, challenges, and potential opportunities. In Proceedings of the 15th Annual International Conference on Digital Government Research (pp. 11-16). ACM.

[30] Helbig, N., Cresswell, A. M., Burke, G. B., & Luna-Reyes, L. (2012). The dynamics of opening government data. Center for Technology in Government.[Online]. Available: http://www. ctg. albany. edu/publications/reports/opendata.

[31] Hogge, B. (2010). Open data study. a report commissioned by the Transparency and Accountability Initiative, available for download at: http://www. soros. org/initiatives/information/focus/communication/articles_publications/publications/open-data-study-20100519.

[32] Canares, M. P. (2014, October). Opening the local: Full disclosure policy and its impact on local governments in the Philippines. In Proceedings of the 8th International Conference on Theory and Practice of Electronic Governance (pp. 89-98). ACM.

[33] Magalhaes, G., Roseira, C., & Strover, S. (2013, October). Open government data intermediaries: A terminology framework. In Proceedings of the 7th International Conference on Theory and Practice of Electronic Governance (pp. 330-333). ACM.

[34] Misra, D., Mishra, A., Babbar, S., & Gupta, V. (2017, March). Open government data policy and Indian ecosystems. In Proceedings of the 10th International Conference on Theory and Practice of Electronic Governance (pp. 218-227). ACM.

[35] Zheng, L., & Gao, F. (2016, June). Assessment on China's Open Government Data Platforms: Framework, Status and Problems. In Proceedings of the 17th International Digital Government Research Conference on Digital Government Research (pp. 408-414). ACM.

[36] Global open data index. https://index.okfn.org/