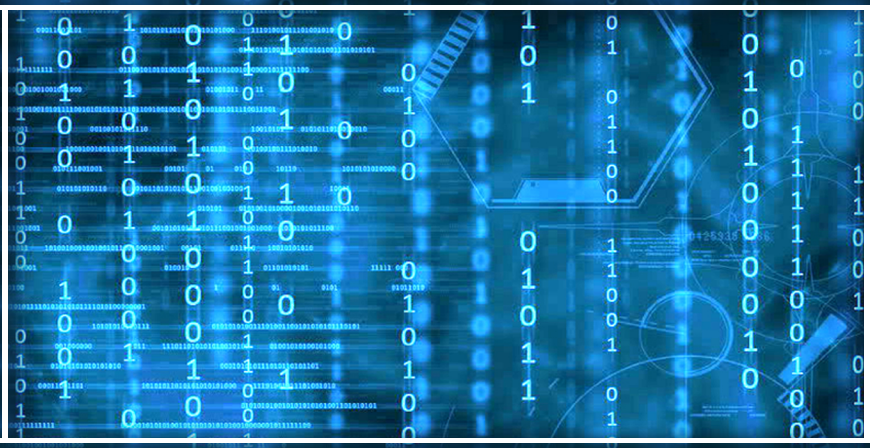
IJACSA
WHERE WISDOM SHARES
International Journal of Advanced Computer Science and Applications

Volume 8 Issue 3
March 2017



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

www.ijacsa.thesai.org

Editorial Preface

From the Desk of Managing Editor…

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

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

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

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

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

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

Thank you for Sharing Wisdom!

Managing Editor
IJACSA
Volume 8 Issue 3 March 2017
ISSN 2156-5570 (Online)
ISSN 2158-107X (Print)

Editorial Board

Reviewer Board Members

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

CONTENTS

Stylometric Techniques for Multiple Author Clustering

Shakespeare's Authorship in *The Passionate Pilgrim*

David Kernot[1 3]
[1]Joint and Operations Analysis Division
Defence Science Technology Group
Edinburgh, SA, Australia

Terry Bossomaier
[2]The Centre for Research in Complex Systems
Charles Sturt University
Bathurst, NSW, Australia

Roger Bradbury
[3]National Security College
The Australian National University
Canberra, ACT, Australia

***Abstract*—In 1598-99 printer, William Jaggard named Shakespeare as the sole author of *The Passionate Pilgrim* even though Jaggard chose a number of non-Shakespearian poems in the volume. Using a neurolinguistics approach to authorship identification, a four-feature technique, RPAS, is used to convert the 21 poems in The Passionate Pilgrim into a multi-dimensional vector. Three complementary analytical techniques are applied to cluster the data and reduce single technique bias before an alternate method, seriation, is used to measure the distances between clusters and test the strength of the connections. The multivariate techniques are found to be robust and able to allocate nine of the 12 unknown poems to Shakespeare. The authorship of one of the Barnfield poems is questioned, and analysis highlights that others are collaborations or works of yet to be acknowledged poets. It is possible that as many as 15 poems were Shakespeare's and at least five poets were not acknowledged.**

Keywords*—Authorship Identification; Principal Component Analysis; Linear Discriminant Analysis; Vector Space Method; Seriation

I. Introduction

William Jaggard first printed The Passionate Pilgrim in 1598-99, and the authorship of the 21 poems within it was attributed to William Shakespeare [1]. However, Bartholomew Griffin's 1596, Fidessa More Chaste Than Kind, already contained poem 11 [2]. Another, poem 19, appeared anonymously in Anne Cornwallis' 1580 personal notebook alongside works from Sir Philip Sidney, Sir Walter Raleigh, Sir Edward Dyer and Edward de Vere, 17th Earl of Oxford [3]. The list grows, and in 1598, Jaggard's brother John printed Richard Barnfield's,

The Encomion of Lady Pecunia, containing poems 8 and 11 [1]. By 1609, only five had been confirmed as Shakespeare's (poems 1, 2, 3, 5, and 17) having appeared in The Sonnets, or his play, Love's Labour's Lost [4]. Then, England's Helicon also printed a version of poem 20, attributing it to Christopher Marlowe, although its reply (signed Ignato) was later said to be by Sir Walter Raleigh [2]. Jaggard persisted with his claim, and in the 1612 third edition added a number of poems from Thomas Heywood, however, after complaints, Jaggard removed Shakespeare's name from the title [1]. By then, the authorship of 12 unknown poems lay in doubt, something that has remained for over 400 years.

Modern scholars are divided on the authorship of the remaining unknown twelve. Reference [5] suggests Jaggard used Shakespeare's name because the majority of the poems were Shakespeare's, including 12 unidentified poems in *The Passionate Pilgrim* said to be his earlier quality work and never meant for publishing. She also adds there is some doubt surrounding the authorship of the Barnfield and Griffin poems. Reference [6] disputes Shakespeare's authorship, while [7] suggest eight, not 12 of the anonymous poems are Shakespeare's. However, [2] suggest poems 7, 10, 13, 14, 15, 16, and 19 use a similar six-line stanza format to Shakespeare's *Venus and Adonis*, and poems 4, 6, and 9 are *about* Venus and Adonis and have Shakespearian similarities, but [5] says poems 7 and 13 resemble Robert Greene's poems.

It is interesting to note that unknown poem 12 gets little attention, even though it appears in Thomas Delany's *The Garland of Goodwill*, and entered into the Stationers Register ledger during 1592-3 [8]. When chosen by Jaggard, Delaney was living with an arrest warrant over his head because of his insightful writing during the London riots and in no position to complain [8], but what is strange are the few references in the literature to Delaney as the author until recently. Either way, Jaggard cannot be asked about the true authorship of the 21 poems, and today, the 12 poems, for the most part, remain unidentified.

Stylometric analysis, the quantitative analysis of a text's linguistic features has been extensively used to determine the authorship of the undocumented collaborations of the playwrights from the Elizabethan period, including Shakespeare [9]. There appears dissension among leading Shakespearean authorship attribution scholars about an agreed method [10], but the most successful and robust methods are based on low-level information such as character n-grams or auxiliary words (function word, stop words such as articles and prepositions) frequencies [11]. The premier work in evaluating authorship in the 16th to mid-17th centuries includes MacDonald P. Jackson, Brian Vickers, and Hugh Craig and Arthur Kinney [9]. Jackson [12] uses common low-frequency word phrases, repetition of phrases, collocation, and images to link word groups to other works. Vickers [13] uses a tri-gram, or n-gram, approach, while Hirch and Craig [14] use function word frequency and other methods, that includes ones based on word probabilities and the Information Theoretic measure Jensen-Shannon divergence (JSD) and unsupervised graph

partitioning clustering algorithms [15]. However, there are other techniques used in this period of Shakespearean analysis, including simple function words [16, 17] and word adjacency networks (WANs) [9]. However, the meaning-extracting method (MEM) from the field of psychology to extract themes from commonly used adjectives and describe a person from their personality, or self is very different [18, 19]. The authors offer a new and alternative approach to authorship identification using personality.

A. An Approach Using RPAS

In this paper, a methodology is employed that adopts a multi-faceted approach to text analysis and reveal details about a person's personality; their sense of self, from subtle characteristics hidden in their writing style [20-22]. The techniques draw on biomarkers for creativity and known psychological states [23-24] to identify characteristics within *The Passionate Pilgrim* poems. It uses a series of four indicators (**RPAS**) identified in [25] to create a stylistic signature from a person's writing: **Richness (R)** [26], the number of unique words used by an author; **Personal Pronouns (P)** [27-30], the pronouns used, closely aligned to gender and self; **Referential Activity Power (A)** [31-32], based on function words, or word particles derived from clinical depression studies; and **Sensory (S)** [33-36], five sensory measures (V-visual A-auditory H – haptic O – olfactory G - gustatory) corresponding to the senses.

RPAS is used to create individual stylistic signatures of the 21 *The Passionate Pilgrim* poems and the known works of William Shakespeare, Christopher Marlowe and Sir Walter Raleigh, Richard Barnfield, and Bartholomew Griffin are labelled. Three clustering techniques are then applied to identify the likely authorship of the 12 unknown poems within *The Passionate Pilgrim*.

II. Methodology

The Passionate Pilgrim contained within the complete works of Shakespeare [37] is used to process the data with the Stanford Parts Of Speech Tagger [38] to remove all punctuation and symbols and then aggregate the works by word frequency. *The Passionate Pilgrim* is further broken down into chunks that represent each known poem, and a decision made to follow the modern approach by editors [2], and divide poem 14 into two poems (labelled as 14 and 15) with a subsequent renumbering of the remaining poems so that there are twenty-one and not twenty poem chunks (refer to Table 1).

The 3,190-word data ends up as an aggregated matrix of 1,032 distinct word types across 21 poems, and the size of each varies between 96 and 377 words (average = 152). Putting this into perspective, they are slightly larger than a Shakespearian sonnet which varies between 91 and 132 words (average = 116).

TABLE I. The List of the Poems by Shakespeare, Barnfield, Griffin, Marlowe Including the 12 Unknown Authored Poems in the Passionate Pilgrim Poems by Author and Abbreviated ID

ID	Abbreviated	Author
1	1S	William Shakespeare
2	2S	William Shakespeare
3	3S	William Shakespeare
4	4U	Unknown
5	5S	William Shakespeare
6	6U	Unknown
7	7U	Unknown
8	8B	Richard Barnfield
9	9U	Unknown
10	10U	Unknown
11	11G	Bartholomew Griffin
12	12U	Unknown (Thomas Delaney)
13	13U	Unknown
14	14U	Unknown
15	15U	Unknown
16	16U	Unknown
17	17S	William Shakespeare
18	18U	Unknown
19	19U	Unknown
20	20M	Christopher Marlowe and Walter Raleigh
21	21B	Richard Barnfield

A 1613 play written after Shakespeare ceased writing is used to provide an independent author perspective and clustering technique. *The Tragedy of Mariam, the Fair Queen of Jewry* by English poet and dramatist, Elizabeth Cary [39], was published 14 years after *The Passionate Pilgrim*, and stylistically very different to Shakespeare's work.

A nine-dimensional array is created from the data using RPAS before applying three complementary techniques to reduce any single bias and overlay the results against Richness (R) and Personal Pronoun (P) to determine the possible authorship of the 12 unknown poems. As a final measure, seriation, an exploratory combinatorial data analysis technique, is used to visualise the nine-dimensional array as a one-dimensional continuum and test the strength of the co-located cluster edges by adding random noise to the data vector.

A. Three Complementary Techniques

Principal Component Analysis (PCA) of the 21 poems (threshold set to 0.30 to ignore any non-significant contributions) determines the variance explained through eigenvalues and identifies any significant factors, known as components, from within the data. Four components are then aggregated to examine the clusters.

Linear Discriminant Analysis (LDA) is used as an alternate classification technique to PCA [29-30]. The unknown works are removed, and all of the individual known authors' poems are numbered from 1 to 4 before training the model and reintroducing the unknown poems. Using the

resultant coefficients from the three Canonical Discriminant Functions, functions 1-2 and 1-3 are aggregated to visually compare the clusters.

The Vector Space Method (VSM) technique [42-43] is used with Elizabeth Cary's, *The Tragedy of Mariam, the Fair Queen of Jewry* as an imposter [44]. Pair-wise comparisons of each of the 21 *Passionate Pilgrim* poems is made against Elizabeth Carey's play (42 pair-wise comparisons) using both cosine and minmax similarity detection, to highlight the clusters that form based on their distance from Cary's play.

B. Seriation

According to [45] "Seriation is an exploratory combinatorial data analysis technique to reorder objects into a sequence along a one-dimensional continuum so that it best reveals regularity and patterning among the whole series." Seriation is the process of placing a linear ordering on a set of N multi-dimensional quantities. The total number of possible orderings is N! (factorial). This grows extremely quickly with N. 5! = 120, 10! = 3.6 million and 20! = $2.4x10^{18}$, or 2.4 billion billion (or quintillion). Thus, even for quite small N, it is not possible to calculate the shortest path by calculating all possible paths. A heuristic or approximation is needed. Inevitably any given approximation will work better with some data than others. Thus, for a robust estimation of the shortest path, it might be necessary to try a range of different estimators and look for consistency among them.

Using the free software environment for statistical computing and graphics, R, and its seriation package [46], and provide the seriation package with the 9x21 matrix consisting of the nine RPAS values for each of the 21 poems of *The Passionate Pilgrim*. Using the Euclidean distance option, seriation attempts to minimise the Hamiltonian path length (the Hamiltonian path on a graph is a path which visits all the nodes just once). The results of the six Hamiltonian path-length calculations produced by the seriation package are evaluated (TSP: *Travelling Salesperson*, Chen: *Rank two ellipse Seriation*, ARSA: *Anti-Robinson Simulated Annealing*, HC: *Hierarchical Clustering*, GW: *Hierarchical Clustering (Gruvaeus Wainer heuristic)*, and OLO: *Hierarchical Clustering (Optimal Leaf Ordering))*. While seriation gives a one-dimensional continuum, Dendrogram branch and leaf visualization are also provided, and clusters can be separated by their Hamiltonian path distances [47]. The technique that provides the shortest Hamiltonian path is selected, and noise introduced into the matrix to examine the strength of the connected groups by using the jitter function in R. The function adds random noise to the vector by drawing samples from the uniform distribution of the original data [48].

III. Analysis

Using **RPAS** Personal Pronouns (P) is plotted against Richness (R) (PtoR) for the 21 *The Passionate Pilgrim* poems (see Fig. 1). PtoR discriminates the unknown poems 14 and 16 with Shakespeare (poems 2 and 3), and they have a low feminine gendered style (P > 10), while all of Shakespeare's known poems have a lower feminine gendered style (P > 30), contrasting this is the group consisting of the cluster with unknown poems 7 and 19 that are similar in style to Griffin (poem 11) and Barnfield (poem 21) who all have a higher masculine style (P >50). The Shakespeare (poem 1) and the Marlowe and Walter Raleigh (poem 20) are similar, as are Barnfield (poem 8) and Shakespeare (poem 5). The unknown poem 12 (from Delaney) has a low Richness score is separate from the main body of poems.

A. Principal Component Analysis (PCA)

The findings show that many PCA correlations are in excess of 0.30. A visual indication of the correlation matrix highlights 24 coefficients are around 0.30 or higher and some are as high as 0.77, and Bartlett's test is significant (p = 0.001) meaning there is some correlation between variables indicating that PCA is worthwhile. Four components are extracted and account for 81.95% of the variance.

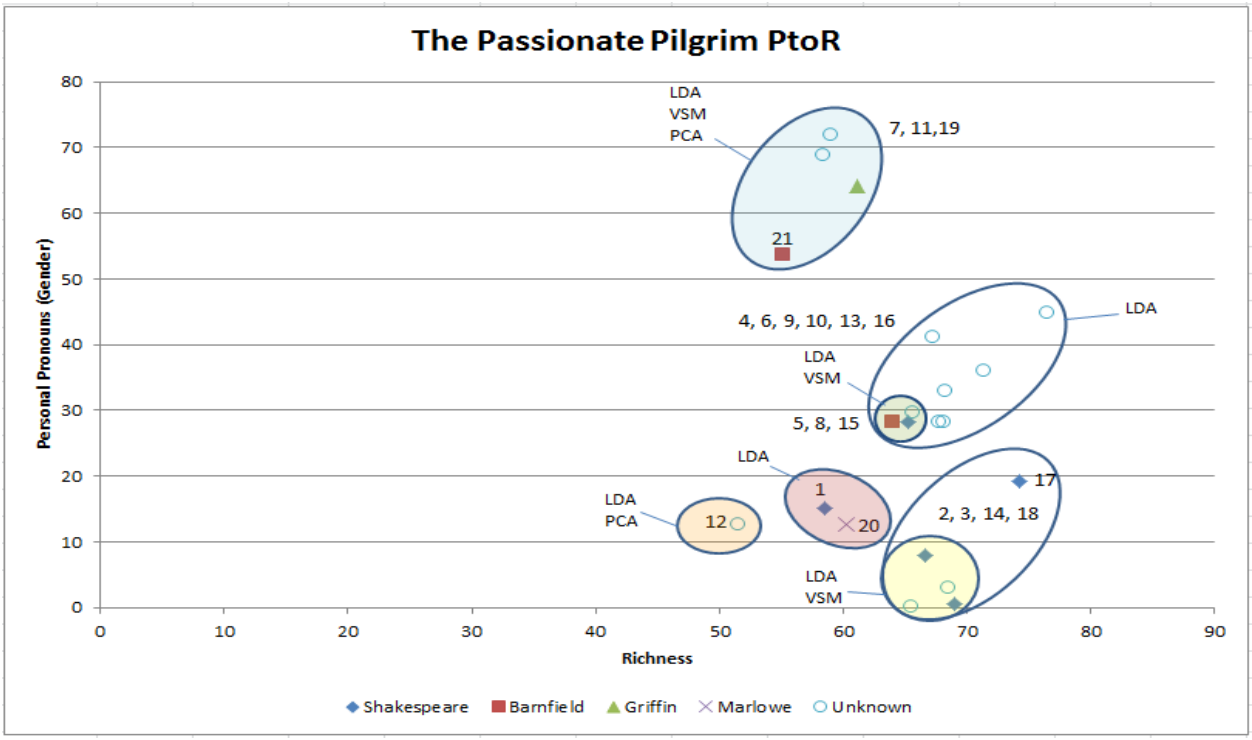


Fig. 1. In this The Passionate Pilgrim gendered Personal pronouns (P) versus Richness (R) diagram, the overlays of the results of LDA, VSM, and PCA analysis highlight the consistency of other results. A Barnfield / Griffin series of poems can be seen (7, 11, 19, and 21) with greater than 50% gendered personal pronouns. This is supported by LDA, VSM and PCA Analysis. A Shakespeare series of poems can be observed (2, 3, 14, 17, and 18), also supported by LDA and VSM analysis. A Shakespeare / Marlowe / Raleigh series is observed (1 and 20) to have less than 20% gendered personal pronouns supported by LDA analysis. Clearly, Delany's poem 12 is supported by LDA, and PCA analysis as a standalone work also has the lowest Richness. In the range of 25-50%, gendered personal pronouns are the Shakespeare / Barnfield poems (5, 8, and 15) supported by LDA and VSM analysis, and these alongside the unknown poems (4, 6, 9) (and 10, 13, 16 supported by LDA analysis). Further, the ellipses are a visual clustering assignment

In Fig. 1, the two common clusters are overlaid. A Barnfield / Griffin group (11 and 21) is found to sit with unknown poems 7 and 19. While unknown poem 12 (Thomas Delaney) was close to Shakespeare (1) and Marlowe and Raleigh (20), it is the furthest poem from the Shakespeare cluster on the Factor 1 and 2 scale that accounts for ~55% of the variance. Additionally, the results highlight all of the known Shakespeare poems cluster (poems 1, 2, 3, 5, 17 with 6, 14, 15, and 16). Poem 4 is close to Barnfield (8), and poems 6, 9, 15, and 16 are close to Shakespeare (5).

B. *Linear Discriminant Analysis (LDA)*

Three functions were extracted, and the first two accounted for 99.6% of the variance (1 = 95.9 and 2 = 3.7). The Wilks' Lambda test of functions 1 through 3 was significant (p=0.009) which highlights that the null hypothesis can be rejected and suggests that all three functions together have a discriminating ability. The second and third functions together are not significant (p=0.190), neither is function 3 on its own (p=0.453). Functions 1-2 and functions 1-3 are plotted to generate six common clustering results (see Fig. 1). It is found that the unknown poems 10 and 13 are again close to Shakespeare (5) and Barnfield (8), as is 15. Unknown poems 7 and 19 are closer to Griffin (11) this time and further from Barnfield (21). Unknown poem 12 (Thomas Delaney) is again closest to Shakespeare (1) and Marlowe and Raleigh (20) but stands alone. Poem 14 is again close to Shakespeare (2 and 3).

While poem 18 is also close to Shakespeare (1, 2, and 3), poem 4 is far from all the poems but closest to Griffin (11). Poem 6 is closest to Shakespeare (17). Poem 16 is closest to Shakespeare (5), and poem 9 is in the middle of Shakespeare (5), Barnfield (21) and Griffin (11). Again, there is some consistency with these results, but there seems to be a lack of clarity with poems 4, 6, 9 and 16.

C. *The Vector Space Method (VSM)*

Pair-wise comparisons of each of the 21 *Passionate Pilgrim* poems against Elizabeth Carey's play, *The Tragedy of Mariam, the Fair Queen of Jewry* (42 pair-wise comparisons) using both cosine and minmax similarity detection, highlights the clusters that form based on their distance from Cary's play. Fig. 1, indicates the three common clustering results. Here, unknown poems, 7 and 19 are in a cluster with Griffin (11). Unknown poem 14 is in a cluster with Shakespeare (1, 2, and 3) and Marlowe / Raleigh (20) and poems 12 and 18, and closest to Shakespeare (1), while Delaney's poem 12 and 14 are closest to Shakespeare (2), but furthest away. Unknown poems 4, 6, 9, 10, 13, 15, and 16 are in a cluster with Shakespeare (5 and 17) and Barnfield (8). In this cluster Barnfield (8) is very close to Shakespeare (5), and poems 10 and 13 have an almost identical score.

Throughout these different analysis techniques, there is a consistency in three to four clusters forming with common

poems in them, but many of the techniques have been dependent on an arbitrary visual clustering size. Therefore, to add further reliability to the results, the data is clustered using seriation to measure cluster distances.

D. Seriation

The R seriation package is fed a 9x21 matrix of the data, and using Euclidean distance seriation of the data minimizes the Hamiltonian path length. Results of the six seriation techniques available highlight that Hierarchical Clustering with Optimal Leaf Ordering (OLO) outperforms the Travelling Salesperson technique (path lengths 214.63 vs. 228.92). Incorporating the clustering of the OLO Dendrogram at a height of 25, the order of the 21 chunks with clusters highlighted is [21 19 7 11] [4 9 6] [5 8 10 13 15 16 17] [20 12 1 3 2 14 18] and it highlights some susceptibility between poems 11-4, 6-5, and 17-20. When the distances between each poem are compared, and either side of poems 11-4 (7-11-4-9), 6-5 (9-6-5-8), and 17-20 (16-17-20-12), the ordering sequence and distance information is important (refer Table 2).

TABLE II. HAMILTONIAN PATH DISTANCES BETWEEN THE 21 *THE PASSIONATE PILGRIM* POEMS. THE OLO DENDROGRAM EDGE CLUSTERS THAT FORM AT A DENDROGRAM HEIGHT OF 25 HIGHLIGHTS A CONSISTENCY IN TWO OF THE THREE SEPARATION POINTS. IN THE CLUSTER SPLIT AT POEMS 11-4, 7-11 AND 4-9 ARE CLOSER THAN 11-4 (27.3 VERSUS 11.8 AND 9.6). IN THE CLUSTER SPLIT AT POEMS 6-5, 9-6 AND 5-8 ARE CLOSER THAN 6-5 (10.61 VERSUS 7.7 AND 3.4), BUT IN THE 17-20 CLUSTER SPLIT, WHILE 16-17 AND 20-12 ARE CLOSER THAN 17-20, THE DIFFERENCES BETWEEN 16-17 AND 17-20 ARE MARGINAL (15.8 AND 12.6 VERSUS 16.8)

Poem edges		Path length
21	19	16.60488
19	7	24.69437
7	11	9.561261
11	4	27.27893
4	9	11.78111
9	6	7.683108
6	5	10.61323
5	8	3.444387
8	10	4.88489
10	13	3.22249
13	15	3.455063
15	16	4.449576
16	17	15.8412
17	20	16.75323
20	12	12.6397
12	1	14.13468
1	3	11.68744
3	2	8.28891
2	14	13.00578
14	18	6.162732

Further, when examining the OLO dendrogram edge clusters that form at a dendrogram height of 25 and find consistency in two of the three separation points. In the cluster split at poems 11-4, it can be seen that 7-11 and 4-9 are closer than 11-4 (27.3 versus 11.8 and 9.6). In the cluster split at poems 6-5, 9-6 and 5-8 are closer than 6-5 (10.61 versus 7.7 and 3.4), but in the 17-20 cluster split, while 16-17 and 20-12 are closer than 17-20, the differences between 16-17 and 17-20 are marginal (15.8 and 12.6 versus 16.8).

To see how stable the results are, in particular, the stability of the clusters connected at the poems 17-20 split, noise is inserted into the initial 9x21 RPAS-poem matrix and recalculate Euclidean distances with various amounts of noise (noise 1 – 8000). An examination of the scene chunk order after seriation (refer Table 3) highlights the high level of stability within the seriation and OLO clustering results. The different OLO seriation results are showing changes in order when noise is added to the RPAS poem matrix. At around noise levels of 500, poems 15 and 16 switch positions, but then revert back with further noise. At noise levels 800 and above, the Barnfield – Griffin cluster (7, 11, 19, and 21) move internally within the cluster but no poems leave. At noise levels 800 and higher the Shakespeare – Marlowe cluster (1, 2, 3, 12, 14, 18, and 20) move internally, and at no point does poem 20 moves out of the cluster and join with poem 17.

TABLE III. THE DIFFERENT OLO SERIATION RESULTS ARE SHOWING CHANGES IN ORDER WHEN NOISE IS ADDED TO THE RPAS POEM MATRIX. AT AROUND NOISE LEVELS OF 500, POEMS 15 AND 16 SWITCH POSITIONS, BUT THEN REVERT WITH FURTHER NOISE. AT NOISE LEVELS 800 AND ABOVE, THE BARNFIELD – GRIFFIN CLUSTER (7, 11, 19, AND 21) MOVE INTERNALLY WITHIN THE CLUSTER BUT NO POEMS LEAVE. AT NOISE LEVELS 800 AND HIGHER THE SHAKESPEARE – MARLOWE CLUSTER (1, 2, 3, 12, 14, 18, 20) MOVE INTERNALLY. THIS SUGGEST A HIGH LEVEL OF STABILITY IN THE SERIATION OLO ORDER AND OLO CLUSTERING RESULTS ([21 19 7 11] [4 9 6] [5 8 10 13 15 16 17] [20 12 1 3 2 14 18])

Noise								
Order	***0***	***100***	***500***	***800***	***1000***	***2000***	***4000***	***8000***
1	21	21	21	7	7	7	7	7
2	19	19	19	11	11	11	11	11
3	7	7	7	19	19	19	19	19
4	11	11	11	21	21	21	21	21
5	4	4	4	4	4	4	4	9
6	9	9	9	9	9	9	9	6
7	6	6	6	6	6	6	6	4
8	5	5	5	5	5	5	5	5
9	8	8	8	8	8	8	8	8
10	10	10	10	10	10	10	10	10
11	13	13	13	13	13	13	13	13
12	15	15	16	15	15	15	15	15
13	16	16	15	16	16	16	16	16
14	17	17	17	17	17	17	17	17
15	20	20	20	14	20	20	14	14
16	12	12	12	18	12	12	18	18
17	1	1	1	20	1	1	20	20
18	3	3	3	12	3	3	12	12
19	2	2	2	1	2	2	1	1
20	14	14	14	3	14	14	3	3
21	18	18	18	2	18	18	2	2

IV. DISCUSSION

Overall, the techniques were generally consistent, and seriation was useful because it was able to provide clustering and distance measures that appeared stable even with a relatively high level of introduced noise. Therefore, the basis of these findings lies in a rigorous multivariate approach to analysis and not a single technique. However, one of the biggest concerns is the influence of the publisher. While Jaggard or his associates cannot be discounted from having a hand in adding their own touches to some of these unknown poems, blending them as it were so they appear as part collaborations, it is an unknown factor. It is known that Jaggard was able to get hold of some of Shakespeare's unpublished work, and both he and his brother John had access to a wide number of Elizabethan works. What cannot be known is how much of this was early unpublished works.

Of the 12 anonymous poems, two are likely Shakespeare's, possibly from his earlier unpublished works (poems 14 and 18 are similar to Shakespeare's poems 2 and 3 and a lesser extent poem 1). However, if they were not earlier Shakespearian poems, then they are from another poet entirely, one that has not been examined. Two other poems (7 and 19) have a blended style similar to Griffin (11) and Barnfield (21), and there is more of Griffin's style (similar to poem 11) in them than Barnfield's, and they are more likely to be Griffin's unpublished work. Again, if they are not an unpublished Griffin poem, then they too are a poet that has not been examined in this paper. Poem (12) has a blended style similar to Shakespeare (1) and Marlowe / Raleigh (20) but consistently shows itself to be different enough to be an independent poet and be the work of Thomas Delaney whose other poems were outside of this analysis.

The remaining seven unknown poems (4, 6, 9, 10, 13, 15, and 16) are all similar in style to a blended Shakespeare (5 and 17) and Barnfield (8). All of these, as are all of Shakespeare's poems here, have a Richness score over 65%. They all have a Personal Pronoun score below 50%, which would be deemed as a feminine writing style which fits Shakespeare. Poems 4, 6, and 9 are very similar in style to each other and closer to Shakespeare's (5) style than Barnfield (8). Poems 10, and 13 are closer to Barnfield's (8) style than Shakespeare (5, 17). Poems 15 and 16 have a higher Shakespeare (5) style than Barnfield's (8) and are higher overall from the Shakespeare poems (5 and 17).

This close style of Barnfield's poem (8) to Shakespeare's (5) is an anomaly, and if it were not for the work sitting in the Shakespeare cluster between 5 and 17, then it could be easily be said that all the poems (4, 6, 9, 10, 13, 15, and 16) are Shakespeare's. The literature around Richard Barnfield is examined more closely. While Barnfield and Shakespeare were certainly friends [49] and could have collaborated, these poems are likely to be Shakespeare's because the style of Barnfield's poem (8) is very similar to Shakespeare's poem (5). It has been suggested, that the 1598 version of Barnfield's manuscript obtained by William Jaggard's brother John was of insufficient length (indicated by the sparse printing layout), and William Jaggard provided his brother two poems from the yet unpublished *The Passionate Pilgrim* to extend Barnfield's *Lady Pecunia* publication. In the 1605 reprint of Richard Barnfield's *Lady Pecunia*, the two poems from the 1598 first edition (poems 8 and 21 from *The Passionate Pilgrim*) were not included [50-51]. According to [52], Barnfield is said to have claimed authorship of only *one* of the two poems (stylistically likely poem 21). If this is true, then it explains the striking similarities between the Shakespeare and Barnfield poems (5 and 8), and a good indication that Shakespeare wrote both 5 and 8, and therefore poems 4, 6, 9, 10, 13, 15, and 16 are Shakespeare's poems. While it further reinforces Jaggard's approach to borrowing from other author's works, from the analysis it is believed that Shakespeare wrote nine of the twelve unknown poems (4, 6, 9, 10, 13, 14, 15, 16, and 18) including 1, 2, 3, 5, 17, and 8.

V. Conclusion

Given Shakespeare's signature in almost three-quarters of the poems, Jaggard may have adopted shrewd marketing tactics in using Shakespeare's name as the sole author. Indeed, when he expanded the third edition with a collection of nine of Heywood's poems, he did not remove Shakespeare's name from the title, nor did he add Heywood as co-author, but in his collection of assorted verses. Jaggard merely adopted what was a standard convention by publishers in the day [53]. The analysis would suggest that the five authors, Barnfield, Delaney, Griffin, Marlowe, and Raleigh were not acknowledged, and several poems may well be collaborative works between Shakespeare and others but this also was common [54]. It is also possible that several poems (7, 14, 18,19) are not early work or collaborations, but other writer's poems not studied here. This failing to acknowledge all author's poems would seem, at least by today's standards, to be an injustice. However, as it can be seen with Jaggard's publication of *The Passionate Pilgrim* and his later publication of Shakespeare's first folio, Jaggard focussed on promoting Shakespeare's work above all others.

In this paper, authors have demonstrated an alternate stylometric technique that can identify self and cluster multiple authors using RPAS. It includes the use of sensory-based adjectives and words that are strong in concreteness and imageability that reflect known psychological states in an individual's personality. They believe that further research is warranted to see if RPAS can identify changes in an individual's stylometric fingerprint over time.

Acknowledgment

The authors thank D. Crone and C. van Antwerpen for critical discussions and reading of the manuscript. This research supported by the Defence Science Technology Group, the Australian Government's lead agency dedicated to providing science and technology support for the country's defence and security needs.

References

[1] Erne, L. (2013). *Shakespeare and the book trade*. Cambridge University Press. Pp. 56-86.

[2] Devington, D. (2007) The Poems by William Shakespeare. Bantam Books, . New York.

[3] Woudhuysen, H. R. (1996). *Sir Philip Sidney and the circulation of manuscripts, 1558-1640*. Oxford University Press.

[4] Connor, F. X. (2014). Shakespeare, poetic collaboration and The Passionate Pilgrim. pp119-130, in Holland, P. (Ed.). (2014). *Shakespeare Survey: Volume 67, Shakespeare's Collaborative Work* (Vol. 67). Cambridge University Press.

[5] Chiljan, K. (2012). Reclaiming The Passionate Pilgrim for Shakespeare. *Oxfordian* , 2012, Vol. 14, p74-81

[6] Bednarz, J.P. (2007) "Canonizing Shakespeare: The Passionate Pilgrim, England's Helicon and the Question of Authenticity," Shakespeare Survey 60 (2007): 255-58,260,262.

[7] Elliott, W. E., & Valenza, R. J. (1991). A Touchstone for the Bard. *Computers and the Humanities*, *25*(4), 199-209.

[8] Korp, C. (2015). Shoemakers, Clowns, and Saints: The Narrative Afterlife of Thomas Deloney. Available at: http://escholarship.org/uc/item/8hk20311

[9] Segarra, S., Eisen, M., Egan, G., & Ribeiro, A. (2015). Stylometric analysis of early modern period English plays. *Digital Scholarship in the Humanities, vol.(submitted)*.

[10] Rudman, J. (2016). Non-Traditional Authorship Attribution Studies of William Shakespeare's Canon: Some Caveats. *Journal of Early Modern Studies*, *5*, 307-328.

[11] Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for information Science and Technology*, *60*(3), 538-556.

[12] Jackson, M. P. (2006). Shakespeare and the quarrel scene in arden of faversham. *Shakespeare Quarterly*, *57*(3), 249-293.

[13] Vickers, B. (2011). Shakespeare and authorship studies in the twenty-first century. *Shakespeare Quarterly*, *62*(1), 106-142.

[14] Hirsch, B. D., & Craig, H. (2014). " Mingled Yarn": The State of Computing in Shakespeare 2.0.

[15] Arefin, A. S., Vimieiro, R., Riveros, C., Craig, H., & Moscato, P. (2014). An information theoretic clustering approach for unveiling authorship affinities in Shakespearean era plays and poems. *PloS one*, *9*(10), e111445.

[16] Matthews, R. A., & Merriam, T. V. (1993). Neural computation in stylometry I: An application to the works of Shakespeare and Fletcher. *Literary and Linguistic Computing*, *8*(4), 203-209.

[17] Merriam, T. V., & Matthews, R. A. (1994). Neural computation in stylometry II: An application to the works of Shakespeare and Marlowe. *Literary and Linguistic Computing*, *9*(1), 1-6.

[18] Boyd, R. L., & Pennebaker, J. W. (2015). Did Shakespeare write Double Falsehood? Identifying individuals by creating psychological signatures with text analysis. *Psychological science*, 0956797614566658.

[19] Chung, C. K., & Pennebaker, J. W. (2008). Revealing dimensions of thinking in open-ended self-descriptions: An automated meaning extraction method for natural language. *Journal of research in personality*, *42*(1), 96-132.

[20] Argamon, S., Koppel, M., Pennebaker, J. W., & Schler, J. (2009). Automatically profiling the author of an anonymous text. Communications of the ACM, 52(2), 119-123.

[21] Iqbal, F., Binsalleeh, H., Fung, B., & Debbabi, M. (2013). A unified data mining solution for authorship analysis in anonymous textual communications. Information Sciences, 231, 98-112.

[22] Northoff, G., Heinzel, A., de Greck, M., Bermpohl, F., Dobrowolny, H., & Panksepp, J. (2006). Self-referential processing in our brain—a meta-analysis of imaging studies on the self. *Neuroimage*, *31*(1), 440-457.

[23] Rosenstein, M., Foltz, P. W., DeLisi, L. E., & Elvevåg, B. (2015). Language as a biomarker in those at high-risk for psychosis. *Schizophrenia research*.

[24] Zabelina, D. L., O'Leary, D., Pornpattananangkul, N., Nusslock, R., & Beeman, M. (2015). Creativity and sensory gating indexed by the P50: Selective versus leaky sensory gating in divergent thinkers and creative achievers. *Neuropsychologia*, *69*, 77-84.

[25] Kernot, D., Bossomaier, T., & Bradbury, R. (2017). Novel Text Analysis for Investigating Personality: Identifying the Dark Lady in Shakespeare's Sonnets, Journal of Quantitative Linguistics (Accepted 18 Jan, 2017).

[26] Tweedie, F. J., & Baayen, R. H. (1998). How variable may a constant be? Measures of lexical Richness in perspective. Computers and the Humanities, 32(5), 323-352

[27] Argamon, S., Koppel, M., Fine, J., Shimoni, A.R. (2003). Gender, genre, and Writing Style in Formal Written Texts. Text, Volume 23, Number 58, August 2003.

[28] Kernot, D. (2016) *Can Three Pronouns Discriminate Identity in Writing* in Data. In Sarker, R., Abbas, H., Dunstall, S., Kilby, P., Davis, R. Young, L. (eds) Data and Decision Sciences in Action: Proceedings of the Australian Society for Operations Research Conference 2016, Springer.

[29] Pennebaker, J. W. (2011). The secret life of pronouns. *New Scientist*, *211*(2828), 42-45.

[30] Pennebaker, J. W., Mehl, M. R., & Niederhoffer, K. G. (2003). Psychological aspects of natural language use: Our words, our selves. *Annual review of psychology*, *54*(1), 547-577.

[31] Bucci, W. (2002). The referential process, consciousness, and the sense of self. *Psychoanalytic Inquiry*, *22*(5), 766-793.

[32] Bucci, W., & Maskit, B. (2004). Building a weighted dictionary for referential activity. In *Spring Symposium of the American Association for Artificial Intelligence in Palo Alto, CA, March.*

[33] Kernot, D. The Identification of Authors using Cross Document Co-Referencing. The University of New South Wales. Nov 2013. Available at: http://www.unsworks.unsw.edu.au/primo_library/libweb/action/dlDisplay.do?vid=UNSWORKS&docId=unsworks_12072

[34] Lynott, D., & Connell, L. (2009). Modality exclusivity norms for 423 object properties. *Behavior Research Methods*, *41*(2), 558-564.

[35] Miller, G. A. (1995). *The science of words*. New York: Scientific American Library.

[36] van Dantzig, S., Cowell, R. A., Zeelenberg, R., & Pecher, D. (2011). A sharp image or a sharp knife: Norms for the modality-exclusivity of 774 concept-property items. *Behavior research methods*, *43*(1), 145-154

[37] Farrow, J. M. (1993) The Collected Works of Shakespeare. http://sydney.edu.au/engineering/it/~matty/Shakespeare/

[38] Toutanova, K., & Manning, C. D. (2000, October). Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics-Volume 13* (pp. 63-70). Association for Computational Linguistics.

[39] Mark, M. (2014) A Celebration of Women Writers. Available at: http://digital.libttrary.upenn.edu/women/cary/Mariam/Mariam.html Accessed 27 October 2014.

[40] Balakrishnama, S., & Ganapathiraju, A. (1998). Linear discriminant analysis-a brief tutorial. *Institute for Signal and information Processing*.

[41] Ye, J., Janardan, R., & Li, Q. (2004). Two-dimensional linear discriminant analysis. In *Advances in neural information processing systems* (pp. 1569-1576).

[42] Koppel, M., & Winter, Y. (2014). Determining if two documents are written by the same author. *Journal of the Association for Information Science and Technology*, *65*(1), 178-187.

[43] Voorhees, E. M. (1998). Using WordNet for text retrieval. *Fellbaum (Fellbaum, 1998)*, 285-303.

[44] Seidman, S. (2013). Authorship verification using the impostors method. In *CLEF 2013 Evaluation Labs and Workshop-Online Working Notes*.

[45] Liiv, I. (2010). Seriation and matrix reordering methods: An historical overview. *Statistical analysis and data mining*, *3*(2), 70-91.

[46] Buchta, C., Hornik, K., & Hahsler, M. (2008). Getting things in order: an introduction to the R package seriation. *Journal of Statistical Software*, *25*(3), 1-34.

[47] Earle, D., & Hurley, C. B. (2015). Advances in dendrogram seriation for application to visualization. *Journal of Computational and Graphical Statistics*, *24*(1), 1-25.

[48] Stahel, W., Maechler, M. (2011). 'Jitter' (Add Noise) to Numbers. R Documentation (1995 – 2011) available at: http://stat.ethz.ch/R-manual/R-devel/library/base/html/jitter.html. Accessed: 2 August 2016.

[49] Sauer, M. M. (2008). *The Facts on File Companion to British Poetry Before 1600*. Infobase Publishing.

[50] Barnfield, R. (1598). Lady Pecunia, Or, The Praise of Money: Also A Combat Betwixt Conscience and Covetousnesse ; Together with The Complaint of Poetry for the Death of Liberality. In Volume 1, Issue 7 of Illustrations of old English literature. pp 1-49. Digitized 25 Oct 2012. Available at: https://books.google.com.au/books?id=0J1TAAAAcAAJ. Accessed on: 11 Nov 2015.

[51] Barnfield, R. (1605). Lady Pecunia, Or, The Praise of Money: Also A Combat Betwixt Conscience and Covetousnesse ; Together with The Complaint of Poetry for the Death of Liberality. In Volume 1, Issue 4 of Illustrations of old English literature. pp 1-38. Digitized 25 Oct 2012. Available at: https://books.google.com.au/books?id=y51TAAAAcAAJ. Accessed on: 11 Nov 2015.

[52] Britannica, E. (2008). Richard Barnfield The Project Gutenberg EBook of Encyclopaedia Britannica, 11th edition, Volume 3, Part 1, Slice 3. Published 10 December, 2008. Page 415.

[53] Reid, L. A. (2012). "Certaine Amorous Sonnets, Betweene Venus and Adonis": fictive acts of writing in The Passionate Pilgrime of 1612. *Études Épistémè. Revue de littérature et de civilisation (XVIe–XVIIIe siècles.*

[54] Thomas, M. W. (2000). Eschewing credit: Heywood, Shakespeare, and plagiarism before copyright. *New Literary History*, *31*(2), 277-293.

ComplexCloudSim: Towards Understanding Complexity in QoS-Aware Cloud Scheduling

Huankai Chen
School of Computing
University of Kent
Canterbury, UK

Frank Z Wang
School of Computing
University of Kent
Canterbury, UK

***Abstract*—The cloud is generally assumed to be homogeneous in most of the research efforts related to cloud resource management and the performance of cloud resource can be determined as it is predictable. However, a plethora of complexities are associated with cloud resources in the real world: dynamicity, heterogeneity and uncertainty. For heterogeneous cloud resources experiencing vast dynamic changes in performance, a critical role is played by the statistical characteristics of execution times, related to different cloud resources, to facilitate decision making in management. The cloud's performance can be considerably influenced by the differences between the estimated and actual execution times, which may affect the robustness of resource management systems.**

Limitation exists in the study of cloud resource management systems' complexities even though extensive research has been done on complexity issues in various fields from decision making in economics to computational biology. This paper concentrates on managing the research question regarding the complexity's role in QoS-aware cloud resource management systems. We present the ComplexCloudSim. Here, CloudSim, a popular simulation tool-kit, is extended through modelling of complexity factors in the cloud, including dynamic changes of run-time performance, resource heterogeneity, and task execution times' uncertainty. The effects of complexity on performance within cloud environments are examined by comparing four widely used heuristic cloud scheduling algorithms, given that the execution time information is inaccurate. Furthermore, a damage spreading analysis, one amongst the available complex system analysis methods, is applied to the system and simulations are run to reveal the system's sensitivity to initial conditions within specific parameter regions. Finally, how small of a damage can spread throughout the system within the region is discussed as well as research is done for the potential ways to avoid such chaotic behaviours and develop more robust systems.

***Keywords*—Cloud Scheduling; Damage Spreading; QoS; Complexity; Chaotic Behaviour; Cloud Simulation**

I. Introduction

The widely popular pay-as-you-go service has been enabled by Cloud Computing [1], which provides access to a shared pool of physical/ virtual, dynamically heterogeneous and scalable computational resources. Computational resources of any scale can be used in a rented module as per need through commercial cloud providers such as Microsoft Azure, Amazon (AWS), Rackspace Open Cloud and Google Compute Engine, which has been made possible by Infrastructure-as-a-Service (IaaS) model of Cloud Computing. Since availing these services on-demand is convenient, over the last years, the use of Cloud Computing has grown exponentially.

Both industry and academia require tailored cloud applications (customised) to meet their demands and use cloud resources efficiently. The main question here is:

"How should map-reduce alike groups of tasks be scheduled in the complex cloud environment that is reliable and efficient while meeting the application requirement for QoS? [2]"

The Cloud Computing community has been facing a real challenge with the above question. The reported scientific advances in both software platform development and Cloud Computing that enable fast data processing in the cloud is certainly a good news. Successful deployment of analysis engines such as Hive, Dremel, MapReduce, Spar and Impala has helped to run analysis jobs in short time across thousands of cloud resources [11]. However, adaptively scheduling groups of tasks based on dynamic changes in resource performance has been a challenge and remains unsolved. Scheduling is a vital mechanism for many cloud analysis engines. Unfortunately, the performance of rented cloud resource is not familiar with the available cloud scheduling systems, as the characteristic is subject to change dynamically, making it difficult to quantify during run-time. Cloud resources are homogeneous and the performance of resources does not change as assumed by most of the current scheduling solutions. In real-world heterogeneous cloud environment, this results in poor performance.

Scheduling comes under NP-complete problem and its complexity increases significantly in a heterogeneous cloud environment [15]. In the simplest form, scheduling, by just allocating appropriate resources based on availability to the incoming tasks, can be performed in a blind way. Nevertheless, advanced and sophisticated schedulers are significantly more reliable and efficient. Moreover, in general, it is expected that schedulers would react to the cloud resource's dynamic performance, most probably by examining the current CPU load of resources [10]. Also, to deal with the massive scale of the cloud, schedulers have to be easily distributed, have low overhead and lightweight.

First, this paper presents an extension to CloudSim [9], i.e. ComplexCloudSim, which offers many capabilities to model the complexity associated with the heterogeneous cloud environment. Then, four heuristics cloud scheduling algorithms are compared by running simulation (as presented in Section III) to demonstrate how resource complexity can make the scheduling system less robust. Section IV presents a damage spreading analysis model for the complex cloud to reveal the

cloud's sensitivity to initial conditions in specific parameter regions. Thus, such hidden chaotic behaviour present in the cloud scheduling system as a result of complexity is discussed. Finally, Section V provides the conclusion for this paper.

II. ComplexCloudSim: Heterogeneity, Dynamicity and Uncertainty

The cloud can be used to share different types of resources, which are typically accessed through applications running in the cloud. A typical cloud scenario can be an application that can generate several jobs. This application may already have sub-tasks that need to be resolved. Each sub-task is sent to a resource for resolving by the scheduling system. In a simple scenario, adequate resources needed to execute the sub-tasks are decided by the user; however, in general, the application will require schedulers that can efficiently and automatically find the most appropriate resources for completing a group of tasks.

One of the most studied research topics in the optimisation community is the scheduling issue related with cloud computing [3] [4] [5]. However, the problem becomes more challenging due to several complexity factors such as:

- **Heterogeneity** : The versatility of the current cloud infrastructures is limited. A crucial feature that needs to be taken into consideration in any cloud system is heterogeneity. Now, a single physical host can run multiple virtual machines (VMs) simultaneously, spurred by the development of virtualisation technology. Nevertheless, virtualisation comes with new challenges that hamper resource scheduling in clouds. This is because of multiple VMs present in the system that share hardware resources (e.g. memory, CPU, network, I/O, etc.) of a physical machine. In such a scenario, accurate measurement of the rented VMs actual performance is difficult. For e.g., in Amazon EC2, instead of fixed performance measures, compute units determine the provisioning of resources to virtual machines. The level of computing power required for provisioning compute unit varies with different host machines, which effectuate heterogeneity amongst VM performance [6]. This suggests that cloud is never homogeneous but always heterogeneous in the real world.
- **Dynamicity** : Another important complexity factor inherent to cloud computing is dynamic changes in resource performance at runtime [7]. In the real world scenario, over- or under-provisioning of resource, hardware/software failures, application misbehaviours and resource CPU overload can lead to such dynamicity of resource performance. The amount of running jobs assigned may also affect the cloud resource and may exhibit local activity. This leads to the creation of complexity. Moreover, the complexity level related to resource dynamicity increased with sharing of common underlying hardware infrastructures with other VMs.
- **Uncertainty** : The availability of complete information about the state of cloud resources is assumed by most of the research efforts related to scheduling. However, in cloud computing, during provisioning, uncertainty can exist between the ready time and the computing capacity of a resource [8]. We argue that the main issues with cloud computing is such uncertainties that bring additional challenges in execution time prediction of tasks, which is vital for many scheduling algorithms. There can be drastic changes in resource states in cloud environments. In most cases, obtaining exact knowledge about a resource is almost impossible. Accurate estimation of runtime tasks, performing prediction correction, undertaking prediction fall-back, improving prediction by historical data, etc. are difficult to execute. Significant uncertainty may rise due to imprecise execution of prediction times in scheduling performance.

A. CloudSim

For Cloud Computing infrastructures, CloudSim is a popular framework to execute simulation of resource scheduling. Mentioning the main entities/concepts regarding CloudSim, in terms of terminology, is vital to introducing it:

- **Datacenter** includes a set of physical hosts that can either be heterogeneous or homogeneous based on hardware configurations (memory, CPUs, storage and bandwidth) and it acts as Cloud Provider. It facilitates resource provision to cloud users.
- **Host**, a physical machine, is defined through the amount of memory present, the list of CPUs (and their types), storage as well as allocated bandwidth. A host allows managing VMs based on a specified VM allocation policy.
- A Cloud Host component manages and hosts the Virtual Machine (**VM**).
- **Cloudlet** is a job assigned by the Cloud User to run on the cloud. A job can be defined by its resource requirement (the number of cores and the amount of memory needed for performing the job), length (millions of instructions), dependencies and type (MapReduce jobs include Map tasks and reduce tasks).
- A **Broker** is the mediator that negotiates between cloud providers and cloud users. it acts on behalf of the cloud user to identify suitable resources that can be obtained from the cloud provider. Broker undertakes online negotiations that are directed towards allocation of resources to meet QoS needs of the user application. Cloudlets are then sent by the broker for scheduling available resources under defined scheduling policies.
- **CloudletScheduler** allows determining the processing power shared amongst Cloudlets based on available resources. Different scheduling policies can be used for implementation of this scheduler.

The hosts and VMs computational capabilities are measured in terms of million instructions per second per core (MIPS) in CloudSim and most of its extensions [12] [14] [13]. Throughout the CloudSim simulation, this measurement plays a crucial role. Provisioned virtual machines are assumed to be stable and predictable, in terms of their performance,

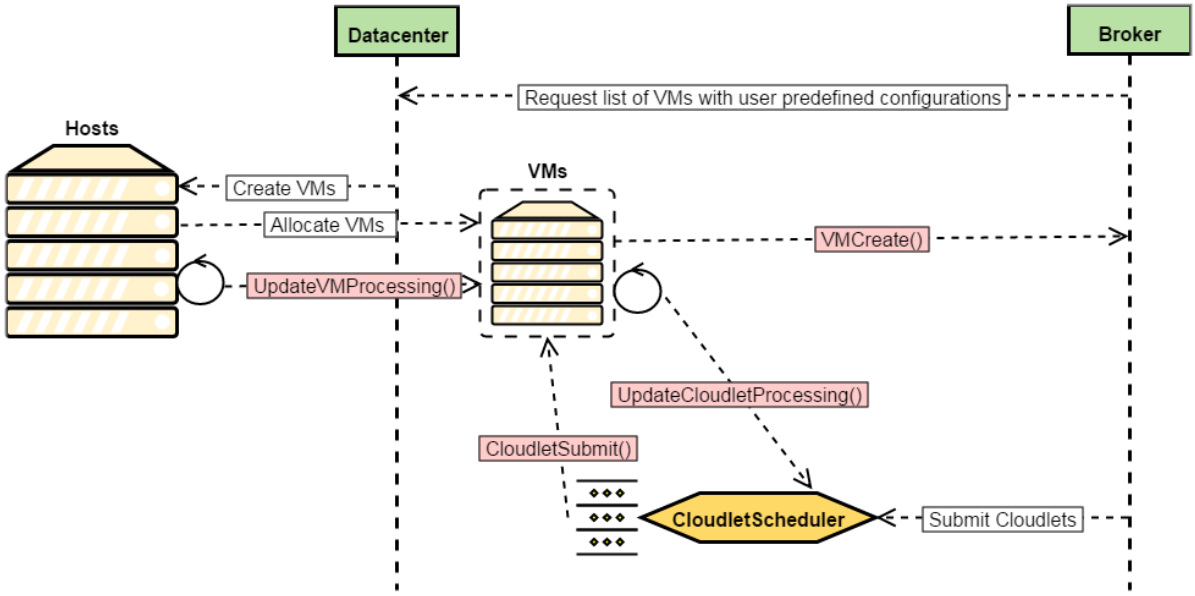


Fig. 1. CloudSim : Simulation Flow Chart

by CloudSim. Guaranteed performance is delivered by VMs, characterised as a fixed amount of MIPS. During a simulation, such a performance does not change, as presented in Figure 1. However, these assumptions do not hold when an actual cloud environment such as Amazon EC2 is used. Even though a certain core speed for each provisioned VM is guaranteed by most cloud providers, the runtime CPU utilisation of the host and the underlying physical hardware is assigned to determine the actual performance of a given VM. Thus, CloudSim may fail to efficiently simulate the cloud environment's complexity due to such incorrect assumptions.

B. ComplexCloudSim

This section explains how cloud simulations can be affected by complexity. This was derived based on a motivational example and a study employing four popular cloud scheduling algorithms. Then, the proposed ComplexCloudSim is presented by including cloud complexity in the original CloudSim.

1) Cloud Scheduling Algorithms: In general, in a cloud scheduler, we integrate a scheduling algorithm that runs on a permanent basis as follows: checking for available resources, receiving new incoming jobs, selecting appropriate resources based on performance (Estimated time to be completed) criteria and feasibility (jobs requirements to resources) as well as generating job plans (to make decision about job priorities and ordering) with selected resources.

Usually, Table I shows a list of terminologies used in relation to scheduling in clouds. For performance evaluation, this paper employs four popularly used heuristic scheduling algorithms related to simulations of cloud-based complexity. The followings are the definitions of these four heuristics.

- **FCFS**: Based on the sequence of submissions, tasks are executed. The task arriving first is prioritised for scheduling based on the available resource, just after submission, following which it is removed from the queue.

TABLE I. TERMINOLOGY FOR SCHEDULING IN CLOUD COMPUTING

Name	Description
QoS	Quality of the service
$MIPS$	Million instructions per second (CPU processing speed)
L_t	Length of task measured in million of instructions
ETC	Estimated time to compute
ERT	Estimated ready time of resource
MCT	Minimum completion time matrix
M_e	Estimated makespan
M_a	Actual makespan

- **Round Robin**: The first task is scheduled on the first resource, and then the second task on the second resource. This goes on through a cycling process for all the available resources.
- **MinMin** : Based on their length (of execution), all tasks in a job are first ordered. Scheduling is first done for the task having the shortest length for which the completion time will be minimum based on the available resource. Then, it is removed from the queue.
- **MaxMin** : Base on their length (of execution time), all tasks in a job are first ordered. Scheduling is done first for the task with the minimum length for which the completion time is maximum based on the available resource. Then, it is removed from the queue.

2) Motivational Example: This section shows how the robustness of a scheduler is affected by the complexity of resources. Let us consider a case in a homogeneous cloud with three VMs where ten independent jobs have to be scheduled (specifications are presented in Tables II and III). To make the complexity of scheduling simple, let us assume the jobs length is fixed and known and also consider that the clouds other performance related features will have no impact on the jobs actual completion timesuch as network bandwidth, memory consumption and disk I/O.

TABLE II. JOBS SPECIFICATIONS

Job Number	Number of Tasks	Task Length (MIs)
1	3	100
2	2	80
3	8	70
4	4	100
5	3	80
6	3	20
7	2	50
8	6	60
9	2	90
10	4	150

TABLE III. VMS SPECIFICATIONS

VMs	Core#	$MIPS_{request}$	$MIPS_{provision}$
VM1 (4 Cores)	1	10	9
	2	10	9
	3	10	9
	4	10	9
VM2 (4 Cores)	5	10	10
	6	10	10
	7	10	10
	8	10	10
VM3 (4 Cores)	9	10	11
	10	10	11
	11	10	11
	12	10	11
Total 3 VMs	12 Cores	120	120

In this example, the Min-Min heuristic is employed to schedule all of these independent jobs. Since this algorithm is efficient and simple, a better schedule (which minimises the jobs' total completion time) is produced when compared with other algorithms in the literature. Also, Algorithm 1 presents the Min-Min algorithm's pseudo code.

Algorithm 1 MinMin Scheduling algorithm

1: **Require:** A set of jobs with n tasks, m different cores, MCT matrix
2: **procedure** MINMIN SCHEDULING ALGORITHM
3: A list of jobs L_j in queue
4: A list of available cores L_c
5: **while** List L_j is no empty **do**
6: For each job in the list L_j
7: **if** The number of avaliable cores meets the job's requirement **then**
8: find the core that gives the minimum ETC
9: Update MCT matrix
10: From the MCT matrix, find the job with the minimum ETC
11: Remove the job from the job list L_j
12: Schedule the job's tasks to the match cores
13: Update the available cores list L_c

As we can see from the difference between the estimated scheduling plan in Figure 2 and the actual scheduling plan in Figure 3, the complexity of resources have a great impact on the job's QoS. In this simple example, the complexity factor of resources is shown to degrade the robustness of scheduling algorithms, i.e. the average job makespan and the

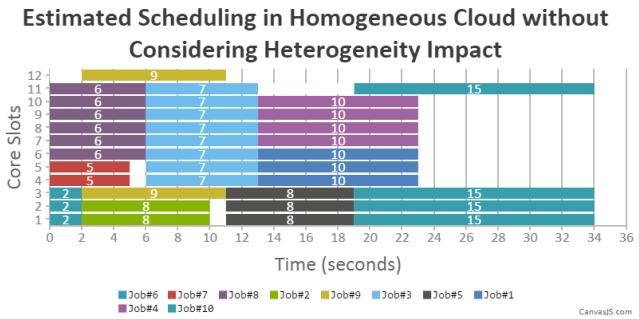


Fig. 2. Motivational Example : Estimated Scheduling Plan

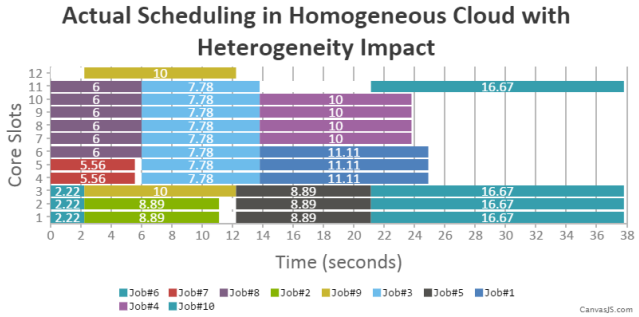


Fig. 3. Motivational Example : Actual Scheduling Plan

total workload runtime in this example, as shown on Table IV. Therefore, in the following section II-B1, we will investigate how different degrees of complexity impact such robustness and how different scheduling heuristics perform under the complex cloud environment.

TABLE IV. JOBS COMPLETION DETAILS

Job Number	M_e	M_a	Makespan Degradation
1	23s	24.89s	1.89s
2	10s	11.11s	1.11s
3	13s	13.78s	0.78s
4	23s	23.78s	0.78s
5	19s	21.11s	2.11s
6	2s	2.22s	0.22s
7	5s	5.56s	0.56s
8	6s	6s	0s
9	11s	12.22s	1.22s
10	34s	37.78s	3.78s (11%)

3) The Implementation for Introducing Complexity: As we have discussed at the beginning of this section, the performance of cloud scheduling is subject to different complexity factors relating to cloud resources: heterogeneity, dynamicity and uncertainty. In the remainder of this section, we will describe, in detail, how ComplexCloudSim attempts to capture these complexity factors.

*a) **Heterogeneity Ratio for VMs Provision**:* In a similar way to the situation with a real-world Cloud Provider, the performance of the provisioning VMs is not guaranteed in ComplexCloudSim. Hence, VMs of equal configuration are likely to have different core performances characterised by the random degradation of request MIPS during provision -

unlike the guaranteed fixed MIPS provision of CloudSim. In ComplexCloudSim, we allocate MIPS to the VMs when they are created, according to the *Heterogeneity Ratio*, as we can see from Algorithm 2.

Algorithm 2 Heterogeneity Ratio for VMs Provision

1: **Require:** VMs MIPS configuration, $MIPS_{request}$
2: **Require:** Heterogeneity Ratio, $0 \leq Ratio_{heterogeneity} \leq 1$
3: **procedure** VMCREATE($MIPS_{request}$,$Ratio_{heterogeneity}$)
4: **if** $Ratio_{heterogeneity} > 0$ **then**
5: $MIPS_{provision} = MIPS_{request} * (1 - Random \in [-Ratio_{heterogeneity}, Ratio_{heterogeneity}])$
6: **else** $MIPS_{provision} = MIPS_{request}$
7: VMProvision($MIPS_{provision}$)

b) ***Dynamicity Ratio for Changes of VM performance at Runtime****:* The idea that there are dynamic changes to performance at runtime, due to the sharing of common resources with other VMs and users, is also an important concept relating to the complexity inherent to Cloud scheduling. In CloudSim, the VM performance is kept to a fixed number of MIPS during simulation. In ComplexCloudSim, we periodically, every second, change the VM's runtime MIPS according to its Dynamicity Ratio and the host's current CPU utilization, as shown in Algorithm 3

Algorithm 3 Dynamicity Ratio for Changes of VM performance at Runtime

1: **Require:** Host's CPU Utilization, U_{host}
2: **Require:** Dynamicity Ratio, $0 \leq Ratio_{dynamicity} \leq 1$
3: **procedure** UPDATEMIPS(U_{host},$Ratio_{dynamicity}$) EVERY SECOND
4: **if** $Ratio_{dynamicity} > 0$ **then**
5: $MIPS_{runtime} = MIPS_{provision} * (1 - U_{host}) * (1 - Random \in [-Ratio_{dynamicity}, Ratio_{dynamicity}])$
6: **else** $MIPS_{runtime} = MIPS_{provision}$

c) ***Uncertainty Ratio for VM Performance Estimation with Inaccurate Information in Scheduling****:* Accurate resource performance prediction is hard or even impossible to achieve in actual complex cloud environments. CloudSim assumes that full information can be obtained and that such information is always correct for the purposes of performance prediction; this is not feasible in real world scenarios. Thus, we introduce a confidence level, the Uncertainty Ratio, to the resource performance predictions, which is used by several scheduling algorithms when making scheduling decisions (e.g. MinMin, MaxMin). In ComplexcloudSim, we inject the Uncertainty Ratio into all the processes which need to perform performance prediction, according to the algorithm 4.

Algorithm 4 Uncertainty Ratio for VM Performance Estimation with Inaccurate Information in Scheduling

1: **Require:** Estimated VM performance, $MIPS_{estimate}$
2: **Require:** Uncertainty Ratio, $0 \leq Ratio_{Uncertainty} \leq 1$
3: **procedure** PREDICTMIPS($MIPS_{estimate}$,$Ratio_{Uncertainty}$)
4: **if** $Ratio_{Uncertainty} > 0$ **then**
5: $MIPS_{actual} = MIPS_{estimate} * (1 - Random \in [-Ratio_{Uncertainty}, Ratio_{Uncertainty}])$
6: **else** $MIPS_{actual} = MIPS_{estimate}$

III. Complexity Simulation: Comparison of Four Heuristics Cloud Scheduling Algorithms

To showcase a possible application of Complexcloudsim, we simulated the execution of a computationally intensive workload (The Montage workflow) using four different heuristic cloud scheduling algorithms and various degrees of complexity in the Cloud resources. We expected the schedulers to differ in their robustness in relation to complexity, and that this should be reflected in diverging workflow execution times. In this section, we outline the experimental setup and evaluate the impacts of resource complexity on Cloud scheduling systems.

A. Experiment Setup

Simulation of the scheduling system was done to examine the robustness of the degradation created by resource complexity. A Montage workflow was employed for this experiment, which comes with CloudSim. This included 1,000 jobs containing a group of random number sub-tasks. To maintain simplicity, we employ a global variable, a degree of complexity, which allows configuring the ratios of dynamicity, heterogeneity and uncertainty simultaneously. For each configuration, the execution of Montage workflow was repeated 100 times on five VMs, after which the statistical results were generated in terms of workflow runtimes. During the course of the experiments, the degree of complexity caused by ComplexCloudSim was incrementally increased, and the impacts of complexity on cloud scheduling systems QoS performance was measured. To compare ComplexCloudSim with the original CloudSim, a baseline simulation we conducted that ran without considering complexity factors; this was also executed 100 times. As expected, under four scheduling algorithms, we determined the workflow runtime for the same workflow with zero variance maintained in the original CloudSim, as presented in Table V.

TABLE V. BASELINE SIMULATION RESULT WITH ORIGINAL CLOUDSIM

Scheduling Algorithms	FCFS	RR	MinMin	MaxMin
Average Runtime (Minutes)	2862	2865	2864	2862
Variance	0	0	0	0
Standard Deviation	0	0	0	0

B. Experiment Result

Here, the impacts on robustness were compared by employing different degrees of resource complexity and scheduling algorithms. Figures 4 and 5 outlined above present the experiment's results. Figure 4 presents the average runtime of the Montage work-flow between 3,220 and 3,505 minutes for all experiments. This indicates degradation of runtime by around 1323% compared with the performance baseline. Apparently, ComplexCloudSim offers complexity factors that have a considerable impact on the cloud scheduling system's QoS.

The average runtime degradation was also found not to change directly in tandem with increased degree of complexity. However, as observed in Figure 5, the degree of complexity ranging from 20% to 120% was found to be proportional with

the increase in the standard deviation for workflow runtime. It was clear that less reliable scheduling performances were obtained due to increase in the standard deviation. Thus, the complexity of the resources determines the reliability of the cloud scheduling system.

Based on the experimental results, the complexity factor had minimum impact on the MinMin scheduling algorithm in terms of both average and standard deviation of the workflow runtime. This suggests that more robust schedules are generated due to MinMin in a complex cloud environment. So when compared with other three heuristics, the overall performance of MinMin was found to be better, which was in line with the earlier research.

Evidently, the effect of complex resources can be simulated by ComplexCloudSim. This is a very desirable property as cloud environments always keep facing complexity issues. We expect this to be important going forward as other cloud simulators did not sufficiently support it.

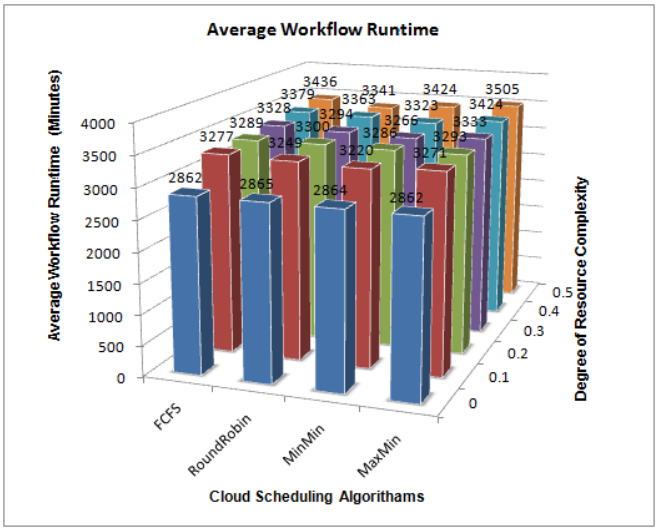


Fig. 4. Complexity Simulation: Average Workflow Runtime

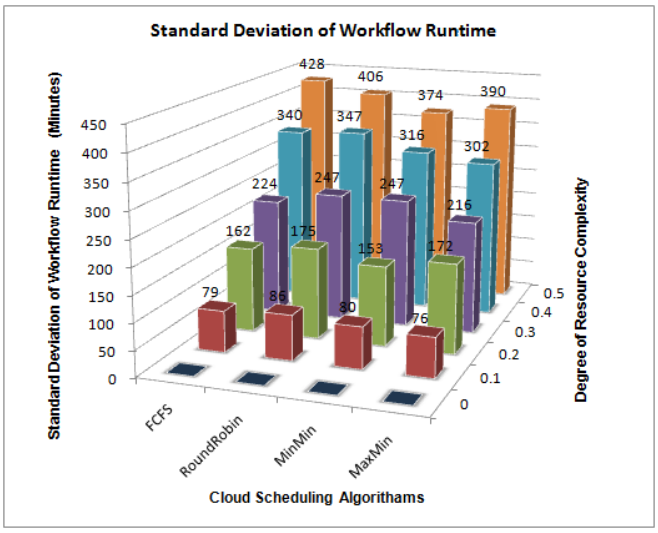


Fig. 5. Complexity Simulation: Standard Deviation of Workflow Runtime

IV. Damage Spreading Evaluation: Chaotic Behaviour in Cloud Scheduling

The original development of the tool Damage Spreading [17] was aimed at studying biologically motivated complex systems. This tool has been commonly referred in the literature for several research areas, including complex network models, for observing systems complex behaviour. In complex systems, the evolution of slightly different configurations of variables can be investigated with this tool, provided they are subjected to the same number sequence. Obtaining information regarding whether or not a small perturbation (damage to the conditions) introduced amongst variables can stay or spread at the same level (even disappears) would assist us in examining a systems robustness in relation to disturbance [16].

Here, "initial damage" is the occurrence of a slight change in the number of VMs C_{vm} and the degree of resource complexity $C_{complexity}$ to run the same workload. We introduced small changes $C_{vm} = 1$ and $C_{complexity} = 0.1$ to a simulation step-wisethe simulation that was executed 100 times with the same workload. Then, we examined if the changes would spread or not by taking into account two important QoS determinants in the scheduling processes - the standard and average deviation of workflow runtime.

To assess the damage spread, the damage was defined as D_{std} (difference in workflow runtime standard deviation R_{std}) and $D_{average}$ (difference in average workflow runtime $R_{average}$) present between two simulation results. As shown through Formulas 1 and 2, these were then calculated, where $j \in [0.1, 0.2, 0.3, 0.4, 0.5]$ represents the degree of complexity and $i \in [5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15]$ represents the number of VMs.

$$D_{average}(i,j) = R_{average}(i + C_{vm}, j) - R_{average}(i,j) \quad (1)$$

$$D_{std}(i,j) = R_{std}(i, j + C_{complexity}) - R_{std}(i,j) \quad (2)$$

Figures 6 and 7 show the results of $D_{average}$ and D_{std} respectively.

As observed in Figure 6, for number of VMs $i < 10$ and various degrees of complexity, the changes of $D_{average}$ are relatively small. The damage does not spread in this region and stays low at initial level.

As seen in Figure 7, the changes of D_{std} for various degrees of complexity, for number of VMs $i < 9$, become highly unstable. However, the situation becomes considerably better with an increase in the number of VMs, when $i > 9$.

Then, the relation between the spreading damage and the number of increased VMs i is examined by employing the standard deviation of $D_{average}$ and D_{std}. The standard deviation of $D_{average}(i)$ as $\sigma_{average}(i)$, and the standard deviation of $D_{std}(i)$ as $\sigma_{std}(i)$ are defined. Hence, the mean values: $Mean(\sigma_{std})$) and $Mean(\sigma_{average})$ of all $\sigma_{average}$ and σ_{std} are calculated, as presented in Tables VI and VII.

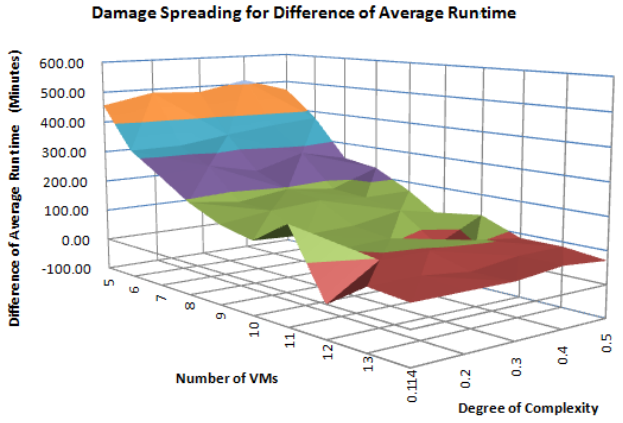


Fig. 6. Damage Spreading Evaluation: $D_{average}$

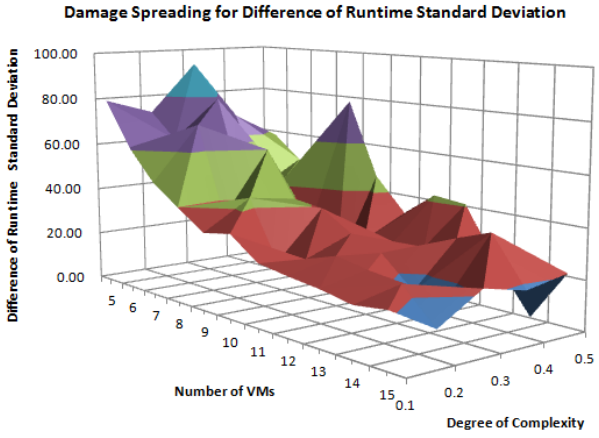


Fig. 7. Damage Spreading Evaluation: : D_{std}

Now, the state of the region is categorised loosely by employing such mean values. We now assign region $\sigma_{average}(i) \leq Mean(\sigma_{average})$ or $\sigma_{std} \leq Mean(\sigma_{std})$ as "Stable Regions". In this region, spreading and initial damages maintain a stable correlation. Reliable improvements in QoS occur with increased number of VMs, which signify smooth and robust running of the scheduling system against degree of complexity changes. We also categorise $\sigma_{average}(i) > Mean(\sigma_{average})$ or $\sigma_{std} > Mean(\sigma_{std})$ as "Chaotic Regions" [18], as highlighted in Tables VI and VII by the red colour. In this region, throughout the scheduling system, small disturbances may spread, which result in significant changes in performance due to the degree of complexity experienced. This suggests that it is difficult to guarantee QoS to an increase in the number of VMs.

Knowing when the scheduling system is in a chaotic region or stable region helps in providing important guidelines to quickly make decisions regarding achieving of a more robust scheduling. For e.g. in a real world situation, we might run a similar workload with more than 9 VMs based on the results from simulation of ComplexCloudSim, but we could also avoid choosing 11 or 12 VMs in a bid to satisfy the requirement of QoS.

TABLE VI. RELATION BETWEEN NUMBER OF VMS AND $D_{average}$

	$D_{average}(i)$					
	Degree of Complexity					Mean($\sigma_{average}$)=23
(i) VMs	0.1	0.2	0.3	0.4	0.5	$\sigma_{average}(i)$
5	456	489	481	514	469	22
6	320	322	344	363	377	25
7	258	271	237	282	248	18
8	193	174	196	178	231	23
9	148	168	180	169	171	12
10	124	117	122	149	94	19
11	198	101	108	64	135	50
12	-1	96	98	104	86	44
13	80	81	65	83	86	8
14	69	68	67	83	71	7

TABLE VII. RELATION BETWEEN NUMBER OF VMS AND D_{std}

	$D_{std}(i)$					
	Degree of Complexity					Mean(σ_{std})=24
(i) VMs	0.1	0.2	0.3	0.4	0.5	$\sigma_{std}(i)$
5	58	69	94	73	80	49
6	48	37	79	63	61	38
7	42	43	39	71	48	31
8	78	23	60	34	40	30
9	46	9	41	44	32	21
10	32	23	39	20	34	18
11	42	25	31	24	26	18
12	41	26	26	28	24	17
13	19	32	15	26	22	13
14	0	37	15	24	20	11
14	21	18	22	11	22	11

V. CONCLUSION AND FUTURE WORK

This paper presents an extension to the CloudSim, which is ComplexCloudSim, to analyse scheduling under a complex cloud environment. The design of a resource complexity module (dynamicity, heterogeneity, and uncertainty) is based on implementation with the primary goal to offer a useful tool for testing and validating the cloud scheduling algorithms robustness. Section III presents the examination results of four cloud scheduling algorithms to showcase the capability of ComplexCloudSim to simulate different complexity factors for the cloud scheduling system as well as replicate the shortcoming and known strengths of these algorithms. Then, based on simulation in Section IV, we found two regions: Stable Region, the region with converged small damage and Chaotic Region, the region where damage spread, in the complex cloud scheduling system.

We find Chaotic Behaviour in the cloud scheduling system to be interesting because it signifies that the future schedules in principle cannot be predicted. Such findings may explain why in the real-world production environment, it is difficult to put most of the scheduling algorithms in research, which rely on prediction and the complexity exists everywhere. Even if we know the precise processing time in advance, it does not guarantee the precise completion time of tasks for complex product systems such as the cloud. Therefore, if the scheduling system decides to plan for a more robust production schedule, it has to first predict if it is in Chaotic Region or Stable Region. Then, suppose the system is under the Chaotic Region, it has

to look out for VMs with a suitable number to meet the QoS requirement of the application.

Even through the ComplexCloudSim can model complexity factors to an extent, still it cannot cover all the situations occurring in the real-world cloud. However, the findings related to chaotic behaviour in cloud scheduling system have inspired new ideas to develop a more robust QoS-aware scheduling algorithm. More detailed analysis is required in further work to understand the cloud scheduling systems chaotic behaviour as well as the damage spreading mechanisms. Such chaotic behaviour needs to be studied for applying in real-world applications. We deem our research work to be one of the many steps towards multiple fruitful research topics.

REFERENCES

[1] Mell, Peter, and Tim Grance. "The NIST definition of cloud computing." (2011): 20-23.

[2] Plestys, Rimantas, et al. "The measurement of grid QoS parameters." Information Technology Interfaces, 2007. ITI 2007. 29th International Conference on. IEEE, 2007.

[3] Braun, Tracy D., et al. "A comparison of eleven static heuristics for mapping a class of independent tasks onto heterogeneous distributed computing systems." Journal of Parallel and Distributed computing 61.6 (2001): 810-837.

[4] Gutierrez-Garcia, J. Octavio, and Kwang Mong Sim. "A family of heuristics for agent-based elastic cloud bag-of-tasks concurrent scheduling." Future Generation Computer Systems 29.7 (2013): 1682-1699.

[5] Bala, Anju, and Inderveer Chana. "A survey of various workflow scheduling algorithms in cloud environment." 2nd National Conference on Information and Communication Technology (NCICT). 2011.

[6] Iosup, Alexandru, Nezih Yigitbasi, and Dick Epema. "On the performance variability of production cloud services." Cluster, Cloud and Grid Computing (CCGrid), 2011 11th IEEE/ACM International Symposium on. IEEE, 2011.

[7] Schad, Jrg, Jens Dittrich, and Jorge-Arnulfo Quian-Ruiz. "Runtime measurements in the cloud: observing, analyzing, and reducing variance." Proceedings of the VLDB Endowment 3.1-2 (2010): 460-471.

[8] Herroelen, Willy, and Roel Leus. "Project scheduling under uncertainty: Survey and research potentials." European journal of operational research 165.2 (2005): 289-306.

[9] Calheiros, Rodrigo N., et al. "CloudSim: a toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms." Software: Practice and Experience 41.1 (2011): 23-50.

[10] Chen, Huankai, et al. "Complexity Reduction: Local Activity Ranking By Resource Entropy For QoS-aware Cloud Scheduling." Services Computing (SCC), 2016 IEEE International Conference on. IEEE, 2016.

[11] Chen, Huankai, and Frank Z. Wang. "Spark on entropy: A reliable & efficient scheduler for low-latency parallel jobs in heterogeneous cloud." Local Computer Networks Conference Workshops (LCN Workshops), 2015 IEEE 40th. IEEE, 2015.

[12] Chen, Weiwei, and Ewa Deelman. "Workflowsim: A toolkit for simulating scientific workflows in distributed environments." E-Science (e-Science), 2012 IEEE 8th International Conference on. IEEE, 2012.

[13] Garg, Saurabh Kumar, and Rajkumar Buyya. "Networkcloudsim: Modelling parallel applications in cloud simulations." Utility and Cloud Computing (UCC), 2011 Fourth IEEE International Conference on. IEEE, 2011.

[14] Bux, Marc, and Ulf Leser. "Dynamiccloudsim: Simulating heterogeneity in computational clouds." Future Generation Computer Systems 46 (2015): 85-99.

[15] Chen, Huankai, Frank Wang, and Na Helian. "A Cost-Efficient and Reliable Resource Allocation Model Based on Cellular Automaton Entropy for Cloud Project Scheduling." system 4.4 (2013).

[16] Ikeda, Hinata. "Chaotic behavior in complex shop scheduling." Soft Computing and Intelligent Systems (SCIS) and 13th International Symposium on Advanced Intelligent Systems (ISIS), 2012 Joint 6th International Conference on. IEEE, 2012.

[17] Kauffman, Stuart A. "Metabolic stability and epigenesis in randomly constructed genetic nets." Journal of theoretical biology 22.3 (1969): 437-467.

[18] Boccaletti, Stefano, et al. "The control of chaos: theory and applications." Physics reports 329.3 (2000): 103-197.

A Novel Design of Patch Antenna using U-Slot and Defected Ground Structure

Saad Hassan Kiani[1], Khalid Mahmood[2], Mehre Munir[3], Alex James Cole[4]
Member IEEE [1, 2, 3]
School of Engineering and Digital Arts, University of Kent, Kent, United Kingdom.[4]

***Abstract*—A novel design of patch antenna is presented with double U slot structure on patch with ground irregularities. As a result tri-band response is achieved with gain reaching 0.785 to 3.75dB respectively and directivity of 5.5 to 5.6dBi. Coaxial cable is mounted with patch as medium of power. The antenna has shown minimum mismatch loss with 0 to 5% with high bandwidth response of 37 to 1200MHZ. The proposed antenna can be used for GSM, W-LAN, GPRS and other radio communication services systems.**

***Keywords*—multiband frequencies; directivity; gain; slots; Bandwidth; reflection coefficient**

I. Introduction

With rapid advancement in communication technology in modern era, antenna designers and researchers have been focused to new designs and structures. Patch antenna due to its low profile structures has been a prominent point of attention to communication technology. With their use as an array, resulting in powerful signals gain, bandwidth and directivity they have been frequently used for far in space communication and satellite systems. Regarding to reduction of size in patch structures and multiband response, several techniques have been proposed and implemented. Some of them are discussed.

With use of CSRR as defection of patch, size reduction was up to only 17% in [1] with altered radiation patterns. Use of synthetic magnetic conductors resulted with lowered gain at desired resonant frequencies [2].Using artificial magnetic conductors, nearly 40% of miniaturization is achieved at expense of efficiency [3]. Introducing incensement in electrical permittivity of a substrate, size reduction can be achieved but at the bandwidth gets worse [4]. Stack configuration with pi-shape fractal patch structure in [5] resulted size reduction with nearly diminishing gain. Hence Metamaterials, Stacked configuration, slotting techniques [6] [7] [8] [9] [10] size reduction and multiband response has been seen as very common topic among designers and researchers. In this paper, a novel design of patch antenna is presented. Two double U slots on patch and with help of ground irregularities, size reduction is achieved with multiband response which will be discussed below.

The antenna is designed in Computer Simulation Technology 2014. This paper is organized as following.

Starting from introduction, follows antenna design terminologies and slot method for size reduction. After result and discussion, Conclusion and Future work to be done is presented.

II. Antenna Design

The fundamental structure of patch antenna is presented in figure 1 which consists of ground, substrate and radiating patch.

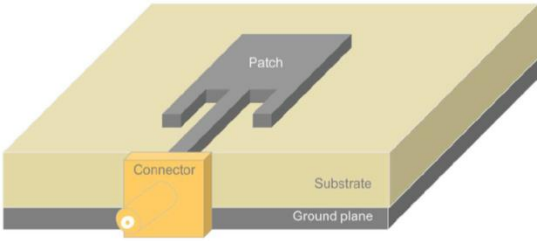


Fig. 1. Fundamental Patch Design

A. Substrate

Substrate plays a very important role in antenna performance parameters. Due to its moisture handling capabilities and commercially availability, FR4 with relative permittivity of 4.3 is selected [11].

B. Width

Antenna width is calculated using equation 1 [11] [12].

$$W = \frac{c}{2 f_0 \sqrt{\frac{(\varepsilon r+1)}{2}}} \quad (1)$$

Where c is the speed of light in free space.f_0 is the resonant frequency and $\boldsymbol{\varepsilon r}$ is the relative substrate permittivity.

C. Length

Patch length is calculated using equation 2 [11] [12].

$$L = L(eff) - 2\Delta L \quad (2)$$

Where

$$L(eff) = \frac{c}{2 f_0 \sqrt{\varepsilon_{(reff)}}} \quad (3)$$

And

$$\varepsilon_{(reff)} = \frac{\varepsilon r+1}{2} + \frac{\varepsilon r-1}{4}\left(1 + \frac{12h}{W}\right)^{-1/2} \quad (4)$$

Where h is the height and W is the width of the patch. After deriving all the basic design parameters, patch antenna resonation at 4.5GHz is designed. The height of the patch and

ground is kept 0.787mm respectively. Substrate thickness is kept 1.6mm.

The table 1 below shows dimension of conventional patch antenna.

TABLE I. DIMENSIONS OF PATCH ANTENNA

Parameters	Values in MM
Patch Length	15.35
Patch Width	20.48
Patch height	0.787
Ground Length	27.35
Ground Width	32.48
Ground Height	0.787

The table 2 below shows of the U slots introduced in patch and ground plane.

TABLE II. DIMENSIONS OF U SLOTS IN ANTENNA

Parameters	Value in MM
Patch upper U slot length and width	6.0
U slot Thickness	1.0
Patch lower U slot length and width	12.0
Patch lower U slot width	6.0
Ground U slot length and width	24

The antenna after introducing slots in its structure showed multiband response with good gain, directivity and radiation patterns. The overall antenna design is shown in figure 2. The (a) part represents patch and (b) represents ground plane.

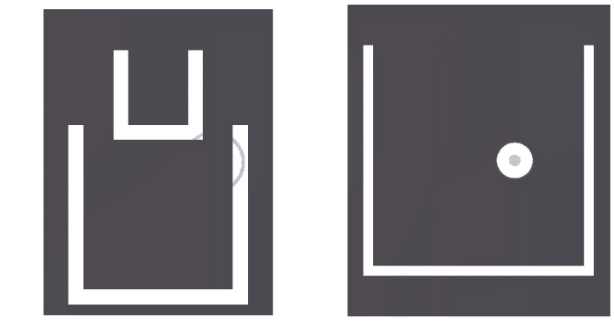


Fig. 2. (a) Frontal view of patch (b) Frontal view of Ground plane

We have used SMA connector for feeding antenna configuration as it is simply executed to 50 ohms of input resistance at required place in patch.

III. RESULTS AND DISCUSSIONS

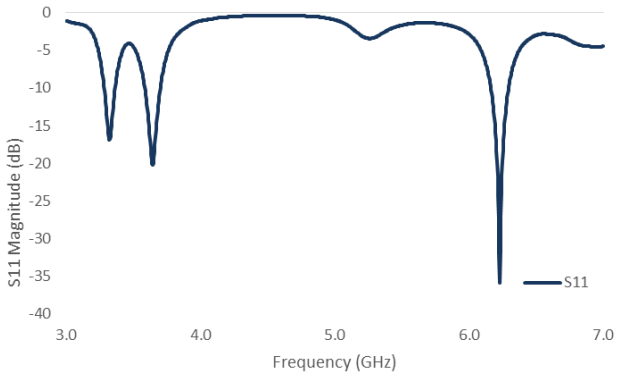


Fig. 3. Return loss graph of antenna

Figure 3 shows the return loss graph of proposed antenna. Resonation at 3 different frequencies, our proposed structure shows multiband response. The results are shown in table 3.

TABLE III. RETURN LOSS VALUES

Frequencies (GHz)	Return Loss
3.32	-19.50dB
3.64	-19.40dB
6.23	-35.79dB

As compared to the conventional antenna for 4.5GHz, the fundamental frequency of our proposed structure shifts down to 3.23GHz. Now the conventional antenna for 3.32 GHz of frequency would require dimensions of 22.18 x28.80 =638.74mm^2 while in our case it's just 15.35x 20.48 =314.368mm^2. So it shows that size is reduced in our proposed design up to 51% as our design is operating at 49% of conventional patch size with multiband response.

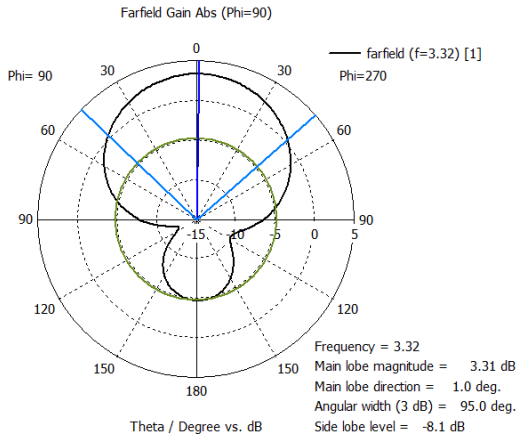


Fig. 4. 3.32GHz 1D Gain radiation pattern

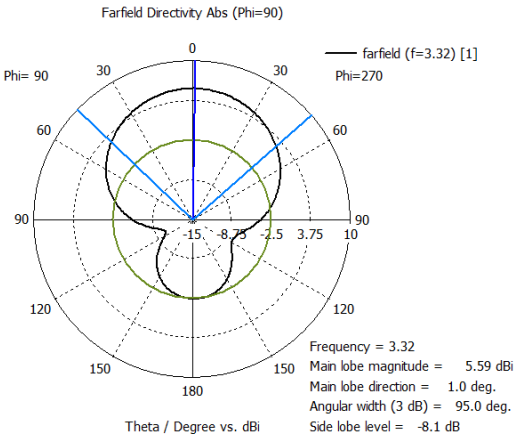


Fig. 7. 3.32GHz 1D radiation pattern of directivity

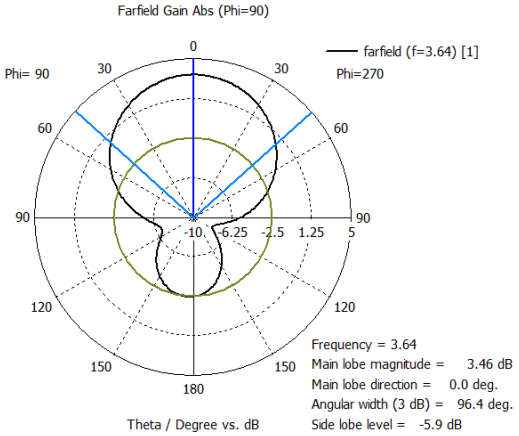


Fig. 5. 3.64GHz 1D gain radiation pattern

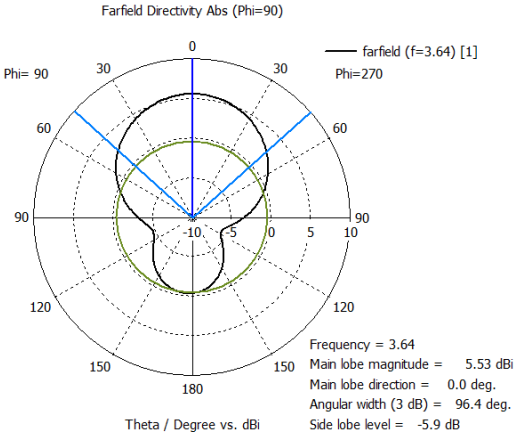


Fig. 8. 3.64GHz 1D radiation pattern of directivity

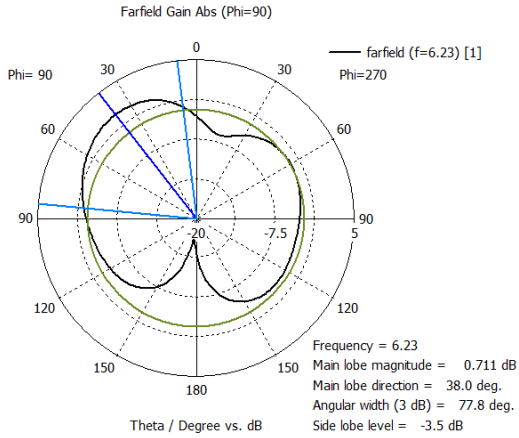


Fig. 6. 6.23GHz 1D gain radiation pattern

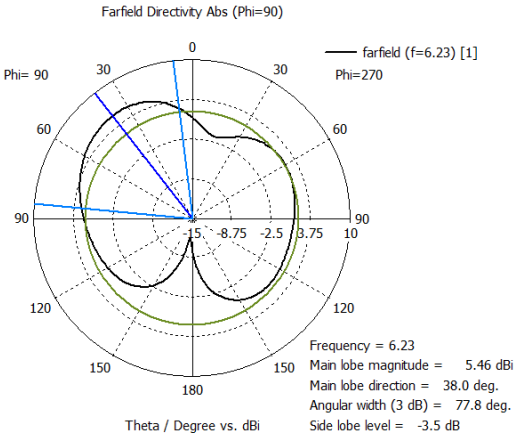


Fig. 9. 6.23GHz 1D radiation pattern of directivity

The radiation patterns of gain are presented in above and directivity are presented in below figures. In all resonating frequencies, there is minimum back lobe radiations better than [10]. The front lobe, back lobe, angular width and side lobe levels are mentioned in table 4 and 5.

TABLE IV. GAIN RADIATION PATTERN PARAMETERS

Frequency (GHz)	Main lobe Direction (Degrees)	Angular Width	Side Lobe Level (dB)
3.32	1.0	95.0	-8.1
3.64	0.0	96.4	-5.9
6.23	38.0	77.8	-3.5

TABLE V. DIRECTIVITY RADIATION PATTERN PARAMETERS

Frequency (GHz)	Main lobe Direction (Degrees)	Angular Width	Side Lobe Level (dB)
3.32	1.0	95.0	-8.1
3.64	0.0	96.4	-5.9
6.23	38.0	77.0	-3.5

From the results mentioned in table 4 and 5 it is cleared that radiation patterns obtained are far better than [9] [10] as antenna is conducting straight form origins. The VSWR ration also tends to be very satisfying in between range of 1 to 1.5 and ensuring no mismatch losses. The gain, directivity, bandwidth and VSWR are presented in table 6.

TABLE VI. ANTENNA PERFORMANCE PARAMETERS

Frequency (GHz)	Gain (dB)	Directivity (dBi)	VSWR	Bandwidth (MHz)
3.32	3.47	5.5	1.35	112.55
3.64	3.50	5.6	1.22	75.000
6.23	0.75	5.5	1.03	37.244

IV. CONCLUSION

A novel combination of U slot patch is presented. Our design showed multiband response with size reduction of 51% also the gain and directivity and other performance parameters have shown satisfied results with nearly zero percent mismatch losses. The proposed antenna can be used for GSM, GPRS, W-LAN and other radio satellite services.

V. FUTURE WORK

The design can also be implemented via other contacting schemes and in multiple input multiple output designs with stack configuration technique.

REFERENCES

[1] Shareef, A. N., & Shaalan, A. B. Size Reduction of Microstrip Patch Antenna by Using Meta-Fractal Technique.

[2] 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.

[3] M. E. Ermutlu, C. R. Simovski, M. K. Karkkainen, P. Ikonen, S. A. Tretyakov and A. A. Sochava, "Miniaturization of patch antennas with new artificial magnetic layers," IWAT 2005. IEEE International Workshop on Antenna Technology: Small Antennas and Novel Metamaterials, 2005., 2005, pp. 87-90.

[4] J. S. Colburn and Y. Rahmat-Samii, "Patch antennas on externally perforated high dielectric constant substrates," IEEE Trans. Antennas Propag., vol. 47, no. 12, pp. 1785–1794, 1999

[5] Munir, M., Altaf, A., & Hasnain, M. (2015, July). Miniaturization of microstrip fractal H-Shape patch antenna using stack configuration for wireless applications. In Recent Trends in Information Systems (ReTIS), 2015 IEEE 2nd International Conference on Jul 9(pp. 44-48).IEEE

[6] Kiani SH, Qureshi SS, Mahmood K, Mehr-e-Munir, Khan SN. Tri-Band Fractal Patch Antenna for GSM and Satellite Communication Systems. INTERNATIONAL JOURNAL OF ADVANCED COMPUTER SCIENCE AND APPLICATIONS. 2016 Oct 1;7(10):182-6.

[7] Elfergani, I. T. E., Abd-Alhameed, R. A., See, C. H., Sadeghpour, T., & Jones, S. M. R. (2011, November). Reconfigurable antenna design approach for mobile applications and a technique for harmonics suppression. In Antennas and Propagation Conference (LAPC), 2011 Loughborough (pp. 1-4). IEEE.

[8] Manzoor, Z., and G. Moradi. "Optimization of Impedance Bandwidth of a Stacked Microstrip Patch Antenna with the Shape of Parasitic Patch's Slots." Applied Computational Electromagnetics Society Journal 30.9 (2015).

[9] Kiani SH, Mahmood K, Shafeeq S, Munir M, Khan KM. A Novel Design of Miniaturaized Patch Antenna Using Different Substrates for S-Band and C-Band Applications. International Journal of Advanced Computer Science and Applications (IJACSA). 2016 Jul 1;7(7).

[10] Saad Hassan Kiani, Khalid Mahmood, Umar Farooq Khattak, Burhan-Ud-Din and Mehre Munir, "U Patch Antenna using Variable Substrates for Wireless Communication Systems" International Journal of Advanced Computer Science and Applications(IJACSA), 7(12), 2016.

[11] Balanis, Constantine A. *Antenna theory: analysis and design*. John Wiley & Sons, 2016.

[12] Pozar, David M. "Microwave Engineering 3e." Tramsmission Lines and Waveguides (2005): 143-149.

A Preliminary Numerical Simulation Study of Developing Ankle Foot Orthosis to Support Sit-To-Stand Movement in Children with Cerebral Palsy

Chihiro NAKAGAWA
Department of Mechanical Engineering
Osaka Prefecture University
1-1 Gakuen, Naka, Sakai, Osaka, 599-8531, Japan

Ryo YONETSU
School of Comprehensive Rehabilitation
Osaka Prefecture University
3-7-30, Habikino, Habikino-shi, Osaka, 583-8555, Japan

Tomohiro ITO
Department of Mechanical Engineering
Osaka Prefecture University
1-1 Gakuen, Naka, Sakai, Osaka, 599-8531, Japan

Shunsuke KUSADA
Department of Mechanical Engineering
Osaka Prefecture University
1-1 Gakuen, Naka, Sakai, Osaka, 599-8531, Japan

Atsuhiko SHINTANI
Department of Mechanical Engineering
Osaka Prefecture University
1-1 Gakuen, Naka, Sakai, Osaka, 599-8531, Japan

***Abstract*—The purpose of this study is to identify an effective method of support for the standing-up motion of children with cerebral palsy (CP). Experiments revealed remarkable differences in the shank and upper-body motions of children with CP compared with normally developed (ND) children. Shank tilt angles of CP children were smaller and their upper-body tilt angles were larger than those of ND children. The large upper-body tilt compensates for the smaller shank tilt but will cause back pain and/or deformation of the hip joint as they grow. It is therefore imperative to find a method of support to help CP children realize more natural motions (similar to those of ND children) to prevent these problems. The standing-up motion of ND children was adopted as the goal. Experiments identified a similarity in the angular variation between ND children's upper bodies and shanks; the standing-up motion of children with CP under that condition was then simulated using a two-dimensional four-link model of the human body. As a result of the numerical simulation, shank angles of CP children increased and their upper-body angles decreased from those measured during the experiments, which indicates that the proposed method of support is qualitatively effective at allowing CP children to realize a more natural standing-up motion.**

Keywords*—Cerebral; palsy; Standing-up motion; Motion analysis; Numerical simulation; Rigid link model

I. INTRODUCTION

Cerebral palsy (CP) is a chronic neurologic disorder caused by a static lesion in the immature brain and is characterized by deficits in movements and postural control. Thus, motor developments of CP have been more overdue or retarded than that of normally developing children [1]. Specially, CP resulted in difficulty in anti-gravity motor developments as in sit-to-stand (STS) movement because this movement requires adequate balance control between upper body and lower limbs, while the base of the support changes from a relatively larger area to a smaller area supported by the feet [2, 3]. CP often observed excessive ankle plantar flexion during STS movement and had difficulty in shifting the body mass over their feet. In these ways, STS movement in children with CP was accomplished by various abnormal compensatory patterns, which include excessive trunk forward inclination and abrupt knee extension [4, 5]. Therefore, improving STS movement is an important rehabilitation goal for children with CP.

Ankle foot orthosis (AFO) are frequently prescribed to correct skeletal malalignment in CP and has been used improving CP motor function in rehabilitation. However, most studies on effectiveness of AFO have focused on gait [6-9], the effectiveness of AFO on STS movement has limited [10, 11]. Compared with STS movement with barefoot, STS movement by using AFO was characterized by increasing ankle dorsiflexion. However, AFO did not change other proximal compensatory patterns of increased trunk forward titling and hip flexion [11]. This finding suggests that conventional AFO could not support STS movement fully in terms of coordinated motion between upper body and lower limbs. In other words, a new AFO which could promote better coordination between upper body and lower limbs would be needed.

In order to solve this issue, this preliminary study tried to construct a human model consists of 4 rigid bodies. From this model, the equations of motion are obtained. The studies on multi-body system are widely carried out especially in the collaborative control system between human movement and mechanical technology [12, 13]. Therefore, the primary purpose of this study was to assess the kinetic characteristics of STS movement in children with CP by using 3 dimensional motion analysis system. Then, this working gives us a center of mass trajectory during STS to find a suitable control system design in AFO device. Moreover, the secondary purpose of this study was to clarify how our designed new AFO changes STS movement in children with CP by numerical simulations. These findings would help us conduct new AFO device in order to perform STS movement coordinately.

II. Experiment on Standing-Up Motion

A. Experimental method

The subjects for the experiment were two children with CP and two ND children; each child was 5 years old. The joint angles were measured while the subjects completed a standing-up motion from a chair. The height of the chair was adjusted based on the subjects' knee positions, such that the knee and hip joints were orthogonal for each. The subjects were instructed to keep their upper bodies as straight as possible and to cross their arms in front of their chest to prevent their upper body from influencing the entire motion.

Using the motion analysis device Kinema Tracer of KISSEI COMTEC, ten markers were put on the subject's acromion, great trochanter, knee, malleolus, and fifth metatarsal head both sides of their body, as shown in Fig. 1. The motion was recorded by four video cameras and each joint angle θ_i (i =1,2,3,4) was measured by tracking the markers. We found that the angles of the lower body joints were dependent upon those of the upper body. In order to define the angles independently in the simulation, Eqs. (1)–(4) were used (see below). Each angle θ_i measured in the experiment was translated to angle ϕ_i (i =1,2,3,4), defined from the horizontal axis shown in Fig. 2.

$$\phi_1 = 270° - \theta_1 + \theta_2 - \theta_2 - \theta_4 \quad (1)$$

$$\phi_2 = 90° + \theta_2 - \theta_3 - \theta_4 \quad (2)$$

$$\phi_3 = 270° - \theta_3 - \theta_4 \quad (3)$$

$$\phi_4 = 90° - \theta_4 \quad (4)$$

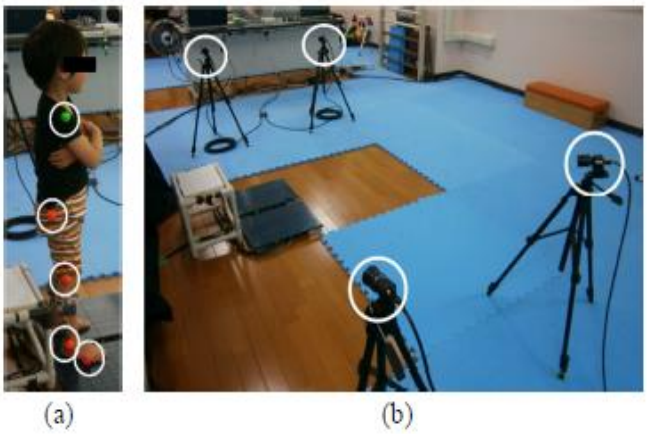


Fig. 1. The 3-dimensional motion analysis system. Positions of the markers are shown in (a). Video cameras were set in the experimental room

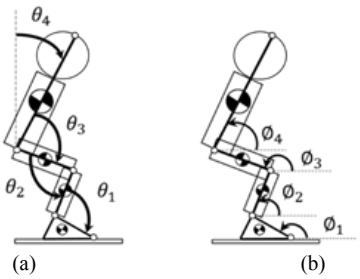


Fig. 2. The angular definition in the experiment is shown in (a) and the angular definition in the simulation is shown in (b)

This research was conducted after the Osaka Prefecture University Research Ethics Committee had approved the study. The purpose of this study was explained to the participants' parents both orally and in writing, and written consent was obtained.

B. Experimental results

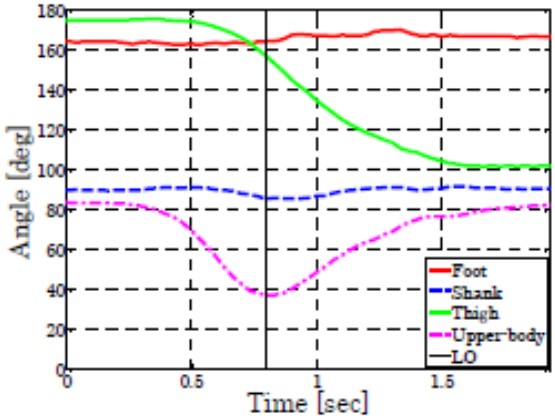
Figure 3 shows the measurement results of each angle ϕ_i derived from Eqs. (1)–(4). The foot and thigh angle variations for both CP and ND children had similar tendencies, and differences between the two groups were not confirmed. The shank angle variations for the CP children were less than those of the ND children; Table 1 shows the initial and maximum tilt angles of the shank. The average maximum tilt angle for the ND subjects was about 10.5°; the maximum tilt angles for the two CP subjects were 4.31° and 8.13°, respectively. We believe that these lower tilt angles for CP children are the result of spasticity making ankle dorsiflexion difficult. Meanwhile, the upper-body angle variations for the CP children were larger than those for the ND children; Table 2 shows the initial and maximum tilt angles of the upper body. The average maximum tilt angle for the ND subjects was about 32.5°; the maximum tilt angles for the two CP subjects were 46.62° and 37.60°, respectively.

TABLE. I. Initial and Peak Angles of the Shank

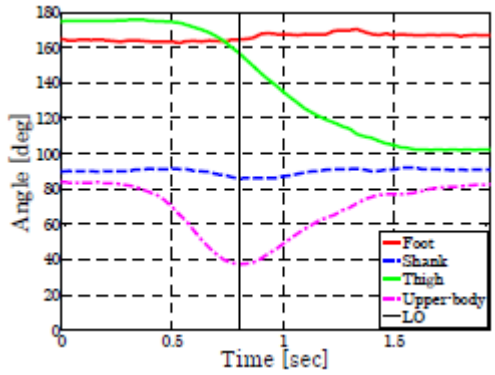
Subject	Initial/°	Peak/°	Max. tilt/°
A (CP)	89.83	85.52	4.31
B (CP)	86.54	78.41	8.13
C (ND)	79.24	68.75	10.49
D (ND)	80.04	69.43	10.61

TABLE. II. Initial and Peak Angles of the Upper Body

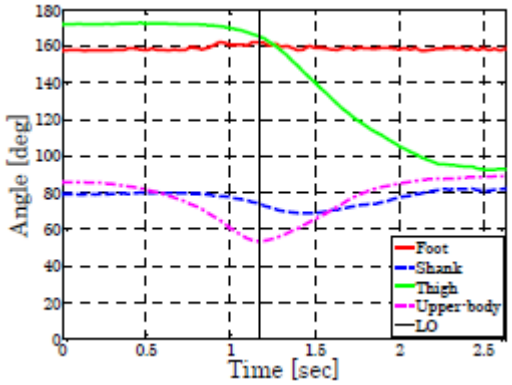
Subject	Initial/°	Peak/°	Max. tilt/°
A (CP)	83.84	37.22	46.62
B (CP)	81.74	44.14	37.60
C (ND)	85.52	53.48	32.04
D (ND)	85.75	52.75	32.97



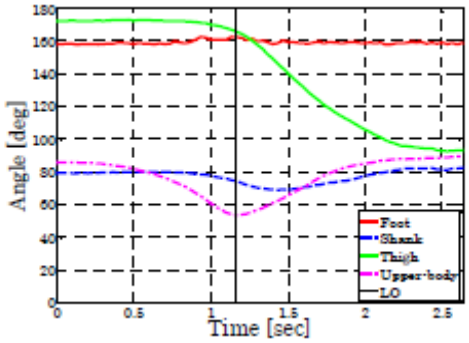
(a) Subject A(CP)



(b) Subject B (CP)



(c) Subject C (ND)



(d) Subject D (ND)

Fig. 3. Angular variations of CP subjects A and B and ND subjects C and D during the standing-up motion. The lines show the angular variations of foot (red), shank (blue), thigh (green), and upper-body (purple); the vertical black dashed line shows the point in time when the subject lifted their hips from the chair (LO)

C. Discussion

These results show that CP children have less shank tilt and more upper-body tilt than ND children during the standing-up motion. Figure 4 shows the correlation diagram of the maximum tilt angles of the shank and upper body; the two have a strong negative correlation. Therefore, it is assumed that the large angle of the upper body plays a considerable role in the smaller tilt angle of the shank, due to the spasticity, during the standing-up motion. The center of gravity (COG)(x_G , y_G) for each subject during the standing-up motion is then defined using a two-dimensional four-link rigid model, shown in Fig. 5. The definitions of each symbol are shown in Table 3 and the index *i* indicates 1: foot, 2: shank, 3: thigh, and 4: upper-body, respectively. The position of the COG for the whole body is defined as Eqs. (5) and (6) using the position of the COG for each body part (x_{iG} , y_{iG}).

$$x_G = \frac{m_1 x_{1G} + m_2 x_{2G} + m_3 x_{3G} + m_4 x_{4G}}{m_1 + m_2 + m_3 + m_4} \quad (5)$$

$$y_G = \frac{m_1 y_{1G} + m_2 y_{2G} + m_3 y_{3G} + m_4 y_{4G}}{m_1 + m_2 + m_3 + m_4} \quad (6)$$

Figure 6 shows the trajectory of the COG of each subject as derived by Eqs. (5) and (6). The initial position is matched. The horizontal axis expresses the variation of the position of the COG for the horizontal direction and the vertical axis expresses that for the vertical direction. The position of the COG transits from the lower left to the upper right. From Fig. 6, we find that the trajectories of the position of the COG for CP children move forward along the horizontal direction more than those for ND children; however, the tendencies are almost similar. From this, we assume that CP children realize a similar transition of the COG as that attained by ND children by tilting their upper bodies excessively during the standing-up motion. This type of movement, used habitually to achieve functional motor skills when a normal movement pattern is unavailable, is called "compensatory movement." It has been suggested that—when there is instability during the standing-up motion—people tilt their upper bodies heavily, placing the COG near the seat (which is the supporting surface) and then standing up under the dynamically stabilized condition (Hirai et al., 2011). Because these motions cause back pain and hip joint deformation (Japanese Association of Rehabilitation Medicine, 2014), suppressing this excessive tilt in the upper body is desirable.

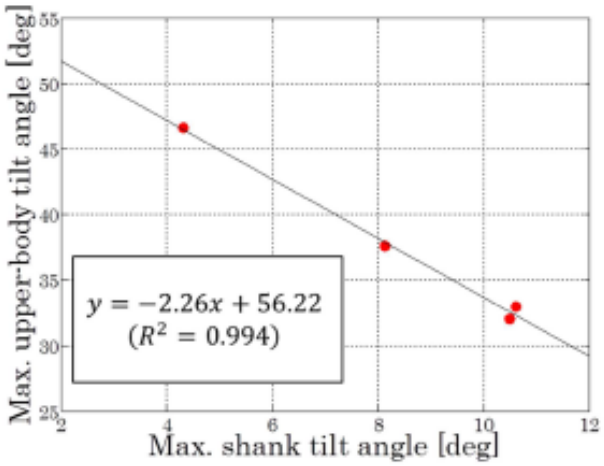


Fig. 4. Correlation between the max. shank tilt angle and max. upper-body tilt angle

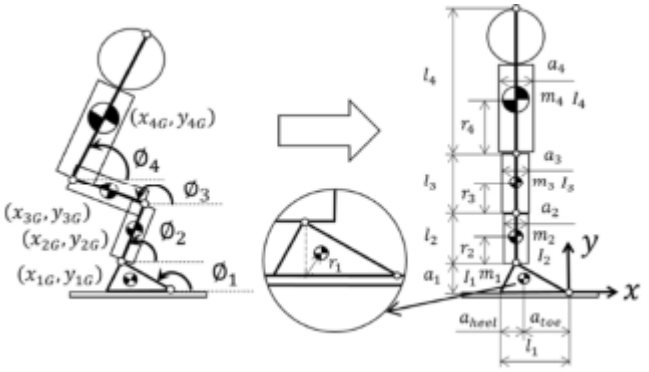


Fig. 5. Definition of the analytical model

TABLE. III. DEFINITIONS OF SYMBOLS IN THE ANALYTICAL MODEL

Symbol	Definition
m_i	Mass of each part i
l_i	Length of each part i
r_i	Length of center of gravity of each part i
h_{heel}	Heel side of foot length
h_{toe}	Toe side of foot length
a_i	Thickness of each part i
I_i	Inertia moment of each part i
l_4	Length from waist to top of head
(x_{ig}, y_{ig})	Center of gravity (COG) of each part i

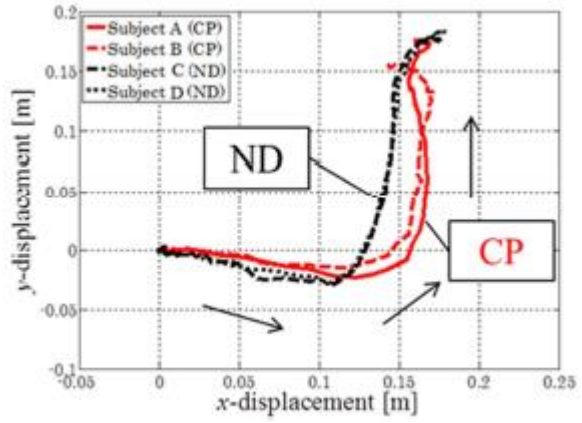


Fig. 6. COG trajectories during the stand-up motion. Red solid and dashed lines show CP subjects A and B; black solid and dashed lines show ND subjects C and D

III. NUMERICAL SIMULATIONS

A. Standing-up motion of ND children

The experiment detailed above shows the problem caused by excessive upper-body tilt in children with CP. Now, the numerical simulations detailed in this section analyze a supporting method designed to allow children with CP to move similarly to ND children in order to prevent the problem caused by excessive upper-body tilt. To do so, we use measurement results from the experiment to clarify the characteristics of ND children's standing-up motions.

Figure 7 shows the angles θ_i (i =1,2,3,4) of each body part of ND children based on the angle definition shown in Fig. 2(a). We find that the tendencies of the knee joint angle θ_2 and hip joint angle θ_3 correspond after the hip is lifted. Therefore, $\theta_2 = \theta_3$ is one characteristic of ND children. This tendency was also confirmed by previous experiments of the standing-up motion in which nine healthy adults (six men and three women) with an average age of 27 years (Doorenbosch et al., 1994) and 47 healthy adults (27 men and 20 women) with an average age of 20.1 years (Tully et al., 2005) were used as subjects. This condition is shown as

$$\theta_2 = \theta_3 \tag{7}$$

By applying the conversion equations shown in Eqs. (1)–(4), following equations are obtained:

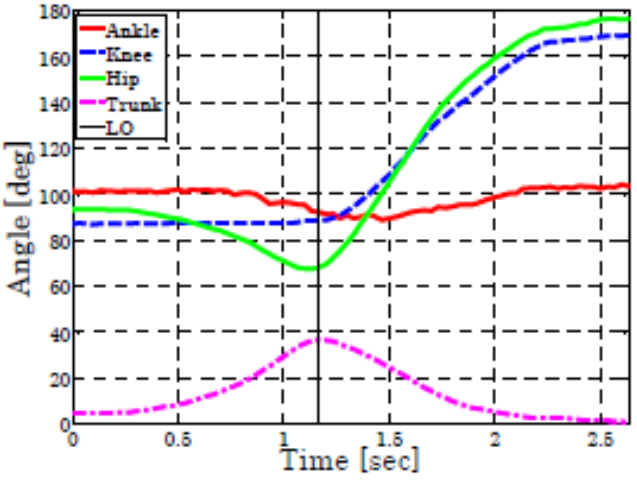
$$\theta_2 = 180° + \phi_2 - \phi_3 \tag{8}$$

$$\theta_3 = 180° - \phi_3 + \phi_4 \tag{9}$$

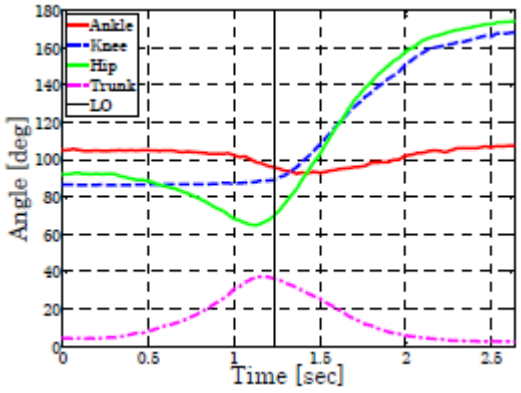
By substituting Eqs. (8) and (9) into Eq. (7), the following condition is obtained:

$$\phi_2 = \phi_4 \tag{10}$$

which means that the lower-leg angle ϕ_2 and upper-body angle ϕ_4 correspond during the standing-up motion. In this study, Eq. (10) is defined as the condition of ND children's standing-up motion.



(a) Subject C (ND)



(b) Subject D (ND)

Fig. 7. Angles (=1,2,3,4) of each body part of ND children

B. Numerical simulations

We have focused on the angles of each body part and on the position of the COG as factors that determine a person's standing-up motion. As shown in Eqs. (5) and (6), the COG's whole-body position is determined by the COG of each rigid body in the analysis model and the COG of each rigid body is determined by the angle of each body part. Therefore, when the COG's position is determined beforehand, the angle of each body part can be derived by inverse operation in the numerical analysis.

As mentioned earlier, there have been studies of AFOs that use actuators to support lower-leg motion. Using a similar idea, we investigate the influence of lower-leg angle changes on the motion of the upper-body angle. By numerically simulating the standing-up motion of CP children, we investigate the possibility of improving both the lack of lower-leg tilt and the excessive upper-body tilt that are characteristic of CP children. We assume that by improving these two characteristics, the standing-up motion of CP children will become more similar to the standing-up motion of ND children. To do so, we use the following three methods:

Method 1.

Target the lower-leg angle of ND children.

Method 2.

Determine the lower-leg angle of CP children based on the standing-up motion of ND children.

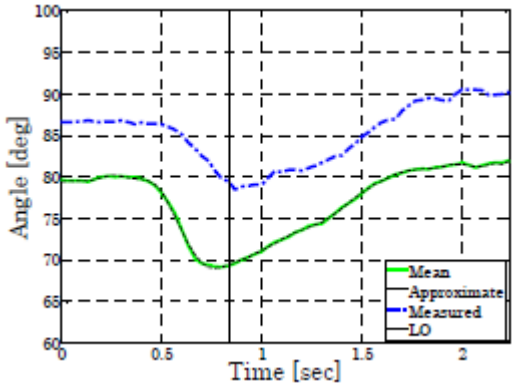
Method 3.

Apply ND children's COG position to CP children.

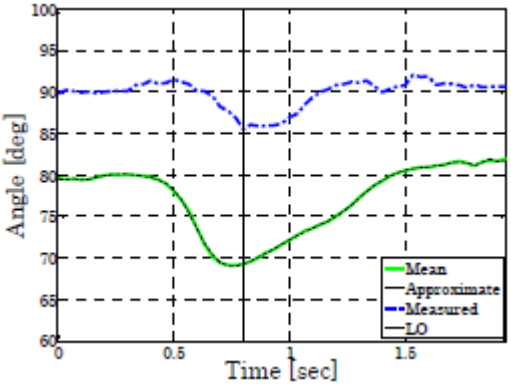
1) Targeting the lower-leg angle of ND children

In this method, ND children's position of COG is regarded as equal to CP children's original position of COG. By inputting the lower-leg angle of ND children, CP children's lower-leg tilts are numerically simulated and upper-body angles derived. From Eq. (5), the definition of COG, the upper-body angle $\emptyset_4$ is derived as

$$\phi_4 = \cos^{-1}\left(\frac{1}{r_4} \left\{ \frac{\left(\sum_{i=1}^{4} m_i\right) x_G - \sum_{i=1}^{3} m_i x_{iG}}{m_4} - \sum_{i=1}^{3} l_i \cos\phi_i \right\} \right) \qquad (11)$$



(a) Subject A (CP)



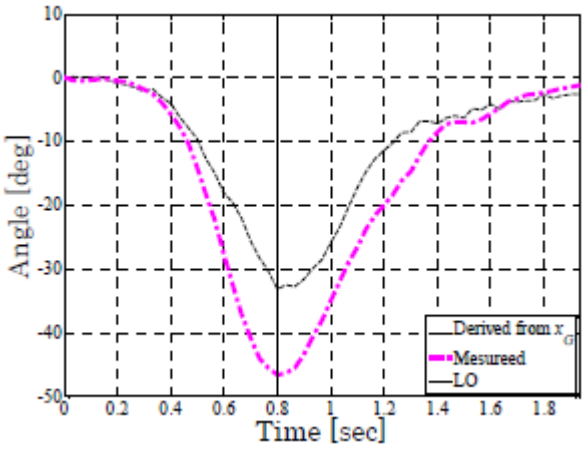
(b) Subject B (CP)

Fig. 8. Comparison of shank angular variations. Each line shows the mean angular variations between ND children (green), the approximate line of mean between ND children (black), and the measured angles of CP subjects A and B (blue)

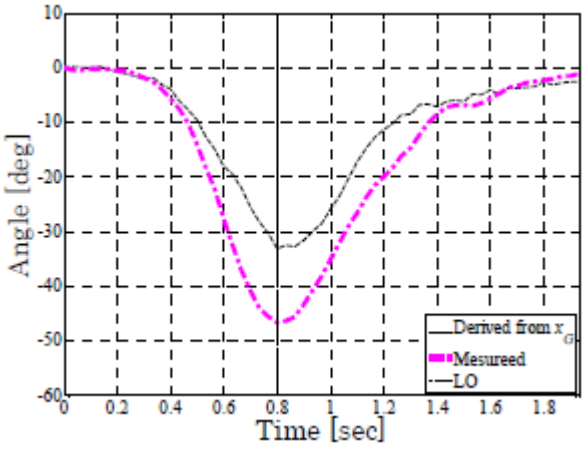
The upper-body angle ϕ_4 is derived by substituting the foot angle ϕ_1, the thigh angle ϕ_3, the CP children's COG position x_G, and the ND children's lower-leg angle. Figure 8 shows the measured values of two CP children's lower-leg angles and the average lower-leg angle of two ND children and its approximate curve. Figure 9 compares the CP children's measured upper-body angles with those derived from Eq. (11). Table 4 shows the maximum tilt angle from the initial condition. We see from Fig. 9 and Table 4 that the maximum tilt angle of the upper body decreases by 13.46° for subject A and 5.26° for subject B. Thus, the upper body's excessive tilt is improved.

TABLE. IV. MAXIMUM MEASURED UPPER-BODY TILT ANGLES, THOSE DERIVED FROM EQ. (11), AND THE DIFFERENCE BETWEEN THEM

Subject	Measured/°	Derived/°	Difference/°
A (CP)	46.62	33.16	13.46
B (CP)	37.60	31.98	5.26



(a) Subject A (CP)



(b) Subject B (CP)

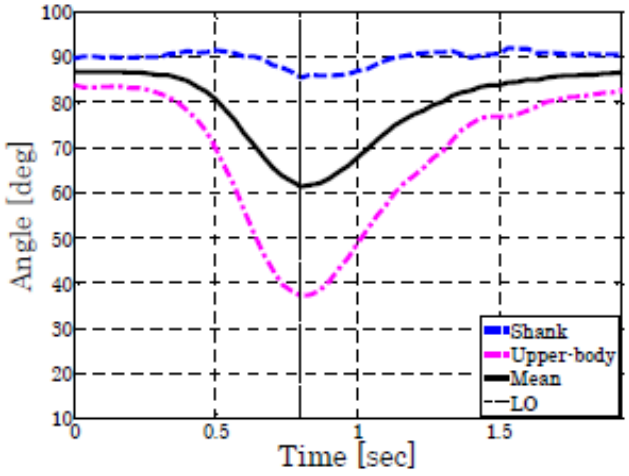
Fig. 9. Comparison of upper-body angular variations. Each line shows the angular variations derived by Eq. (11) (black) and measured (purple) of CP subjects A and B

2) Determining the lower-leg angle based on the standing-up motion of ND children

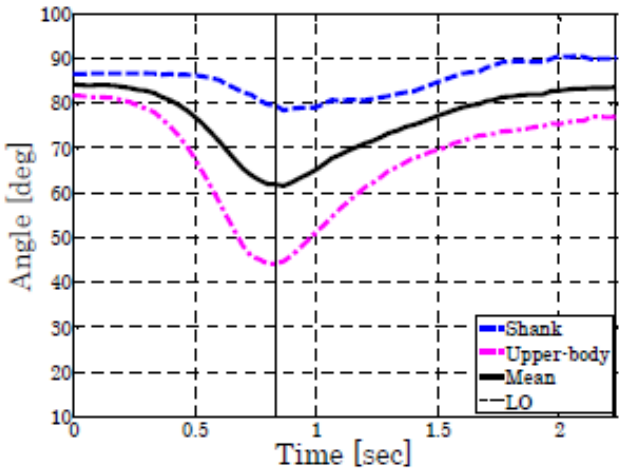
In this method, the position of the COG in CP children is regarded as equal to the original motion of CP children. The upper-body angle is derived from Eq. (11), regarding the average of the lower-leg angle and upper-body angle as the lower-leg angle, based on the condition of ND children's standing-up motion. The angle of the lower leg is defined as

$$\phi_{24} = \frac{\phi_2 + \phi_4}{2} \tag{12}$$

Figure 10 shows the mean of the lower-leg and upper-body angles ϕ_{24}. We see that the mean angle ϕ_{24} is larger than the lower-leg angle ϕ_2 measured in the experiment. Thus, using ϕ_{24} as the lower-leg angle will improve the excessive upper-body tilt. Figure 11 shows the upper-body angle ϕ_4 derived by substituting ϕ_{24} into the lower-leg angle ϕ_2 in Eq. (11).



(a) Subject A (CP)



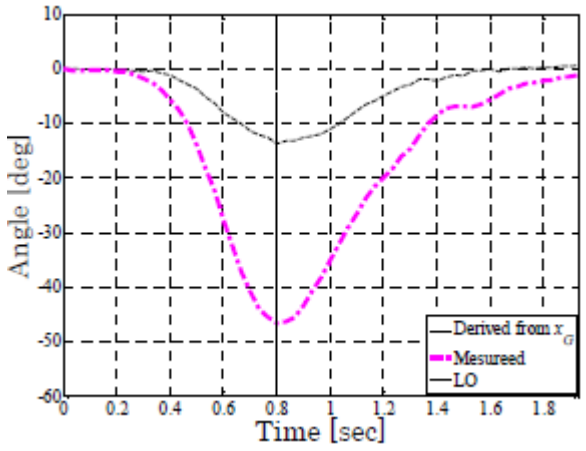
(b) Subject B (CP)

Fig. 10. Comparison of angular variations. Each line shows the angular variations of the measured shank (blue), the measured upper body (purple), and the mean between the shank and upper body (black) of CP subjects A and B

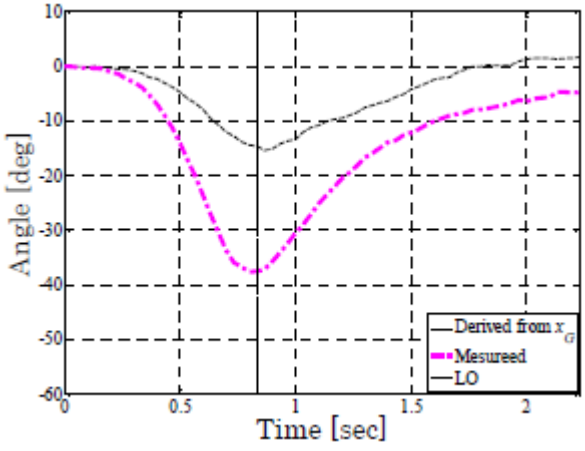
Table 5 shows the maximum tilt angle from the initial posture. From Fig. 11 and Table 5, we see that the maximum upper-body angle decreases. However, the maximum tilt angles of ND children were about 32°, so the angles resulting from the simulation (13.75 and 15.44°) are too small. Therefore, the excessive upper-body tilt was improved but is still different from the standing-up motion of ND children.

TABLE. V. MAXIMUM UPPER-BODY TILT ANGLE MEASURED IN THE EXPERIMENT AND DERIVED USING THE MEAN ANGLE BETWEEN THE MEASURED SHANK AND UPPER-BODY ANGLES, AND THE DIFFERENCE BETWEEN THEM

Subject	Measured/°	Derived/°	Difference°
A (CP)	46.62	13.75	32.87
B (CP)	37.60	15.44	22.16



(a) Subject A (CP)



(b) Subject B (CP)

Fig. 11. Comparison of angular variations. Each line shows the angular variations derived by Eq. (11) (black) and the measured upper-body angles (purple) of CP subjects A and B

3) Applying ND children's COG position

In Section 3.1, conditional Eq. (11) was defined using the measurement data of ND children as a reference. Assuming that lower-leg angle ϕ_2 and upper-body angle ϕ_4 correspond, then angle ϕ_{24} is derived using the equation $\phi_2 = \phi_4 = \phi_{24}$ and the definitional equation of the position of the COG, found in Eq. (5).

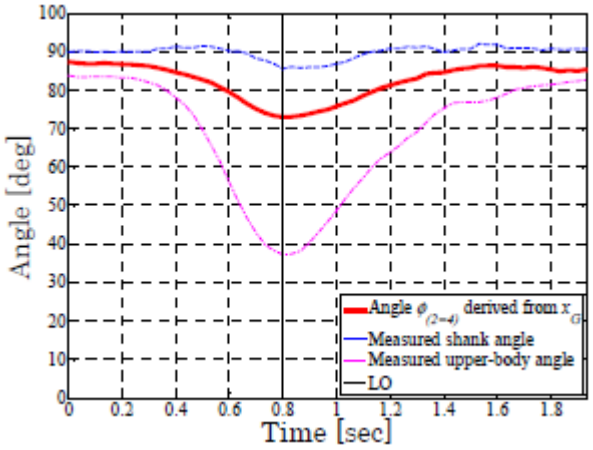
$$\phi_{24} = \cos^{-1}\left[\left\{\frac{\left(\sum_{i=1}^{4} m_i\right)x_G - \left(m_2 r_1 + \sum_{i=2}^{4} m_i l_2\right)\cos\phi_1}{m_2 r_2 + m_3 l_2 + m_4 l_2 + m_4 r_4}\right\} - \left\{\frac{(m_3 r_3 + m_4 l_3)\cos\phi_3}{m_2 r_2 + m_3 l_2 + m_4 l_2 + m_4 r_4}\right\}\right] \quad (13)$$

We substitute the approximate curve of ND children's COG into Eq. (13) and investigate the effect that has on bringing the motions of CP children closer to those of ND children. Figure 12 shows the derived lower-leg and upper-leg angle ϕ_{24}. Table 6 shows the maximum tilt angle from the initial posture.

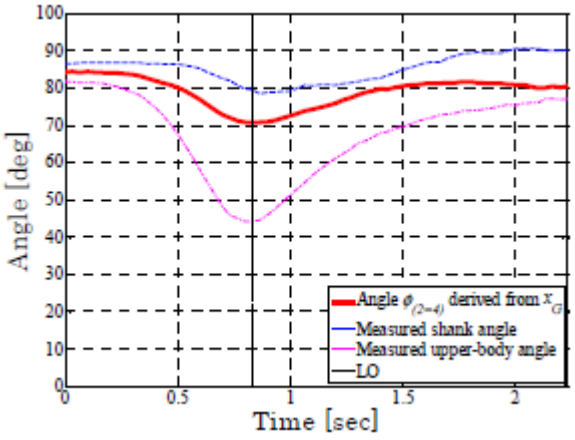
From Fig. 12 and Table 6, we find that the maximum tilt angle of the lower leg increases while that of the upper body decreases. Table 6 shows that the measured value of ND children's maximum lower-leg tilt angle is about 10.5° and that of the upper body is about 32°; however, the derived lower-leg angle in this simulation is larger (to a certain degree), while the derived upper-body angle is much smaller. Therefore, the increased lower-leg tilt and reduced upper-body tilt are improved although the motions of CP children are still different from those of ND children.

TABLE. VI. THE MAXIMUM SHANK AND UPPER-BODY TILT ANGLES MEASURED IN THE EXPERIMENT AND DERIVED BY EQ. (13)

Subject	Measured shank/°	Measured upper-body /°	Derived/°
A (CP)	46.62	4.31	14.34
B (CP)	37.60	8.13	13.50



(a) Subject A (CP)



(b) Subject B (CP)

Fig. 12. Comparison of angular variations. Each line shows the angular variations of the angles derived by Eq. (13) (red), the measured shank (blue), and the measured upper-body (purple) of CP subjects A and B

C. Discussion of the simulation result

Because the measured lower-leg angles of ND children are directly substituted for CP children in the first method (Section 3.2.1), it makes sense that those simulation results most closely resemble the behavior of ND children. We therefore conclude

that AFO support for CP children will be most effective when the lower-leg angles of ND children are taken into consideration.

On the other hand, we also tried to reproduce the behavior of ND children using characteristics of their standing-up motion (Sections 3.2.2 and 3.2.3). In both cases, the numerical simulations confirm improvements in lower-leg and upper-body tilt; however, compared with the actual behavior of ND children, the simulation results show larger lower-leg tilt and less upper-body tilt. We believe this is caused by our not setting a threshold for all angles. By calculating the COG position in the numerical simulation using the substituted lower-leg angle and by setting a threshold tilt angle that avoids the risk of falling down, we believe that behavior closer to that of ND children can be achieved, as can AFO application.

Thus, the simulation results show improvement in the evident characteristics of CP children during their standing-up motion and the possibility of obtaining CP behavior that is nearer that of ND children in a qualitative manner.

IV. Conclusions

Supportive methods for obtaining standing-up motions in CP children that are similar to those of ND children have been investigated with the assumption that AFOs use actuators to support ankle dorsiflexion. Through numerical simulations, we found that it is qualitatively possible to achieve CP standing-up behavior that is near to that of ND children by applying certain characteristics of ND children's behavior, specifically, that the lower-leg angle corresponds to the upper-body angle. The experiments using the proposed methods will be the future works.

Acknowledgment

The part of study is supported by Adaptable and Seamless Technology transfer Program through target-driven R&D (A-STEP) Grant (AS242Z01706K). We appreciate the support of Japan Science and Technology Agency.

References

[1] Milani-Comparetti A, Gidoni AE. "Pattern analysis of motor development and its disorders," Dev Med child Neurol 9, pp.625-630, 1967.

[2] Riley PO, Schneckman ML, Mann RW, Hodge WA. "Mechanics of a constrained chair-rise," J Biomech, 24, pp. 77-85, 1991.

[3] Schultz AB, Alexander NB, Ashton-Miller JA. "Biomechanical analyses of rising from a chair". J Biomech, 25, pp.1383–1391, 1992.

[4] Park ES, Park CI, Lee HJ, Kim DY, Lee DS, Cho SR. "The characteristics of sit-to-stand transfer in young children with spastic cerebral palsy based on kinematic and kinetic data", Gait Posture, 17, pp.43-49, 2003.

[5] Yonetsu R, Iwata A, Surya J, Unase K, Shimizu J. "Sit-to-stand movement changes in preschool-aged children with spastic diplegia following one neurodevelopmental treatment session- a pilot study" Disabil Rehabil, 37, pp.1643-1650, 2015.

[6] Radtka SA, Skinner SR, Dixon DM, et al. "A comparison of gait with solid, dynamic, and no ankle-foot orthoses in children with spastic cerebral palsy", Phys Ther, 77, pp.395-409, 1997.

[7] Crenshaw S, Herzog R, Castagno P, et al. "The efficacy of tone-reducing features in orthotics on the gait of children with spastic diplegic cerebral palsy", J Pediatr Orthop, 20, pp.210-216, 2000.

[8] Desllovere K, Molenaers G, Gestel LV, et al. "How can push-off be preserved during use of an ankle foot orthosis in children with hemiplegia? A prospective controlled study", Gait Posture, 24, pp.142-151, 2006.

[9] Van Gestel L, Molenaers G, Huenaerts C, et al. "Effect of dynamic orthoses on gait: a retrospective control study in children with hemiplegia", Dev Med Child Neurol, 50, pp.63-67, 2008.

[10] Wilson H, Harideri N, Song K, Telford D. "Ankle-foot orthoses for preambulatory children with spastic diplegia", J Pediatr Orthop, 17, pp.370-376, 1997.

[11] Park ES, Park CI, Chang HJ, Choi JE, Lee DS. "The effect of hinged ankle-foot orthoses on sit-to-stand transfer in children with spastic cerebral palsy", Arch Phys Med Rehabil, 85, pp.2053-2057, 2004.

[12] Nakagawa C, Morita Y. Shintani A, Ito T., "Standing posture analysis of a human on a four-wheel stand-up-type personal mobility vehicle", Transaction of the JSME (in Japanese), 82: No,838: 1-13, 2016.

[13] Kim JH, Xiang Y, Yang J, Arora JS, Abdel-Malek K, "Dynamic motion planning of overarm throw for a biped human multibody system", Multibody Syst Dyn, 24, pp.1–24, 2010.

A Survey of Spam Detection Methods on Twitter

Abdullah Talha Kabakus
Abant Izzet Baysal University
IT Center
Bolu, Turkey

Resul Kara
Duzce University
Faculty of Engineering,
Department of Computer Engineering
Duzce, Turkey

***Abstract*—Twitter is one of the most popular social media platforms that has 313 million monthly active users which post 500 million tweets per day. This popularity attracts the attention of spammers who use Twitter for their malicious aims such as phishing legitimate users or spreading malicious software and advertises through URLs shared within tweets, aggressively follow/unfollow legitimate users and hijack trending topics to attract their attention, propagating pornography. In August of 2014, Twitter revealed that 8.5% of its monthly active users which equals approximately 23 million users have automatically contacted their servers for regular updates. Thus, detecting and filtering spammers from legitimate users are mandatory in order to provide a spam-free environment in Twitter. In this paper, features of Twitter spam detection presented with discussing their effectiveness. Also, Twitter spam detection methods are categorized and discussed with their pros and cons. The outdated features of Twitter which are commonly used by Twitter spam detection approaches are highlighted. Some new features of Twitter which, to the best of our knowledge, have not been mentioned by any other works are also presented.**

Keywords*—Twitter spam; spam detection; spam filtering; mobile security

I. INTRODUCTION

Twitter is one of the most popular social media platforms which provide a social network of users post messages up to 140 characters called as "tweet". Twitter lets users share their messages about everything related to the real life including news, events, celebrities, politics [1–5]. According to Twitter, Twitter has 313 million monthly active users that post 500 million tweets per day which equal 350,000 tweets per minute [6–8]. Thanks to this huge social network, users are able to stay connected with the topics they are interested in. Twitter provides a list of most talked topics at a given point in time called "Trending Topics (TT)" to let users be aware of most popular topics on Twitter. "Hashtag" is a term which starts with "#" character is commonly used to mention the topic of the tweet and let users track the topics they are interested in [9] . Thanks to its popularity and design, Twitter immediately reflects noteworthy events in real-time. This structure of Twitter lets real-time search systems and meme-tracking services mine real-time tweets to find out what is happening in the world with minimum delay [10,11]. Sentiment analyzing services are able to make a conclusion about topics in Twitter which turns Twitter into a real-time poll system [12–16]. The success of those services completely relies on filtering spammers from legitimate users. Consumers tend to use Twitter to learn ideas of others about the products they are going to buy. Similarly, companies use Twitter to measure the satisfaction of their customers for their products [17–21]. However, this popularity and practicalness also attract the attention of spammers. In April of 2014, Twitter was flooded by an avalanche of malicious tweets that were sent by thousands of compromised user accounts [22]. In August of 2014, Twitter revealed that 8.5% of its monthly active users which equals approximately 23 million users have automatically contacted their servers for regular updates [23,24]. A report shows that 83% users of social networks have received at least one unwanted friend request or message [25]. Most common definition of spam is unsolicited one [26–28]. Spammers share links within their tweets in order to spread advertise to generate sales, propagate pornography, share malicious links which direct users to malicious software, hijack trending topics for their purposes, abuses reply or mention functions to post unsolicited messages to legitimate users to attract their attention, and phish legitimate users [1,21,28–37]. According to the report by statista, 80% of Twitter users access Twitter via their mobile devices [38]. Thus, users who access Twitter via their mobile devices should more care about spam than the users who access Twitter via web browsers since it may (1) collect excessive amount of personal information such as user location, call history, SMS, bank account details, calendar events, (2) access the data located in the device's memory or SD card, (3) send premium-rate SMS messages, (4) capture key-strokes by key logging, (5) make calls, and (6) detect user's location via Internet or GPS and share [39–45]. Another issue with users of social media is that according to the reports, users of social media do not show an adequate understanding of the threats of social media as much as they are on other platforms. Bilge et al. [46] report that 45% of users on a social media platform readily click on links posted by their "friends", even though they may not know that person in real life. Content-filtering approaches are not effective for Twitter since spammers tend to share shorten URLs in order to (1) overcome the character limitation defined by Twitter, and (2) manipulate spam filtering methods based on URL blacklisting [28,36,47–52]. The major contributions of this paper are given as follows:

- Features of Twitter which can be used to detect spam are presented with discussing their effectiveness,
- A comprehensive review of Twitter spam detection methods are discussed with considering their pros and cons in order to give a clear idea to the researchers who are interested in spam detection in Twitter,
- The new features of Twitter which, to the best of our knowledge, have not been used by any spam detection

approaches yet that can be used to detect spam are presented,

- The outdated features of Twitter which are commonly used by spam detection approaches in literature are presented.

The rest of the paper is structured as follows: Section 2 describes the background including features of Twitter and how Twitter deals with spam. Section 3 presents the features of Twitter spam detection. Section 4 presents the Twitter spam detection methods. Section 5 presents discussion. Finally, Section 6 concludes the paper.

II. BACKGROUND

In this section, features of Twitter and the way Twitter deals with spam are presented.

A. Features of Twitter

Twitter lets accounts to "follow" other accounts which they are interested in. Unlike other social media platforms, the relationship between users is bi-directional instead of unidirectional links which mean one user may not be following one of his followers. The user can "like" or "retweet (RT)" a tweet which means sharing that tweet with his "followers". The relationship between users in Twitter is presented in Fig. 1. Each user has a unique Twitter username, and users can post tweets that refer others by adding their usernames with starting "@" character which is called as "mention" on Twitter. Users are immediately informed with notifications when a mention, like, or RT happens to one of his tweets.

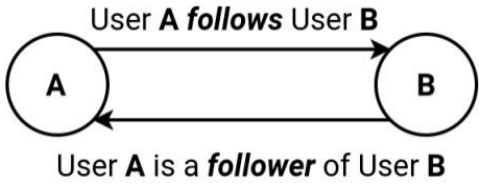


Fig. 1. The relationship between users in Twitter

Another feature of Twitter is letting users create user public or private lists in order to organize their interests by grouping users whose interests are same or similar [53–55]. Similarly, it is possible to manage lists by adding users to the lists or removing users from the lists which the user is the owner of. The lists the user subscribed are categorized as "subscribed to" while the lists the user is added by their owners are categorized as "member of" which are presented in Fig. 2.

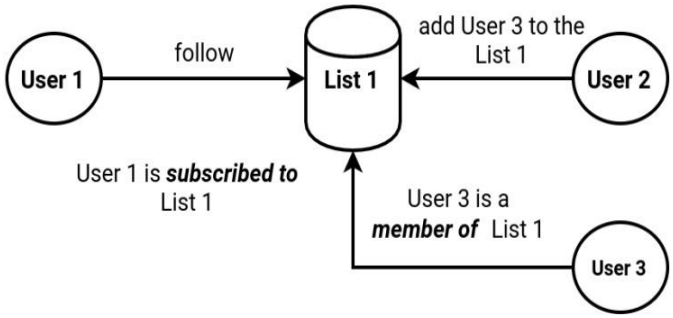


Fig. 2. The relationships between lists and users

B. How Twitter Deals with Spam

Twitter uses both manual and automated services to compete spammers in order to provide a spam-free environment. The manual way is that Twitter lets users report spammers through the spammers' profile pages. Twitter provides a user interface as it is presented in Fig. 3 to report the account by selecting the reason. Another way which is commonly reported in the literature is mentioning spammers to the official "@spam" account [28,29,37,56–58] but according to the recent report by Twitter, this method of reporting spam is outdated [30]. Also, Wang reports that this method is abused by both hoaxes and spam [29]. These manual approaches are labor-intensive and would not be enough to detect all spammers considering billions of users. Twitter uses various factors such as (1) posting duplicate messages over multiple accounts or multiple duplicate messages on one account, (2) following/unfollowing large number of accounts in a short time period, (3) having large number of spam complaints filed against the account, (4) aggressively liking, following, and retweeting, (5) posting malicious links, (6) posting tweets which mainly consist of links instead of also posting personal updates, and (7) posting unrelated tweets to a trending topic to determine what conduct is considered to be spamming [59].

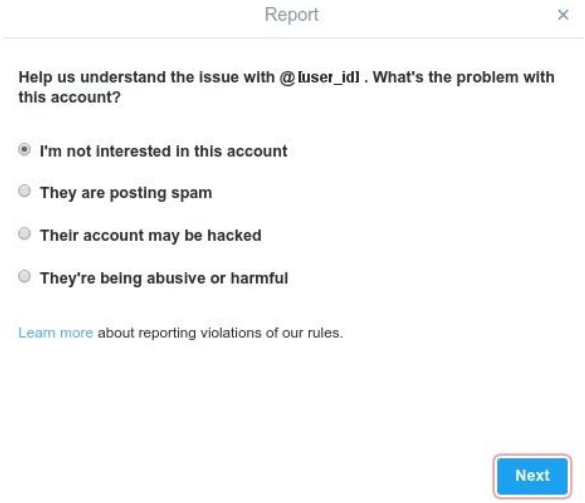


Fig. 3. The user interface of Twitter which is used to report an account by selecting the reason

III. FEATURES OF TWITTER SPAM DETECTION

The features of Twitter spam detection are categorized as follows: (1) Account-based features, (2) tweet-based features, and (3) relationship between the tweet's sender and receiver. These features are the mainframes of the features used by the related works in literature. Each feature category is discussed in the following subsections.

A. Account-based Features

Spammers can be detected by analyzing their Twitter accounts which contain the features listed in Table 1. Since some of these features such as biography, location, homepage, and creation date are user-controlled, they are useless in term of spam detection

TABLE. I. ACCOUNT-BASED SPAM DETECTION FEATURES

Feature	Description	Is User-controlled?
Username	The unique identifier of the account	Yes
Biography	The biography of the account	Yes
Profile photo	The profile photo of the account	Yes
Header photo	The header photo of the account which is displayed at the top of the profile	Yes
Theme color	The theme color choice of the account	Yes
Birth date	The birth date information of the account	Yes
Homepage	The website of the account	Yes
Location	The location of the account	Yes
Creation date	The date the account is created	Yes
Number of tweets	Total number of tweets the account has	No
Number of following	Total number of accounts the account follows	No
Number of followers	Total number of followers the account has	No
Number of likes	Total number of likes the account's tweets have	No
Number of retweets	Total number of retweets the account's tweets have	No
Number of lists	Total number of lists the account has	Yes
Number of moments	Total number of moments the account has	Yes

When the behaviors of spammers are analyzed within the scope of account-based features, these facts are observed:

- Since spammers tend to follow too many legitimate accounts in order to attract attention, the number of following is expected to be high compared to legitimate users.
- Since spammers are not followed by legitimate users, the number of followers is expected to be less compared to legitimate users.
- Since spammers' tweets are unsolicited, the number of likes and retweets for their tweets are expected to be less compared to legitimate users.
- Since spammers tend to post lots of tweets to attract the attention of legitimate users, the number of tweets sent by the account is expected to be high compared to legitimate users.
- Spammers' tweets mostly contain links and hashtags to attract the attention of legitimate users.
- Since spammers' tweets are ignored by legitimate users, the number of replies and mentions spammers get are expected to be low compared to legitimate users.
- Spammers tend to post same or similar tweets which are posted by one or more controlled accounts.
- Legitimate users tend to be added to the lists unlike spammers unless bots under the command and control (C&C) architecture add them to the lists they intentionally created in order to manipulate spam detection approaches.

B. Tweet-based Features

Spammers tend to post lots of unsolicited tweets to legitimate users to attract attention. Spammers can be detected by analyzing their tweets. This is necessary to filter spam tweets from legitimate ones and provide users a spam-free environment which is the aim of Twitter [60]. Each tweet contains the information listed in Table 2.

TABLE. II. TWEET-BASED SPAM DETECTION FEATURES

Feature	Description	Is User-controlled?
Sender	The sender of the tweet	Yes
Mentions	The mention(s) used in the tweet	Yes
Hashtags	The hashtag(s) used in the tweet	Yes
Link	The link used in the tweet	Yes
Number of likes	The number of likes the tweet has	No
Number of retweets	The number of retweets the tweet has	No
Number of replies	The number of replies the tweet has received	No
Sent date	The date tweet is sent	Yes
Location	The detected location of the place the tweet is posted	Yes

When the behaviors of spammers are analyzed within the scope of tweet-based features, these facts are observed:

- Spammers tend to use links to direct legitimate users to their malicious purposes.
- Spammers tend to use lots of mentions to attract the attention of more legitimate users.
- Spammers tend to use lots of hashtags (especially the trending ones) to reach more users.
- Since spammers' tweets are unsolicited, the number of likes and retweets their tweets have received are much lower compared to legitimate users.

C. Graph-based Features

Twitter is a network of users with relationships between them and tweets. This structure can be represented as a graph. For the graph model, users and tweet can be represented as nodes and relationships can be represented links between nodes. These relationships show how the tweet's sender and mentions are connected to each other. Also, these relationships are clear indicators of legitimate conversations. By constructing a graph model to represent users and their relationships, the distance between the tweet's sender and mentions can be calculated for spam analysis. Graph-based features are listed in Table 3.

TABLE. III. GRAPH-BASED FEATURES

Feature	Description	Is User-controlled?
Distance	The length of the shortest path between users	No
Connectivity	The strength of the connection	No

When the behaviors of spammers are analyzed within the scope of graph-based features, these facts are observed:

- The distance between a spammer and a legitimate user is further than the distance between two legitimate users.
- The connectivity between a spammer and a legitimate user is more robust than the connectivity between two legitimate users.
- Graph-based features provide the most robust performance to detect spam and spammers since they are hard to manipulate and not user-controlled.

IV. TWITTER SPAM DETECTION METHODS

In this section, Twitter spam detection methods in literature are presented and discussed. The proposed methods are categorized as follows: (1) Account-based spam detection methods, (2) tweet-based spam detection methods, (3) graph-based spam detection methods, and (4) hybrid spam detection methods.

A. Account-based Spam Detection Methods

Account-based spam detection methods are based on the features (or a combination of them) of Twitter account which are listed in Table 1. Lee et al. [61] propose a honeypot-based approach to detect spam in social media platforms. The features they consider detecting spam are the longevity of the account on Twitter, the average tweets per day, the ratio of the number of following and number of followers, the percentage of bi-directional friends, the ratio of the number of URLs in the 20 most recently posted tweets, the ratio of number of unique URLs in the 20 most recently posted tweets, the ratio of the number of usernames in the 20 most recently posted tweets, and the ratio of the number of unique usernames in the 20 most recently posted tweets. Lin and Huang [62] propose a method to detect spam in Twitter on the basis of two features: (1) URL rate which defines the ratio of the number of tweets with URL in the total number of tweets, and (2) interaction rate which defines the ratio of the number of tweets interacting over the total number of tweets. Gee and Hakson [58] propose a method based on account-based features such as followers-to-following ratio, the number of tweets to account lifetime ratio, the average time between posts, posting time variation, max idle hours, and link fraction. The limitation of this work is that they utilize the manual way of reporting spam in Twitter which is outdated as it is discussed before. Many Twitter spam detection methods use account-based features but alongside with other spam detection features in order to provide more robust spam detection methods which are called as "hybrid" spam detection methods in this paper.

B. Tweet-based Spam Detection Methods

Tweet-based spam detection methods are based on the features (or combinations of them) of a tweet which are listed in Table 2. URL filtering approaches use static or dynamic crawlers to investigate newly observed URLs. Also, they use URL or domain blacklisting in order to detect suspicious URLs from a knowledge base. These approaches use several features such as URL and DNS information, URL redirections, and the landing website's source code (HTML). McGrath and Gupta [47] present a phishing detection method based on lexical features of an URL. The features they consider detecting phishing are the length of URL and the domain name, the character composition of the domain name, the presence of brands in URLs, and misuse of URL-aliasing and free web hosting services. Ma et al. [63] propose a method to detect malicious websites by analyzing their URLs. The features they use detecting malicious websites contain WHOIS properties such as who is the registrar of the website, who is the registrant of the website, when the website is registered, domain name properties such as the time-to-live (TTL) value for DNS records, and geographic properties such as in which country does the IP address belong, the speed of the uplink connection alongside lexical features of URL. Prophiler [64] is a filter that uses static analysis techniques to detect the malicious content of a website. The features Prophiler considers are derived from (1) the HTML content of the website such as the number of elements with small area, the number of elements contain suspicious content, the number of included URLs, and the number of known malicious patterns, (2) the associated JavaScript code such as keywords-to-words ratio, the number of long strings presence of decoding routines, probability of shellcode presence, and the number of DOM-modifying function, and (3) the corresponding URL such as the number of suspicious URL patterns, presence of subdomains or IP addresses in URLs, and the TTL value for DNS A and NS record. Since Prophiler uses static analysis techniques, it is not able to detect malicious URLs embedded into dynamic content such as part of JavaScript which is currently the most commonly used programming language [65,66], Flash, and Java applets. Methods based on dynamic analysis techniques [67–70] use virtual machines and automated web browsers such as Selenium for in-depth content analysis. Chhabra et al. [49] present a phishing detection method based on URL analysis. Their method is specially designed to be able to analyze shortened URLs which are commonly used in Twitter to manipulate spam tweets as it is discussed before. The features the proposed method use detecting phishing through an URL are the number of clicks, geographical spread, temporal spread, and web popularity. WarningBird [71] is a suspicious URL detection system for Twitter which investigates correlations of URL redirect chains. WarningBird uses 14 features to detect suspicious URL such as the length of URL redirect, the number of different landing URLs, the relative number of different Twitter accounts, the similarity in the account creation dates, the similarity in the number of followers and following, the similarity in the follower-following ratio, and the similarity of tweets. Martinez-Romo

and Ajauro [72] propose a tweet-based spam detection method which focuses on the analysis of the language used in tweets. Specifically, the language models they use are (1) the language model of the tweets related to a trending topic, (2) the language model of the tweet, and (3) the language model of the page linked by the tweet. Similar to the account-based spam detection methods, many Twitter spam detection methods use tweet-based features alongside with other spam detection features in order to provide more robust spam detection.

C. Graph-based Spam Detection Methods

Graph-based spam detection methods are based on the features (or combinations of them) of a tweet which are listed in Table 2. Song et al. [28] extract the distance and connectivity between the tweet's sender and mentions. While distance defines the length of the shortest path between the tweet's sender and mentions, connection defines the strength of the connection between users. Graph-based spam detection methods use graph data structures to model features of Twitter as nodes and edges. Graph data models are the perfect solution to represent the data where information about data interconnectivity or topology is at least as important as the data itself [73]. Thus, graphs are commonly used by social networks such as Facebook, Twitter [74–81] which are mostly built on users, topics, and bi-directional interactions. Despite that graph-based features provide the best performance in terms of accuracy and sensitivity to differentiate spammers from legitimate users, other graph-based spam detection methods are presented in hybrid spam detection methods since they are combined with other spam detection methods.

D. Hybrid Spam Detection Methods

Hybrid spam detection methods use a combination of spam detection methods described in previous subsections in order to provide more robust spam detection which investigates the possibility of spam in a more comprehensive way. Stringing et al. [51] propose an approach based on both account-based and tweet-based features which are the ratio of the number of friend requests that the user sent to the number of friends she has, the ratio of the number of tweets which contain URLs to the total number of tweets the user has, the similarity of tweets sent by the user, the number of tweets sent by the user, the number of friends the user has, and the possibility of whether an account likely used a list of names to pick its friends or not. Gao et al. [82] propose a tweet-based spam detection approach based on the social degree of the tweet's sender, the history of interaction, the size of the cluster, the average time interval, the average number of URL in tweets, and the unique number of URL in tweets. Chen et al. [83] present a real-time spam detection method for Twitter based on 12 lightweight features which are extracted from a dataset contains 6.5 million spam tweets. The features they consider detecting spam on Twitter are age of the account, the number of followers, the number of following, the number of likes the account received, the number of the account's lists, the number of tweets of the account, the number of retweets of the tweet, the number of hashtags used in the tweet, the number of mentioned users in the tweet, the number of URLs used in the tweet, the number of characters used in the tweet, and the number of digits used in the tweet. Wang [29] proposes a hybrid Twitter spam detection method based on graph-based and tweet-based features. The graph-based features considered in the proposed method are the number of followers, the number of following, a reputation score which is calculated as the ratio between the number of followers over the total sum of the number of followers and following, and the number of following. The tweet-based features considered in the proposed method are tweet similarity, the number of tweets which contain URLs in the most recent 20 tweets, the number of tweets contains mentions in the most recent 20 tweets, and the number of tweets contains hashtags. Yang et al. [84] propose a Twitter spam detection method based on a combination of graph-based, tweet-based, and account-based features. The proposed method uses more robust features including the number of bi-directional links, the ratio of bi-directional links, betweenness centrality, clustering coefficient alongside tweet-based and account-based features such as the number of followers, the number of following, the number of tweets sent by the account, the age of the account, the ratio of the number of tweets contain URL, the ratio of the number of tweets contain hashtags, the number of duplicate tweets, the ratio of spam word, the ratio of the number of tweets used to reply to others, and the ratio of the number of retweets. Benevenuto et al. [1] propose a hybrid spam detection method based on account-based features such as the number of followers, the number of following, the ratio between followers over following, the number of tweets sent by the account, the number of mentions the account received, the number of replies, and the ratio of tweets received from the account's followers. The tweet-based based features of the proposed method are the number of words in each tweet, the number of URLs per word, the number of words of each tweet, the number of characters of each tweet, the number of hashtags on each tweet, the number of mentions on each tweet, the number of URLs of each tweet, and the number of times the tweet is retweeted. Chu et al. [48] present a method to categorize Twitter accounts as human, bot, and cyborg which is based on both account-based and tweet-based features. The features they consider categorizing the Twitter account into human, bot or cyborg are the number of the ratio of tweets contain URLs, device makeup, the number of the ratio of followers to friends, link safety, and whether the account is verified. Amleshwaram et al. [85] propose a hybrid Twitter spam detection method based on both account-based and tweet-based features. They categorize spammers into two: (1) users centric, and (2) URL-centric. The features they consider for spam analysis are the number of unique mentions, unsolicited mentions, hijacking trends, intersection with famous trends, variance in tweet intervals (VaTi), variance in number of tweets per unit time (VaTw), ratio of VaTi and VaTw, tweet sources, duplicate URLs, duplicate domain names, IP/domain fluxing, tweet's language dissimilarity, similarity between tweets, URL and tweet similarity, followers-to-following ratio, and profile description’s language dissimilarity. Chakraborty et al. [86] propose a hybrid method based on account-based and tweet-based features which use some new features such as spam score of profile description, name, and screen name, presence or absence of profile image and average same hashtag count. McCord and Chuah [9] present a hybrid method based on account-based and tweet-based features to facilitate spam detection. The features they use in the proposed method are the distribution of tweets over a

24-hour period, the number of URLs, the total number of replies/mentions in the most 100 recent tweets, the number of retweets in the 20-100 most recent tweets, the total number of hashtags in the 100 most recent tweets. Wang et al. [87] propose a spam detection method based on account-based, tweet-based, natural language processing (NLP), and sentiment features. Some unique features they use while detecting spam are length of the profile name, automatically or manually created sentiment lexicons, the number of exclamation marks, the number of question marks, maximum word length, mean word length, the number of capitalization words, the number of white spaces, and part of speech (POS) tags per tweet. Outline of the related works including their methodologies, the categories their metrics are based on, and accuracies are listed in Table 4.

TABLE. IV. OUTLINE OF THE RELATED WORKS INCLUDING THEIR METHODOLOGIES, THE CATEGORIES THEIR METRICS ARE BASED ON, AND ACCURACIES

Title	Methodology	Metrics Based on	Accuracy
"Uncovering Social Spammers: Social Honeypots + Machine Learning" [61]	Decorate, LogitBoost, HyperPipes, Bagging, RandomSubSpace, BFTree, FT, SimpleLogistic, LibSVM, ClassificationViaRegression	Account	99.21%
"Beyond blacklists: learning to detect malicious web sites from suspicious URLs" [63]	Naive Bayesian, SVM with linear kernel, SVM with an RBF kernel, l1-regularized logistic regression	Tweet	95-99%
"Prophiler: A Fast Filter for the Large-Scale Detection of Malicious Web Pages" [64]	Naive Bayesian, Random Forest, Decision Tree, Logistic Regression	Tweet	90.41%
"WarningBird: A Near Real-Time Detection System for Suspicious URLs in Twitter Stream" [71]	LIBLINEAR	Tweet	0.9028
"Spam Filtering in Twitter using Sender-Receiver Relationship" [28]	Bagging, LibSVM, Decision Tree, Bayes Network, FT	Graph	99.7%
"Towards Online Spam Filtering in Social Networks" [82]	Decision Tree	Hybrid	TPR with 80.8%, FPR with 0.32%
"6 Million Spam Tweets: A Large Ground Truth for Timely Twitter Spam Detection" [83]	Random Forest, Decision Tree, Bayes Network, Naive Bayesian, k-NN, SVM	Hybrid	TPR with 90%
"Don't follow me: Spam detection in Twitter" [29]	Naive Bayesian, Neural Network, SVM, Decision Tree	Hybrid	93.5%
"Die Free or Live Hard? Empirical Evaluation and New Design for Fighting Evolving Twitter Spammers" [84]	Random Forest, Decision Tree, Decorate, Naive Bayesian	Hybrid	88.6%
"Detecting spammers on Twitter" [1]	SVM	Hybrid	87.6%
"Who is Tweeting on Twitter: Human, Bot, or Cyborg?" [48]	Bayesian	Hybrid	TPR with 90.47%
"CATS: Characterizing Automation of Twitter Spammers" [85]	Random Forest, Decision Tree, Decorate, Naive Bayesian	Hybrid	93.6%
"SPAM: A Framework for Social Profile Abuse Monitoring" [86]	Random Forest, Decision Tree, SVM, Naive Bayesian	Hybrid	89%
"Spam Detection on Twitter Using Traditional Classifiers" [9]	Random Forest, Decision Tree, Naive Bayesian, k-NN	Hybrid	95.7%
"A study of effective features for detecting long-surviving Twitter spam accounts" [62]	Decision Tree	Account	Precision with 86%
"Twitter Spammer Profile Detection" [58]	Naive Bayesian, SVM	Account	89.6%
"Detecting Spammers on Social Networks" [51]	Random Forest	Hybrid	90.93%
"Making the Most of Tweet-Inherent Features for Social Spam Detection on Twitter" [87]	Naive Bayesian, k-NN, SVM, Decision Tree, Random Forest	Hybrid	Precision with 94.6%
"Detecting malicious tweets in trending topics using a statistical analysis of language" [72]	Decision Tree, Naive Bayesian, Logistic Regression, SVM, Decorate, Random Forest	Tweet	94.5%

V. DISCUSSION

Spam detection in Twitter needs different ways from traditional spam detection methods for email and the web since (1) spammers tend to use shortened URLs instead of the full form of URL, and (2) Twitter is based on a huge and detailed network which is built on tweets, accounts, lists, moments, and the relationships between them. Thus, a more robust approach is required to detect spam in Twitter to with considering the variety of legitimate users who may behave similarly to spammers under certain circumstances. Even Twitter itself has false positive (spammers which are classified as legitimate users) detections as it is reported that Twitter has recommended a legitimate user to follow bots instead of related accounts [88]. In this paper, the features of Twitter spam detection are presented with discussing their effectiveness in detecting spam. Then, the proposed works in literature are categorized into four: (1) Account-based, (2) tweet-based, (3) graph-based, and (4) hybrid spam detection methods which use a combination of others.

Methods based on account-based features analyze account by using features related with accounts which some of them can be manipulated by spammers such as the number of following, the number of tweets sent by the account, the number of lists created by the account, the number of moments created by the account which is a brand new feature and, to the best of our knowledge, it has not been used by any works in

literature yet [89–91], the number of mentions the account received, the number of likes received by the tweets of account, and the number of retweets received by the tweets of account. Similarly, the number of followers, the ratio between the number of followers over the number of following, the ratio of the number of tweets liked by others, the ratio of the number of tweets retweeted also can be slightly manipulated by using a group of bots. Bots use various tools to do automated tasks such as following a user, sending a tweet. Some works investigate a number of last tweets of an account in order to reveal if the account posts spam tweets whose contents are almost identical to the tweets recently posted which is useful to detect spam distributed by bots, a set of accounts under the command and control (C&C) infrastructure. Account-based features are lightweight enough to be used detecting real-time spam which requires instant analysis. The number of lists the user is a member of can be considered a useful metric to detect spammers since it is an obvious sign of the user's impact on others but it is open to manipulation by creating fake lists and adding the fake accounts which are under the C&C infrastructure into these lists. Account-based features are lightweight enough to be used detecting real-time spam which requires instant analysis but they can be easily manipulated by spammers [37].

Tweet-based spam detection methods use parts of a tweet such as mentions, hashtags, the number of likes the tweet received, the number of retweets the tweet received, the content of tweet, lexical analysis of the tweet, the URL of the tweet, the location of the tweet, the post date of the tweet. Since the most common way to spread spam is sharing via a malicious URL [92], URLs of tweets are needed to be inspected. Therefore, almost all Twitter spam detection methods inspect URLs of tweets. The traditional ways to filter spam are based on IP blacklisting [93], domain and URL blacklisting [94]. Since spammers tend to use shortened URLs, traditional URL or IP blacklisting methods are not able to filter malicious URLs in Twitter. Also, Grier et al. [36] show that methods based on blacklisting are too slow to protect users since there is a delay before the malicious URLs are included in the database. Similar to account-based features, tweet-based features are lightweight enough to be used detecting real-time spam which requires instant analysis.

Graph-based spam detection methods use features of relationships between the sender and the mentions of a tweet such as connectivity and distance to analyze how these accounts are connected each other and to measure strengths of their connections in order to reveal the possibility of a spam connection. Graph-based features are hard to be manipulated [21], unlike account-based and tweet-based features. However, extracting of these features require in-depth analysis on the huge and complex Twitter graph which is time and resource intensive. Therefore, unlike account-based and tweet-based features, graph-based features are not lightweight enough for real-time spam detection. Another limitation of the graph-based approaches is that they assume that tweets come from friends are benign regardless of their content [21] which is not valid when attackers steal the accounts of legitimate users for their malicious aims.

VI. Conclusion

Twitter is the most popular microblogging platform which provides easy-to-use user experience thanks to its architecture. This popularity attracts the attention of spammers who post tweets to phish legitimate users by directing them to malicious websites through the URLs shared in tweets, spread malicious software and advertises through URLs shared within tweets, aggressively follow/unfollow legitimate users and hijack trending topics to attract their attention, propagate pornography. In August of 2014, Twitter has revealed that 8.5% of its monthly active users which equals approximately 23 million users have automatically contacted their servers for regular updates. Since Twitter has unique characteristics from email services and websites, traditional spam filtering methods are not able to detect spam in Twitter. Thus, a more robust spam detection approach which is specially designed for Twitter is needed. In order to provide a spam-free environment, tweets of spammers are needed to be detected and filtered as well as the owners. By doing this, it is critical to reduce false positive detections in order to prevent legitimate users to be classified as spammers. In this paper, the features of Twitter spam detection and proposed approaches in the literature are discussed with considering their advantages and disadvantages. Also, the outdated features of Twitter which are commonly used by Twitter spam detection approaches are highlighted. Some new features of Twitter which, to the best of our knowledge, have not been mentioned by any other works are also presented.

References

[1] F. Benevenuto, G. Magno, T. Rodrigues, V. Almeida, Detecting spammers on Twitter, in: CEAS 2010 - Seventh Annu. Collab. Electron. Messag. Anti-Abuse Spam Conf., Redmond, Washington, USA, 2010: pp. 12–21. doi:10.1.1.297.5340.

[2] N.K. Alex Cheng, Mark Evans, Inside the Political Twittersphere, Sysomos. (2009). https://sysomos.com/inside-twitter/political-twittersphere (accessed February 5, 2017).

[3] A. Tumasjan, T. Sprenger, P. Sandner, I. Welpe, Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment, in: Proc. Fourth Int. AAAI Conf. Weblogs Soc. Media, Washington, DC, USA, 2010: pp. 178–185. doi:10.1074/jbc.M501708200.

[4] F. Bravo-Marquez, M. Mendoza, B. Poblete, Combining strengths, emotions and polarities for boosting Twitter sentiment analysis, in: Proc. Second Int. Work. Issues Sentim. Discov. Opin. Min. (WISDOM '13), Chicago, IL, USA, 2013: pp. 1–9. doi:10.1145/2502069.2502071.

[5] A. Pak, P. Paroubek, Twitter as a Corpus for Sentiment Analysis and Opinion Mining, Computer (Long. Beach. Calif). 2010 (2010) 1320–1326. doi:10.1371/journal.pone.0026624.

[6] Company | About, Twitter. (2017). https://about.twitter.com/company (accessed February 5, 2017).

[7] Twitter Usage Statistics - Internet Live Stats, InternetLiveStats. (2017). http://www.internetlivestats.com/twitter-statistics/ (accessed February 5, 2017).

[8] D. Sayce, Number of tweets per day?, (2016). http://www.dsayce.com/social-media/tweets-day/ (accessed February 5, 2017).

[9] M. McCord, M. Chuah, Spam Detection on Twitter Using Traditional Classifiers, in: Auton. Trust. Comput. - 8th Int. Conf. (ATC 2011), Banff, Canada, 2011: pp. 175–186. doi:10.1007/978-3-642-23496-5_13.

[10] B. Stone, Google Adds Live Updates to Results, New York Times. (2009). http://www.nytimes.com/2009/12/08/technology/companies/08google.html (accessed February 5, 2017).

[11] A. DuVander, Which APIs Are Handling Billions of Requests Per Day? | ProgrammableWeb, Program. Web. (2012). http://www.programmableweb.com/news/which-apis-are-handling-billions-requests-day/2012/05/23 (accessed February 5, 2017).

[12] L. Jiang, M. Yu, M. Zhou, X. Liu, T. Zhao, Target-dependent Twitter Sentiment Classification, in: Comput. Linguist., 2011: pp. 151–160.

[13] R.P. Schumaker, A.T. Jarmoszko, C.S. Labedz, Predicting wins and spread in the Premier League using a sentiment analysis of twitter, Decis. Support Syst. (2016). doi:10.1016/j.dss.2016.05.010.

[14] A. Go, L. Huang, R. Bhayani, Twitter Sentiment Analysis, Entropy. (2009) 17. doi:10.1007/978-3-642-35176-1_32.

[15] A. Montejo-Ráez, E. Martínez-Cámara, M.T. Martín-Valdivia, L.A. Ureña-López, Ranked WordNet graph for Sentiment Polarity Classification in Twitter, Comput. Speech Lang. 28 (2014) 93–107. doi:10.1016/j.csl.2013.04.001.

[16] S. Liu, X. Cheng, F. Li, F. Li, TASC: Topic-Adaptive Sentiment Classification on Dynamic Tweets, IEEE Trans. Knowl. Data Eng. 27 (2015) 1696–1709. doi:10.1109/TKDE.2014.2382600.

[17] E. Haddi, X. Liu, Y. Shi, The Role of Text Pre-processing in Sentiment Analysis, Procedia Comput. Sci. 17 (2013) 26–32. doi:10.1016/j.procs.2013.05.005.

[18] W. Chow, S. Shi, Investigating customers' satisfaction with brand pages in social networking sites, J. Comput. Inf. Syst. 55 (2015) 48–58.

[19] H. Saif, Y. He, M. Fernandez, H. Alani, Semantic Patterns for Sentiment Analysis of Twitter, in: Proc. 13th Int. Semant. Web Conf., Trentino, Italy, 2014: pp. 324–340.

[20] M.H.M. Sharif, I. Troshani, R. Davidson, Public Sector Adoption of Social Media, J. Comput. Inf. Syst. 55 (2015) 53–61. doi:10.1017/CBO9781107415324.004.

[21] C.D. Gowri, V. Mohanraj, A Survey on Spam Detection in Twitter: A Review, Int. J. Comput. Sci. Bus. Informatics. 14 (2014) 92–102. http://ijcsbi.org/index.php/ijcsbi/article/view/418.

[22] D. Goodin, Mystery attack drops avalanche of malicious messages on Twitter, Ars Tech. (2014). http://arstechnica.com/security/2014/04/mystery-attack-drops-avalanche-of-malicious-messages-on-twitter/ (accessed February 5, 2017).

[23] Z.M. Seward, Twitter admits that as many as 23 million of its active users are automated, Quartz. (2014). http://qz.com/248063/twitter-admits-that-as-many-as-23-million-of-its-active-users-are-actually-bots/ (accessed February 5, 2017).

[24] L. Whitney, Twitter says as many as 23 million accounts connect with automated services, CNET. (2014). https://www.cnet.com/news/twitter-reveals-23-million-of-accounts-active/ (accessed February 5, 2017).

[25] A Study of Social Network Scams, 2008.

[26] H. Drucker, D. Wu, V.N. Vapnik, Support Vector Machines for Spam Categorization, IEEE Trans. Neural Networks. 10 (1999) 1048–1054. doi:10.1109/72.788645.

[27] E. Blanzieri, A. Bryl, A Survey of Learning-Based Techniques of Email Spam Filtering, Artif. Intell. Rev. 29 (2008) 63–92.

[28] J. Song, S. Lee, J. Kim, Spam Filtering in Twitter using Sender-Receiver Relationship, in: RAID'11 Proc. 14th Int. Conf. Recent Adv. Intrusion Detect., Menlo Park, CA, USA, 2011: pp. 301–317. doi:10.1007/978-3-642-23644-0_16.

[29] A.H. Wang, Don't follow me: Spam detection in Twitter, in: SECRYPT 2010 - Proc. Int. Conf. Secur. Cryptogr., Athens, Greece, 2010: pp. 1–10. doi:978-989-8425-18-8.

[30] Reporting Spam on Twitter, Twitter. (2017). https://support.twitter.com/articles/64986 (accessed February 5, 2017).

[31] X. Zhang, S. Zhu, W. Liang, Detecting Spam and Promoting Campaigns in the Twitter Social Network, in: IEEE Int. Conf. Data Min. (ICDM 2012), IEEE, Brussels, Belgium, 2012: pp. 1194–1199. doi:10.1109/ICDM.2012.28.

[32] D. Boyd, J. Heer, Profiles as Conversation: Networked Identity Performance on Friendster, in: HICSS '06 Proc. 39th Annu. Hawaii Int. Conf. Syst. Sci., Kauai, Hawaii, USA, 2006. doi:10.1109/HICSS.2006.394.

[33] T.N. Jagatic, N.A. Johnson, M. Jakobsson, F. Menczer, Social Phishing, Commun. ACM. 50 (2007) 94–100. doi:10.1145/1290958.1290968.

[34] K. Lee, B.D. Eoff, J. Caverlee, Seven Months with the Devils: A Long-Term Study of Content Polluters on Twitter, in: Fifth Int. AAAI Conf. Weblogs Soc. Media, AAAI Press, Barcelona, Spain, 2011: pp. 185–192.

[35] C.M. Zhang, V. Paxson, Detecting and Analyzing Automated Activity on Twitter, in: PAM'11 Proc. 12th Int. Conf. Passiv. Act. Meas., Atlanta, GA, USA, 2011: pp. 102–111. doi:10.1007/978-3-642-19260-9_11.

[36] C. Grier, K. Thomas, V. Paxson, M. Zhang, @spam: The Underground on 140 Characters or Less, in: Proc. 17th ACM Conf. Comput. Commun. Secur., Chicago, IL, USA, 2010: pp. 27–37. doi:10.1145/1866307.1866311.

[37] P. Kaur, A. Singhal, J. Kaur, Spam Detection on Twitter: A Survey, in: 2016 Int. Conf. Comput. Sustain. Glob. Dev., IEEE, New Delhi, India, 2016: pp. 2570–2573.

[38] M. Brandt, 80% Of Twitter's Users Are Mobile, Statista. (2015). https://www.statista.com/chart/1520/number-of-monthly-active-twitter-users/ (accessed February 5, 2017).

[39] A.P. Felt, M. Finifter, E. Chin, S. Hanna, D. Wagner, A Survey of Mobile Malware in the Wild, in: SPSM '11 Proc. 1st ACM Work. Secur. Priv. Smartphones Mob. Devices, Chicago, IL, USA, 2011: pp. 3–14. doi:10.1145/2046614.2046618.

[40] M. Chandramohan, H.B.K. Tan, Detection of Mobile Malware in the Wild, Computer (Long. Beach. Calif). 45 (2012) 65–71. doi:10.1109/MC.2012.36.

[41] Y. Zhou, Z. Wang, W. Zhou, X. Jiang, Hey, You, Get Off of My Market: Detecting Malicious Apps in Official and Alternative Android Markets, in: Proc. 19th Annu. Netw. Distrib. Syst. Secur. Symp., San Diego, California, USA, 2012. http://www.csd.uoc.gr/~hy558/papers/mal_apps.pdf.

[42] G. Delac, M. Silic, J. Krolo, Emerging Security Threats for Mobile Platforms, in: 2011 Proc. 34th Int. Conv. MIPRO, Opatija, Croatia, 2011: pp. 1468–1473.

[43] T. Cannon, Android Data Stealing Vulnerability, (2010). https://thomascannon.net/blog/2010/11/android-data-stealing-vulnerability/ (accessed February 5, 2017).

[44] X. Wei, L. Gomez, I. Neamtiu, M. Faloutsos, Malicious Android Applications in the Enterprise: What Do They Do and How Do We Fix It?, in: ICDEW '12 Proc. 2012 IEEE 28th Int. Conf. Data Eng. Work., IEEE, Arlington, Virginia, USA, 2012: pp. 251–254.

[45] A.T. Kabakus, I.A. Dogru, A. Cetin, APK Auditor : Permission-based Android malware detection system, Digit. Investig. 13 (2015) 1–14. doi:10.1016/j.diin.2015.01.001.

[46] L. Bilge, T. Strufe, D. Balzarotti, E. Kirda, S. Antipolis, All Your Contacts Are Belong to Us: Automated Identity Theft Attacks on Social Networks, in: 18th Int. World Wide Web Conf. (WWW '09), Madrid, Spain, 2009: pp. 551–560. doi:http://doi.acm.org/10.1145/1526709.1526784.

[47] D.K. McGrath, M. Gupta, Behind Phishing: An Examination of Phisher Modi Operandi, in: LEET'08 Proc. 1st Usenix Work. Large-Scale Exploit. Emergent Threat., San Francisco, CA, USA, 2008: p. 4. http://portal.acm.org/citation.cfm?id=1387713.

[48] Z. Chu, S. Gianvecchio, H. Wang, S. Jajodia, Who is Tweeting on Twitter: Human, Bot, or Cyborg?, in: 26th Annu. Comput. Secur. Appl. Conf. (ACSAC 2010), Austin, Texas, USA, 2010: pp. 21–30. doi:10.1145/1920261.1920265.

[49] S. Chhabra, A. Aggarwal, F. Benevenuto, P. Kumaraguru, Phi.sh/$oCiaL: The phishing landscape through short URLs, in: 8th Annu. Collab. Electron. Messag. Anti-Abuse Spam Conf. (CEAS 2011), Perth, Australia, 2011.

[50] F. Klien, M. Strohmaier, Short Links Under Attack: Geographical Analysis of Spam in a URL Shortener Network, in: 23th ACM Conf. Hypertext Soc. Media (HT 2012), Milwaukee, WI, USA, 2012: pp. 83–87. doi:10.1145/2309996.2310010.

[51] G. Stringhini, C. Kruegel, G. Vigna, Detecting Spammers on Social Networks, in: 26th Annu. Comput. Secur. Appl. Conf. (ACSAC 2010), Austin, Texas, USA, 2010: pp. 1–9.

[52] D. Antoniades, E. Athanasopoulos, T. Karagiannis, we.b: The web of short URLs, in: WWW '11 Proc. 20th Int. Conf. World Wide Web, Hyderabad, India, 2011: pp. 715–724. doi:10.1145/1963405.1963505.

[53] D. Kim, Y. Jo, I.-C. Moon, A. Oh, Analysis of Twitter Lists as a Potential Source for Discovering Latent Characteristics of Users, in: CHI 2010 Work. Microblogging What How Can We Learn From It, Atlanta, Georgia, USA, 2010. doi:10.1.1.163.7391.

[54] Y. Yamaguchi, T. Amagasa, H. Kitagawa, Tag-based User Topic Discovery Using Twitter Lists, in: 2011 Int. Conf. Adv. Soc. Networks Anal. Min. (ASONAM 2011), Kaohsiung, Taiwan, 2011: pp. 13–20. doi:10.1109/ASONAM.2011.58.

[55] Using Twitter lists, Twitter. (2017). https://support.twitter.com/articles/76460 (accessed February 5, 2017).

[56] C. Chen, J. Zhang, Y. Xie, Y. Xiang, W. Zhou, M.M. Hassan, A. AlElaiwi, M. Alrubaian, A Performance Evaluation of Machine Learning-Based Streaming Spam Tweets Detection, IEEE Trans. Comput. Soc. Syst. 2 (2016) 65–76. doi:10.1109/TCSS.2016.2516039.

[57] M. Verma, S. Sofat, Techniques to Detect Spammers in Twitter - A Survey, Int. J. Comput. Appl. 85 (2014) 27–32. doi:10.5120/14877-3279.

[58] G. Gee, T. Hakson, Twitter Spammer Profile Detection, Stanford, California, USA, 2010. http://cs229.stanford.edu/proj2010/GeeTeh-TwitterSpammerProfileDetection.pdf.

[59] The Twitter Rules, Twitter. (2017). https://support.twitter.com/articles/18311 (accessed February 5, 2017).

[60] C. Yang, R. Harkreader, G. Gu, Empirical Evaluation and New Design for Fighting Evolving Twitter Spammers, IEEE Trans. Inf. Forensics Secur. 8 (2013) 1280–1293. doi:10.1109/TIFS.2013.2267732.

[61] K. Lee, J. Caverlee, S. Webb, Uncovering Social Spammers: Social Honeypots + Machine Learning, in: Proc. 33rd Annu. Int. ACM SIGIR Conf. Res. Dev. Inf. Retr., Geneva, Switzerland, 2010: pp. 435–442. doi:10.1145/1835449.1835522.

[62] P.-C. Lin, P.-M. Huang, A Study of Effective Features for Detecting Long-surviving Twitter Spam Accounts, in: 2013 15th Int. Conf. Adv. Commun. Technol., PyeongChang, Korea, 2013: pp. 841–846. http://ieeexplore.ieee.org/xpl/articleDetails.jsp?tp=&arnumber=6488315&matchBoolean=true&rowsPerPage=30&searchField=Search_All&queryText=(%22twitter+spam%22)%5Cnpapers3://publication/uuid/60707410-4AE4-4FBE-A667-C91C41C51802.

[63] J. Ma, L.K. Saul, S. Savage, G.M. Voelker, Beyond blacklists: learning to detect malicious web sites from suspicious URLs, in: KDD '09 Proc. 15th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., Paris, France, 2009: pp. 1245–1253. doi:10.1145/1557019.1557153.

[64] D. Canali, M. Cova, G. Vigna, C. Kruegel, Prophiler: A Fast Filter for the Large-Scale Detection of Malicious Web Pages, in: WWW '11 Proc. 20th Int. Conf. World Wide Web, Hyderabad, India, 2011: pp. 197–206. doi:10.1145/1963405.1963436.

[65] Developer Survey Results 2016, 2016. http://stackoverflow.com/research/developer-survey-2016 (accessed February 5, 2017).

[66] D. Rowinski, It's Official: JavaScript Is The Most Commonly Used Programming Language On Earth, Appl. Resour. Cent. from Applause. (2016). https://arc.applause.com/2016/03/22/javascript-is-the-worlds-dominant-programming-language/ (accessed February 5, 2017).

[67] C. Whittaker, B. Ryner, M. Nazif, Large-Scale Automatic Classification of Phishing Pages, in: 17th Annu. Netw. Distrib. Syst. Secur. Symp. (NDSS '10), San Diego, California, USA, 2010. http://www.isoc.org/isoc/conferences/ndss/10/pdf/08.pdf%5Cnhttp://research.google.com/pubs/pub35580.html.

[68] Y. Wang, D. Beck, X. Jiang, R. Roussev, Automated Web Patrol with Strider HoneyMonkeys: Finding Web Sites That Exploit Browser Vulnerabilities, in: 13th Annu. Netw. Distrib. Syst. Secur. Symp. (NDSS '06), San Diego, California, USA, 2005: pp. 1–11.

[69] K. Thomas, C. Grier, J. Ma, V. Paxson, D. Song, Design and evaluation of a real-time URL spam filtering service, in: SP '11 Proc. 2011 IEEE Symp. Secur. Priv., Oakland, California, USA, 2011: pp. 447–462. doi:10.1109/SP.2011.25.

[70] M. Cova, C. Kruegel, G. Vigna, Detection and Analysis of Drive-by-Download Attacks and Malicious JavaScript Code, in: WWW '10 Proc. 19th Int. Conf. World Wide Web, Raleigh, North Carolina, USA, 2010: pp. 281–290.

[71] S. Lee, J. Kim, WarningBird: A Near Real-Time Detection System for Suspicious URLs in Twitter Stream, IEEE Trans. Dependable Secur. Comput. 10 (2013) 183–195. doi:10.1109/TDSC.2013.3.

[72] J. Martinez-Romo, L. Araujo, Detecting malicious tweets in trending topics using a statistical analysis of language, Expert Syst. Appl. 40 (2013) 2992–3000. doi:10.1016/j.eswa.2012.12.015.

[73] R. Angles, C. Gutierrez, Survey of graph database models, ACM Comput. Surv. 40 (2008) 1–39. doi:10.1145/1322432.1322433.

[74] J. Ugander, B. Karrer, L. Backstrom, C. Marlow, P. Alto, The Anatomy of the Facebook Social Graph, Arxiv Prepr. arXiv. abs/1111.4 (2011) 1–17. doi:10.1.1.31.1768.

[75] J. Weaver, P. Tarjan, Facebook Linked Data via the Graph API, Semant. Web. 4 (2013) 245–250. doi:10.3233/SW-2012-0078.

[76] B. Krishnamurthy, P. Gill, M. Arlitt, A few chirps about twitter, in: Proc. 1st Work. Online Soc. Networks, Seattle, WA, USA, 2008: pp. 19–24. doi:10.1145/1397735.1397741.

[77] S. Myers, A. Sharma, P. Gupta, J. Lin, Information Network or Social Network? The Structure of the Twitter Follow Graph, in: WWW'14 Companion Proc. 23rd Int. Conf. World Wide Web, Seoul, Korea, 2014: pp. 493–498. doi:10.1145/2567948.2576939.

[78] M. Gabielkov, A. Legout, The Complete Picture Of the Twitter Social Graph, in: Conex. Student '12 Proc. 2012 ACM Conf. Conex. Student Work., Nice, France, 2012: pp. 20–21. doi:10.1145/2413247.2413260.

[79] L. Zou, L. Chen, J.X. Yu, Y. Lu, A novel spectral coding in a large graph database, in: EDBT '08 Proc. 11th Int. Conf. Extending Database Technol. Adv. Database Technol., Nantes, France, 2008: pp. 181–192. doi:10.1145/1353343.1353369.

[80] R. a Hanneman, M. Riddle, Introduction to Social Network Methods, University of California Press, Riverside, CA, USA, 2005. doi:10.1016/j.socnet.2006.08.002.

[81] U. Brandes, T. Erlebach, Network Analysis, Springer Berlin Heidelberg, Heidelberg, Germany, 2005. doi:10.1007/b106453.

[82] H. Gao, Y. Chen, K. Lee, D. Palsetia, A. Choudhary, Towards Online Spam Filtering in Social Networks, in: 19th Annu. Netw. Distrib. Syst. Secur. Symp. (NDSS 2012), San Diego, California, USA, 2012.

[83] C. Chen, J. Zhang, X. Chen, Y. Xiang, W. Zhou, 6 Million Spam Tweets: A Large Ground Truth for Timely Twitter Spam Detection, in: 2015 IEEE Int. Conf. Commun., IEEE, London, UK, 2015: pp. 7065–7070. doi:10.1109/ICC.2015.7249453.

[84] C. Yang, R.C. Harkreader, G. Gu, Die Free or Live Hard? Empirical Evaluation and New Design for Fighting Evolving Twitter Spammers, in: RAID'11 Proc. 14th Int. Conf. Recent Adv. Intrusion Detect., Menlo Park, CA, USA, 2011: pp. 318–337. doi:10.1007/978-3-642-23644-0_17.

[85] A.A. Amleshwaram, N. Reddy, S. Yadav, G. Gu, C. Yang, CATS: Characterizing Automation of Twitter Spammers, in: 2013 5th Int. Conf. Commun. Syst. Networks (COMSNETS 2013), Bangalore, India, 2013. doi:10.1109/COMSNETS.2013.6465541.

[86] A. Chakraborty, J. Sundi, S. Satapathy, SPAM: A Framework for Social Profile Abuse Monitoring, in: CEAS '11 8th Annu. Collab. Electron. Messag. Anti-Abuse Spam Conf., Perth, Australia, 2011: pp. 46–54. http://www.cs.sunysb.edu/~aychakrabort/courses/cse508/report.pdf.

[87] B. Wang, A. Zubiaga, M. Liakata, R. Procter, Making the Most of Tweet-Inherent Features for Social Spam Detection on Twitter, in: Proc. 5th Work. Mak. Sense Microposts, Florence, Italy, 2015: pp. 10–16.

[88] D. Hernandez, Why can't Twitter kill its bots?, Fusion. (2015). http://fusion.net/story/195901/twitter-bots-spam-detection/ (accessed February 5, 2017).

[89] A. Read, Everyone Can Now Create Twitter Moments: Here's All You Need to Know, Buffer. (2016). https://blog.bufferapp.com/twitter-moments (accessed February 5, 2017).

[90] J. Roettgers, Twitter Now Lets Everyone Create and Share Moments, Variety. (2016). http://variety.com/2016/digital/news/twitter-moments-1201872731/ (accessed February 5, 2017).

[91] T. Huddleston, Now Twitter Wants You to Create Your Own "Moments," Fortune. (2016). http://fortune.com/2016/09/28/twitter-create-your-own-moments/ (accessed February 5, 2017).

[92] M. Egele, G. Stringhini, C. Kruegel, G. Vigna, COMPA: Detecting Compromised Accounts on Social Networks, in: 20th Annu. Netw. Distrib. Syst. Secur. Symp. (NDSS 2013), San Diego, California, USA, 2013.

[93] Z. Qian, Z.M. Mao, Y. Xie, F. Yu, On Network-level Clusters for Spam Detection, in: 17th Annu. Netw. Distrib. Syst. Secur. Symp. (NDSS '10), San Diego, California, USA, 2010.

[94] Y. Xie, F. Yu, R. Panigrahy, Spamming Botnet: Signatures and Characteristics, in: ACM SIGCOMM 2008, Seattle, WA, USA, 2008.

Prediction Method for Large Diatom Appearance with Meteorological Data and MODIS Derived Turbidity and Chlorophyll-A in Ariake Bay Area in Japan

Kohei Arai [1]
1 Graduate School of Science and Engineering
Saga University
Saga City, Japan

***Abstract*—Prediction method for large diatom appearance in winter with meteorological data and MODIS derived turbidity and chlorophyll-a in Ariake Bay Area in Japan is proposed. Mechanism for large diatom appearance in winter is discussed with the influencing factors, meteorological condition and in-situ data of turbidity, chlorophyll-a data with the measuring instruments equipped at the Saga University own Tower in the Ariake Bay area. Particularly, the method for estimation of turbidity is still under discussion. Therefore, the algorithm for estimation of turbidity with MODIS data is proposed here. Through experiments, it is found that the proposed prediction method for large diatom appearance is validated with the meteorological data and MODIS derived turbidity as well as chlorophyll-a data estimated for the winter (from January to March) in 2012 and 2015.**

Keywords*—chlorophyl-a concentration; red tide; diatom; MODIS; satellite remote sensing

I. INTRODUCTION

The Ariake Sea is the largest productive area of Nori (*Porphyra yezoensis*[1]) in Japan. In winters in 2012, 2013, 2014 and 2015, as well as 2016, a massive diatom bloom appeared in the Ariake Bay, Japan [1]. In case of above red tides, bloom causative was *Rhizosolenia imbricate*[2] and *Eucampia zodiacus*[3]. This bloom has been occurred several coastal areas in Japan and is well reported by Nishikawa et al. for Harima-nada sea areas [2]-[10]. Diatom blooms have recurrently appeared from late autumn to early spring in the coastal waters of western Japan, such as the Ariake Bay [11] and the Seto Inland Sea [12], where large scale "Nori" aquaculture occurs. Diatom blooms have caused the exhaustion of nutrients in the water column during the "Nori" harvest season. The resultant lack of nutrients has suppressed the growth of "Nori" and lowered the quality of "Nori" products due to bleaching with the damage of the order of billions of Japanese yen [3].

The chlorophyll-a concentration algorithm developed for MODIS[4] has been validated [13]. The algorithm is applied to MODIS data for a trend analysis of chlorophyll-a distribution in the Ariake Bay in winter during from 2010 to 2015 is made [13]. Also, locality of red tide appearance in Ariake Sea (Ariake Bay), Isahaya Bay and Kumamoto offshore is clarified by using MODIS data derived chlorophyll-a concentration [14]. On the other hand, red tide appearance (location, red tide species, the number of cells in unit water volume by using microscopy) are measured from the research vessel of the Saga Prefectural Fishery Promotion Center: SPFPC by once 10 days. The location and size of the red tide appearance together with the red tide source would be clarified by using SPFPC data. Match-up data of MODIS derived chlorophyll-a concentration is used for investigation of relations between MODIS data and truth data of the red tide appearance. Through time series data analysis of MODIS derived chlorophyll-a concentration, one of the possible causes of diatom appearance is clarified with the evidence of Research Bessel observations. Time series data analysis is made for large size diatom appearance events and meteorological data as well as MODIS derived Photosynthetically Available Radiance: PAR[5], Chlorophyll-a concentration. The results from the time series analysis say that large size of diatoms appear after a long period time of relatively small size of red tide appearance [15]. Also, it depends on the weather conditions and tidal effect as well as water current in the bay area. Also, a relation between large sized diatom appearance and meteorological data in Ariake Bay Area in Japan in the winter in 2016 is discussed [16].

In this paper, a prediction method for large diatom appearance in winter with meteorological data and MODIS derived turbidity and chlorophyll-a in Ariake Bay Area in Japan is proposed. Although the mechanism for large diatom appearance in winter is discussed in the previous papers [1]-[12], there is no paper which deals with a prediction method for diatom appearance so far. This paper proposes a prediction method through discussions of mechanism for diatom appearance with the influencing factors, meteorological condition and in-situ data of turbidity, chlorophyll-a data with the measuring instruments equipped at the Saga University own Tower[6] in the Ariake Bay area. There is no reliable

[1] http://en.wikipedia.org/wiki/Porphyra
[2] https://microbewiki.kenyon.edu/index.php/Rhizosolenia
[3] http://www.eos.ubc.ca/research/phytoplankton/diatoms/centric/eucampia/e_zodiacus.html
[4] http://modis.gsfc.nasa.gov/
[5] https://modis.gsfc.nasa.gov/data/dataprod/dataproducts.php?MOD_NUMBER=22
[6] http://www.ilt.saga-u.ac.jp/aripro/tower/index.html

estimation method for turbidity with MODIS data [17]. Therefore, the algorithm for estimation of turbidity with MODIS data is also proposed here.

In the next section, the proposed prediction method and procedure is described followed by experimental data and prediction results. Then conclusion is described with some discussions.

II. Proposed Method and Experiments

A. Intensive Study Areas

Fig.1 shows the intensive study areas of Ariake Bay, Kyushu, Japan.

Fig. 1. Intensive study areas

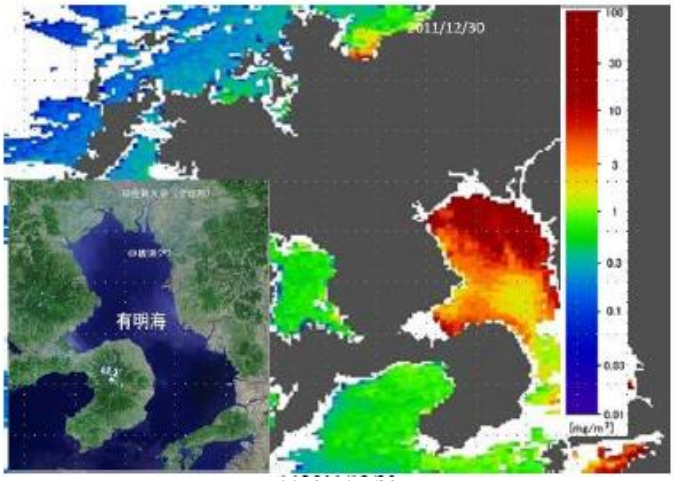
Ariake Bay is a portion of Ariake Sea of which the width is around 20km (in direction of east to west) and the length is approximately 100km (in direction of north to south). It is almost closed sea area because the mouth of Ariake Sea is quite narrow. Sea water exchanges are, therefore, very small.

B. Mechanism for Diatom Appearance in Winter

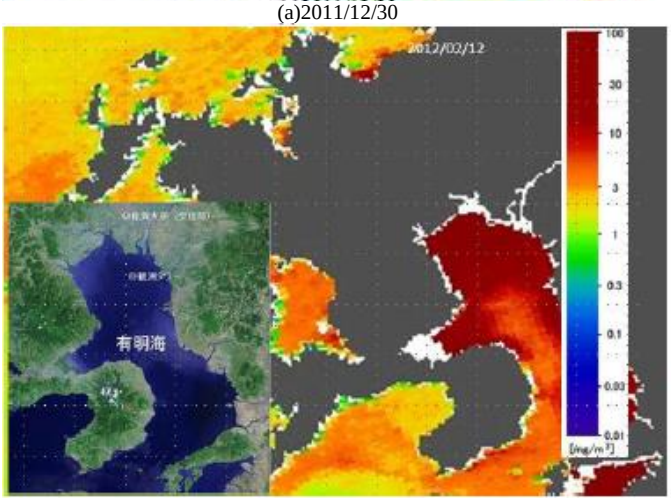
Examples of the MODIS derived chlorophyll-a concentration which are acquired in the late of December 2011 to the begging of April 2012 are shown in Fig.2. The MODIS clear scenes of data acquisition dates for the period from December 2011 to April 2012 are followings,

(1) 2011/12/5, 13, 19, 25, 27, 30
(2) 2012/1/7, 12, 17, 21, 23, 26, 30, 31
(3) 2012/2/4, 11, 12, 20, 24, 27
(4) 2012/3/12, 14, 21, 25, 26, 27, 29
(5) 2012/4/1, 7, 12, 15, 17, 22, 26, 27, 28

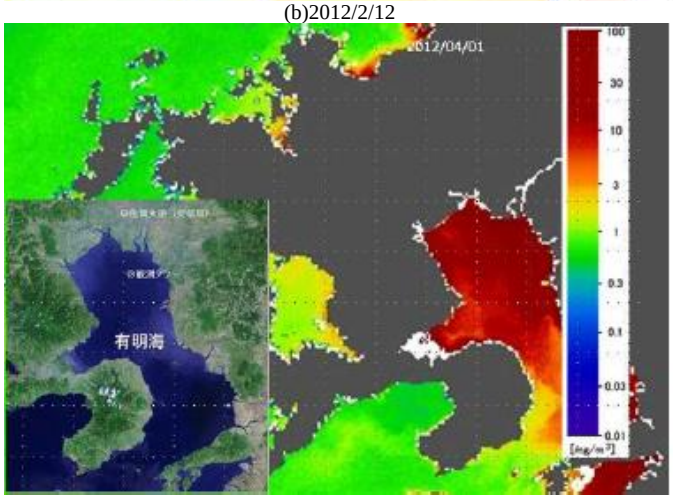
The examples of the chlorophyll-concentrations for 2011/12/30, 2012/2/12, 2012/4/1 are just example of sparsely distributed chlorophyll-a concentration on December 30 in 2011 and examples of densely distributed chlorophyll-a concentration on February 12 in 2012 and April 1 2012.



(a)2011/12/30



(b)2012/2/12

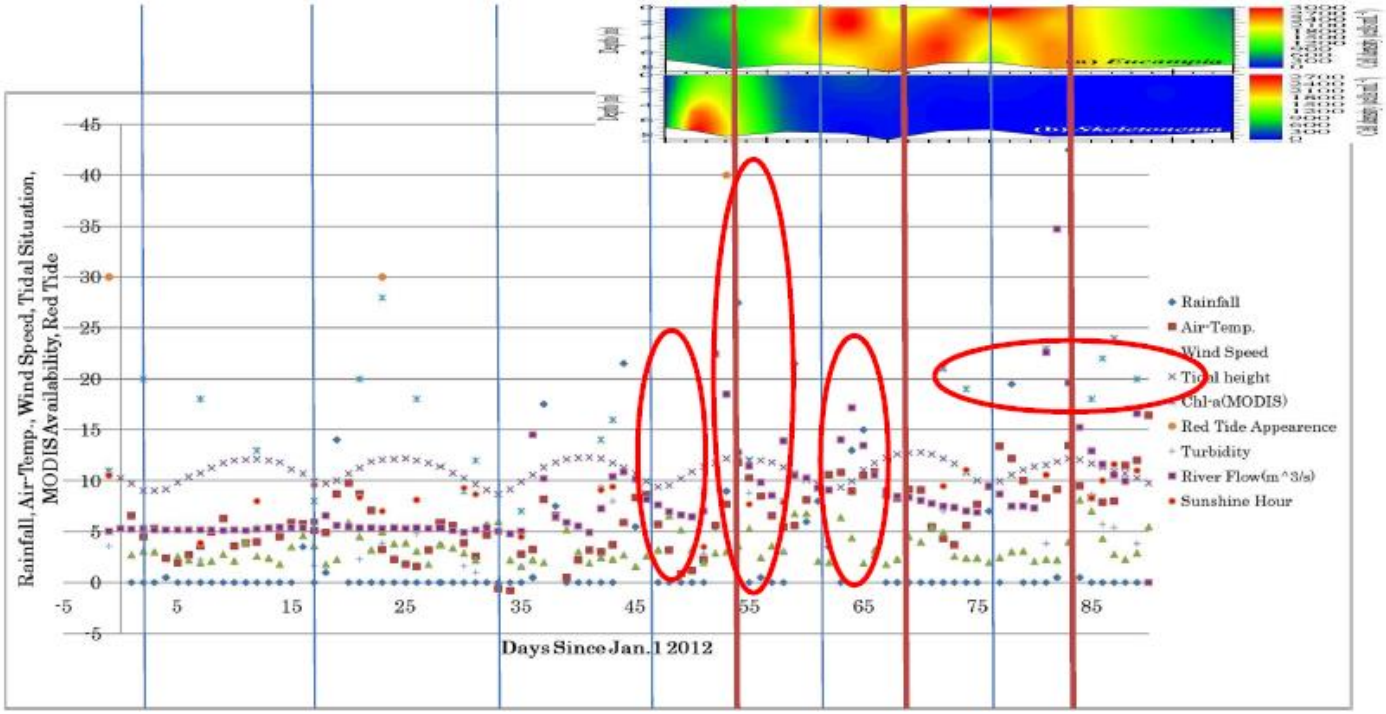


(c)2012/4/1

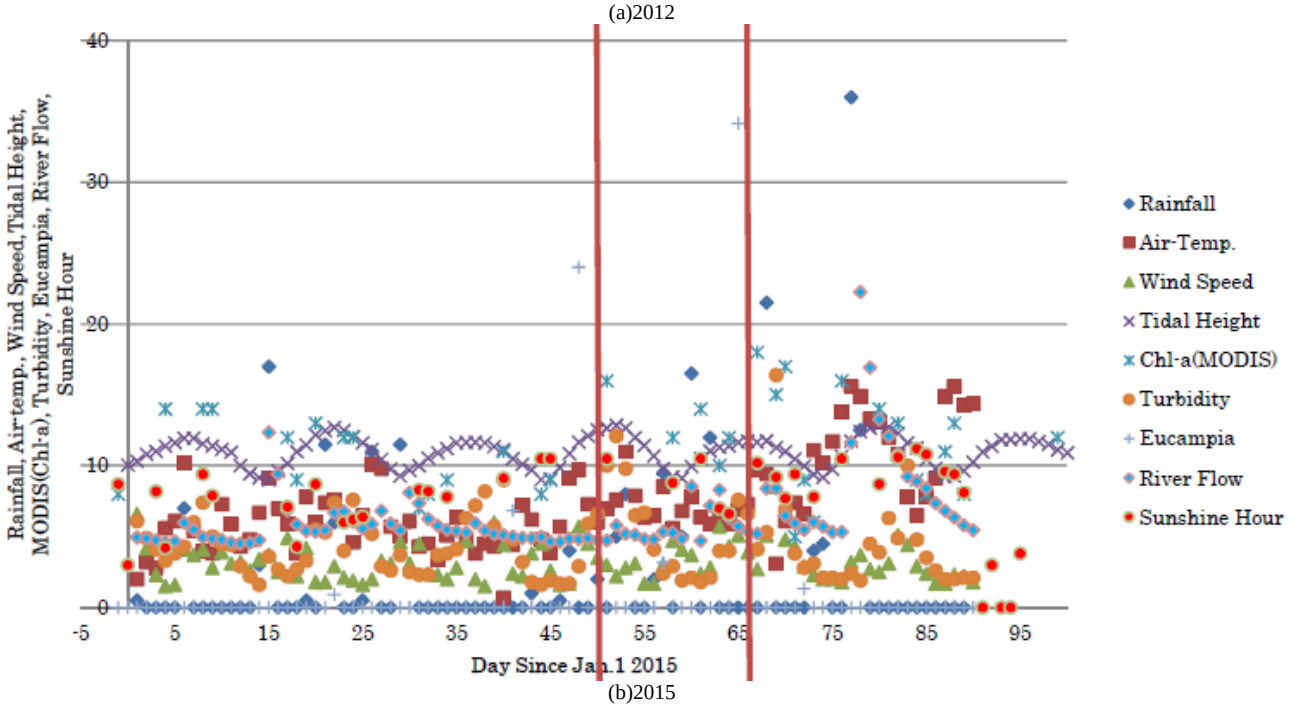
Fig. 2. MODIS derived chlorophyll-a concentration in the winter in 2012

Relatively small size diatoms appear at western side of Ariake bay area in the middle of February and then large size diatoms appear in the whole area of Ariake bay area from 25 February to the middle of April 2012. The influencing factors, meteorological condition, turbidity, chlorophyll-a, river water flow, tidal height is collected from the Japanese Meteorological Agency: JMA, MODIS data and diatom

appearance. These data of 2012 are plotted in Fig.3 (a). In the figure, *Eucampia zodiacus* (top) and *Skeletonema spp.* (bottom) appearances which are reported by the Ito et al. [1] are also plotted in the top right corner of the figure. These are time series of the depth distributions of diatom for the period



(a)2012



(b)2015

Fig. 3. Measured meteorological data and tidal height as well as river flow data together with MODIS derived turbidity and chlorophyll-a data

Fig.3 (b) shows the influencing factors and diatom appearance in 2015. The MODIS clear scenes of data acquisition dates for the period from the late of December 2014 to the end of April 2015 are followings,

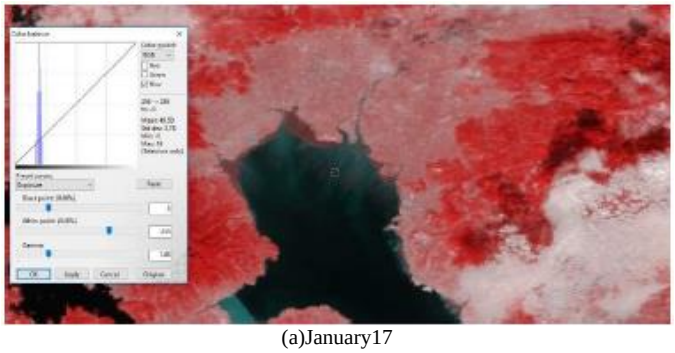
(1) 2014/12/26, 30
(2) 2015/1/7, 8, 17, 18, 20, 24, 31
(3) 2015/2/1, 3, 9, 13, 14, 27
(4) 2015/3/2, 4, 8, 11, 12, 14, 17, 21, 23, 25, 26, 28, 29, 30
(5) 2015/4/15, 17, 21, 23, 25, 26

As a matter of fact, diatom needs nutrients, sunshine, appropriate sea temperature (22 to 26 degree Celsius) and salinity (15 to 28 ‰), as well as diatom seeds. Nutrients are

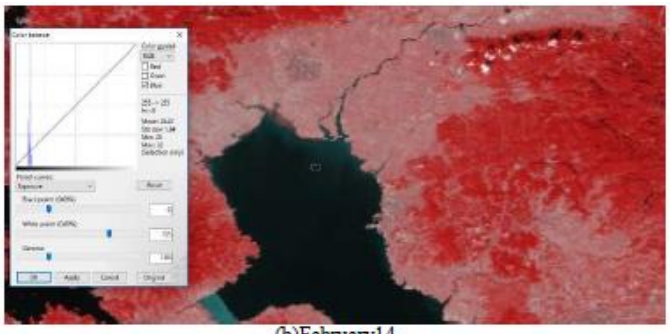
provided by river water (source of nutrients) which is mainly caused by rainfall and run-off water. Therefore, river water flow is a key component for nutrients. Relatively large diatom (*Eucampia zodiacus*) seeds are situated in the bottom layer situated in Ariake bay while relatively small diatom seeds are situated from the sensory ranges of the specific rivers, Shiota-River for *Skeletonema spp.* and *Asteroplanus karianus.* Therefore, convection or vertical mixing in the sea water of Arirake bay is a key for the large diatom appearance at the sea surface. The convection is usually occurred due to spring tide or strong winds from the north. Therefore, diatom bloom is used to be occurred in spring tide. Also, diatom seeds need sunshine, nutrients for blooming. Therefore, diatom bloom occurs after a relatively large river water flow followed by relatively small turbidity and sunshine as well as spring tide. These are mechanism for diatom appearance and bloom.

C. MODIS Data Derived Turbidity in Particular in 2015

As is mentioned before, method for estimation of turbidity with MODIS data is still under discussion. To create a method for turbidity estimation, a regressive analysis is conducted with MODIS 547nm band data and in-situ data of the measured turbidity at the Saga University Tower. Fig.4 shows examples of MODIS imagery data (False color representation) acquired on (a) January 17, (b) February 14 and (c) March 30 in 2015. In these figures, histogram of the white square sea area (Saga University Tower is situated) of MODIS 547nm band data (colored in blue) is shown. 29 of in-situ data of turbidity are used for the regressive analysis. The result is shown in Fig.5.



(a)January17



(b)February14

(c)March30

Fig. 4. False color representation of MODIS imagery which includes MODIS 547nm band data as blue color component in the false color component

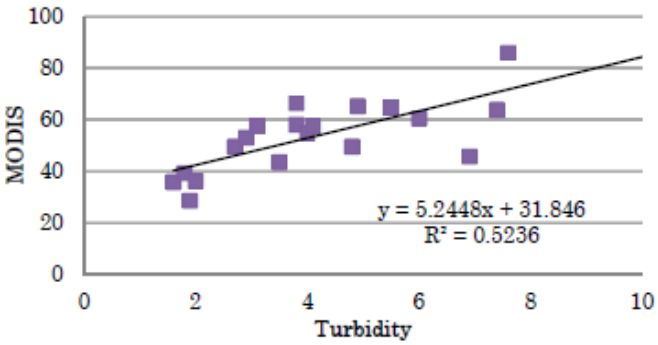


Fig. 5. Relation between in-situ data of turbidity measured at the Saga University observation tower and the estimated turbidity derived from MODIS 547nm band of data

As the result, it is found that the turbidity can be estimated with the following regressive equation,

$$\text{Turbidity} = 5.2448\text{MODIS}_{547\text{nm}} + 31.846 \quad (1)$$

In the regressive analysis, it is confirmed that $R^2 = 0.5236$ or 0.7236 of correlation coefficient.

D. Warning or Caution for Diatom Appearance

Warning or caution of diatom appearance can be provided to fishermen (Nori farmers) in accordance with the following scenario,

1) Check river water inflow during tidal wave

2) If the river water inflow is greater than the previously set threshold, then check the MODIS derived turbidity and the sunshine hour as well as the MODIS derived chlorophyll-a concentration

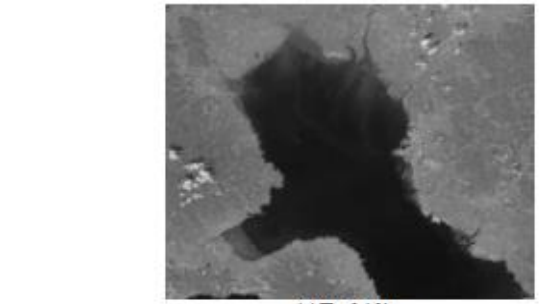
3) If not, then go back to the process (1)

4) And if the MODIS derived turbidity is less than the previously set threshold and sunshine hour and the MODIS derived chlorophyll-a concentration is greater than the previously set threshold, then warning or caution is provided for the next sprig tide

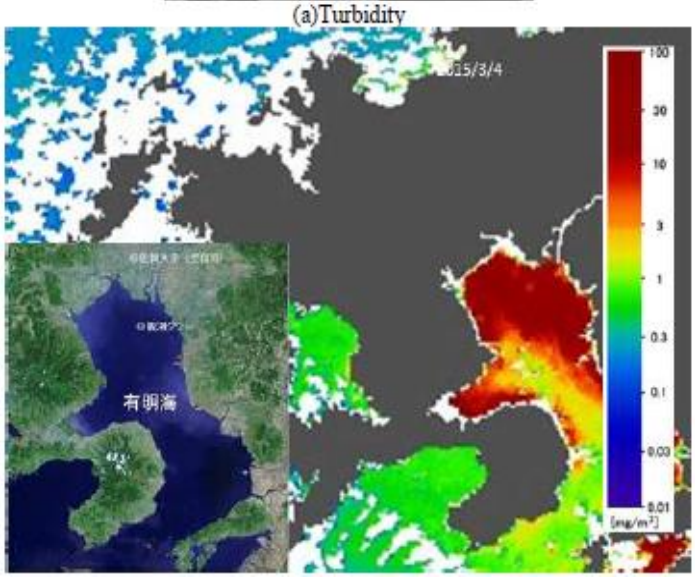
5) If not, then go back to the process (1).

E. Spatial Distribution of Diatom Appearance

Spatial distribution of diatom appearance can be estimated with turbidity distribution in the tidal wave and chlorophyll-a concentration distribution after the tidal wave. Fig.6 shows an example of the MODIS derived turbidity distribution in the tidal wave and the MODIS derived chlorophyll-a concentration distribution just after the tidal wave. The areas of which the turbidity is less than the threshold and the chlorophyll-a concentration is greater than the threshold is determined as warning areas of diatom appearance.



(a)Turbidity



(b)Chlorophyll-a concentration

Fig. 6. Example of the MODIS derived turbidity distribution in the tidal wave and the MODIS derived chlorophyll-a concentration distribution just after the tidal wave

III. CONCLUSION

Prediction method for large diatom appearance in winter with meteorological data and MODIS derived turbidity and chlorophyll-a in Ariake Bay Area in Japan is proposed. Because the proposed large diatom prediction method is totally new, there is no comparison between the proposed method and the other method at all. Mechanism for large diatom appearance in winter is discussed with the influencing factors, meteorological condition and in-situ data of turbidity, chlorophyll-a data with the measuring instruments equipped at the Saga University own Tower in the Ariake Bay area. The discussed mechanism here is almost same as the previous discussions [1]-[12]. The method for estimation of turbidity is proposed here. Estimation accuracy of turbidity with MODIS data depends on ocean areas, seasons, etc. 0.7236 of the correlation coefficient of the regression analysis is not so bad in comparison to the other trials (around 0.7 of correlation coefficients).

Through experiments, it is found that the proposed prediction method for large diatom appearance is validated with the meteorological data and MODIS derived turbidity and chlorophyll-a data estimated for the winter (from January to March) in 2012 and 2015.

Further investigations are required for confirmation of the mechanism of relatively large diatom appearance with not only 2012 and 2015 but also other years, 2010, 2011, 2013, 2014, and 2016 as well as 2017. Also, it is required to clarify the mechanism of relatively small size diatom appearance.

ACKNOWLEDGMENT

The author would like to thank Dr. Toshiya Katano of Tokyo University of Marine Science and Technology, Dr. Joji Ishizaka of Nagoya University, Dr. Minoru Wada of Nagasaki University, Dr. Yuichi Hayami, Dr. Kei Kimura, Dr. Kenji Yoshino, Dr. Naoki Fujii of Institute of Lowland and Marine Research, Saga University and Dr. Takaharu Hamada of the University of Tokyo for their great supports through the experiments.

REFERENCES

[1] Ito Y., Katano T., Fujii N., Koriyama M., Yoshino K., Hayami Y., Decreases in turbidity during neap tides initiate late winter large diatom blooms in a macrotidal embayment, Journal of Oceanography,69: 467-479. 2013.

[2] Nishikawa T, Effects of temperature, salinity and irradiance on the growth of the diatom Eucampia zodiacus caused bleaching seaweed Porphyra isolated from Harima-Nada, Seto Inland Sea, Japan. Nippon Suisan Gakk 68: 356–361. (in Japanese with English abstract), 2002.

[3] Nishikawa T, Occurrence of diatom blooms and damage tocultured Porphyra thalli by bleaching. Aquabiology 172: 405–410. (in Japanese with English abstract), 2007.

[4] Nishikawa T, Hori Y., Effects of nitrogen, phosphorus and silicon on the growth of the diatom Eucampia zodiacus caused bleaching of seaweed Porphyra isolated from Harima-Nada, Seto Inland Sea, Japan. Nippon Suisan Gakk 70: 31–38. (in Japanese with English abstract), 2004.

[5] Nishikawa T, Hori Y, Nagai S, Miyahara K, Nakamura Y, Harada K, Tanda M, Manabe T, Tada K, Nutrient and phytoplankton dynamics in Harima-Nada, eastern Seto Inland Sea, Japan during a 35- year period from 1973 to 2007. Estuaries Coasts 33: 417–427, 2010.

[6] Nishikawa T, Hori Y, Tanida K, Imai I, Population dynamics of the harmful diatom Eucampia zodiacus Ehrenberg causing bleachings of Porphyra thalli in aquaculture in Harima- Nada, the Seto Inland Sea, Japan. Harmful algae 6: 763–773, 2007.

[7] Nishikawa T, Miyahara K, Nagai S., Effects of temperature and salinity on the growth of the giant diatom Coscinodiscus wailesii isolated from Harima-Nada, Seto Inland Sea, Japan. Nippon Suisan Gakk 66: 993–998. (in Japanese with English abstract), 2000.

[8] Nishikawa T, Tarutani K, Yamamoto T., Nitrate and phosphate uptake kinetics of the harmful diatom Eucampia zodiacus Ehrenberg, a causative organism in the bleaching of aquacultured Porphyra thalii. Harmful algae 8: 513–517, 2009.

[9] Nishikawa T, Yamaguchi M., Effect of temperature on lightlimited growth of the harmful diatom Eucampia zodiacus Ehrenberg, a causative organism in the discoloration of Porphyra thalli. Harmful Algae 5: 141–147, 2006.

[10] Nishikawa T, Yamaguchi M., Effect of temperature on lightlimited growth of the harmful diatom Coscinodiscus wailesii, a causative organism in the bleaching of aquacultured Porphyra thalli. Harmful Algae 7: 561–566, 2008.

[11] Syutou T, Matsubara T, Kuno K., Nutrient state and nori aquaculture in Ariake Bay. Aquabiology 181: 168–170. (in Japanese with English abstract), 2009.

[12] Harada K, Hori Y, Nishikawa T, Fujiwara T., Relationship between cultured Porphyra and nutrients in Harima-Nada, eastern part of the Seto Inland Sea. Aquabiology 181: 146–149. (in Japanese with English abstract), 2009.

[13] Arai K., T. Katano, Trend analysis of relatively large diatoms which appear in the intensive study area of the ARIAKE Sea, Japan, in winter (2011-2015) based on remote sensing satellite data, Internationa Journal of Advanced Research in Artificial Intelligence (IJARAI), 4, 7, 15-20, 2015.

[14] Arai, K., Locality of Chlorophyll-a Concentration in the Intensive Study Area of the Ariake Sea, Japan in Winter Seasons Based on Remote Sensing Satellite Data, Internationa Journal of Advanced Research in Artificial Intelligence (IJARAI), 4, 8, 18-25, 2015.

[15] Kohei Arai, One of the Possible Causes for Diatom Appearance in Ariake Bay Area in Japan in the Winter from 2010 to 2015 (Clarified with AQUA/MODIS), International Journal of Advanced Research on Artificial Intelligence, 5, 4, 1-8, 2016.

[16] Kohei Arai, Relation between Large Sized Diatom Appearance and Meteorological Data in Ariake Bay Area in Japan, in Particular, in the Winter in 2016, International Journal of Engineering Science and Research Technology, 2, 2, 1-9, 2016

[17] Ng H.G., Matja fri M.Z., Abdullah K., Alias A.N., Comparison of turbidity measurements by MOIS and AVHRR images, Proceeding of the CGIV '08 Proceedings of the 2008 Fifth International Conference on Computer Graphics, Imaging and Visualisation, 398-403, 2008.

AUTHORS PROFILE

Kohei Arai, He received BS, MS and PhD degrees in 1972, 1974 and 1982, respectively. He was with The Institute for Industrial Science and Technology of the University of Tokyo from April 1974 to December 1978 also was with National Space Development Agency of Japan from January, 1979 to March, 1990. During from 1985 to 1987, he was with Canada Centre for Remote Sensing as a Post Doctoral Fellow of National Science and Engineering Research Council of Canada. He moved to Saga University as a Professor in Department of Information Science on April 1990. He was a councilor for the Aeronautics and Space related to the Technology Committee of the Ministry of Science and Technology during from 1998 to 2000. He was a councilor of Saga University for 2002 and 2003. He also was an executive councilor for the Remote Sensing Society of Japan for 2003 to 2005. He is an Adjunct Professor of University of Arizona, USA since 1998. He also is Vice Chairman of the Commission "A" of ICSU/COSPAR for 8 years, 2008-2016 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.html

Modeling of High Speed Free Space Optics System to Maintain Signal Integrity in Different Weather Conditions; System Level

Rao Kashif
Department of Electronic Science and Technology
University of Science and Technology of China
Hefei, China

Fujiang Lin
Department of Electronic Science and Technology
University of Science and Technology of China
Anhui, China

Oluwole John
Department of Electronic Science and Technology
University of Science and Technology of China
Hefei, China

Abdul Rehman Buzdar
Department of Electronic Science and Technology
University of Science and Technology of China
Anhui, China

***Abstract*—Free space optical (FSO) also known as free space photonics (FSS) is a technology widely deployed in Local Area Network (LAN), Metro Area Network (MAN), and in Inter & Intra chip communications. However satellite to satellite and other space use of FSO requires further consideration. Although FSO is highly beneficial due to its easy deployment and high security in narrow beam as well as market demand for 10 GB+, some factors especially rain, snow and fog attenuation causes signal integrity problem in FSO. To get better Signal Integrity in FSO we need to consider all components while designing the system. In this paper a comparative analysis has been performed on 10 GB and 40 GB FSO system over 1 Km. Firstly for selecting suitable modulation technique we compared NRZ and RZ modulation and get spectrum analysis. NRZ modulation was found more data efficient. Signal Integrity in FSO system with 10 GB/s was analyzed by eye diagram and Q-Factor of both APD and PIN Photo detector was presented in graph. Same experiment was repeated with 40 GB/s and Bit error rate of both photo detectors were presented.**

Keywords*—Free Space Optical; NRZ; RZ; PIN; APD; Photo Detector; BER; Q-factor

I. Introduction

Free space optical (FSO) is a transmission system which provide point to point, mesh and point to multi point communication by using laser and photodiodes. It can be a good candidate for high bandwidth future broadband and communication systems. Due to low BER, high bandwidth and easy installation FSO is popular in optical and wireless research community One more advantage of FSO is its Unlicensed Frequency Spectrum (800-1700nm) [1]. FSO is also a smart selection for intra satellite communication and due to small terminal and low power it has advantage over microwave links [2]. The first laser link to handle commercial traffic was built in Japan by Nippon Electric Company (NEC) around 1970. FSO is also efficient and being used for underwater communication, indoor wireless optical network, Intra chip to chip and board to board communication [3] and intra satellite communication [2].

Beside this, FSO also have number of challenges. One of those challenges is weather attenuation to optical signal. Rain, Snowfall and FOG are big challenges and they need researcher's consideration to maintain signal integrity in FSO system as the data-rate and coverage distance increases. To keep signal integrity throughout the system we need to consider few factors by selecting right blocks and components for modulation, receiver photodiodes, and Noise mechanism. In this paper we design a simple point to point FSO system in OpticSystem 14.0 and check signal integrity by PIN & APD photodiode in different weather conditions with 10 GB/s and 40 GB/s data rate over 1 Km.

II. NRZ And RZ Modulation Techniques

Selecting right modulation technique which will convert electrical signal into bit stream is the first step for optical system designing. Using INTERCONNECT (Lumerical) with 25 GB/s PRBS bit rate generator we compared NRZ (Non-Return to Zero) and RZ (Return to Zero) modulation as shown in Fig.1. NRZ modulation does not have rest state Fig.2a while signal drops to zero between each pulse in RZ Fig.2b. NRZ is more data efficient as it requires only half the bandwidth as compared to RZ Fig.3(a) & (b).

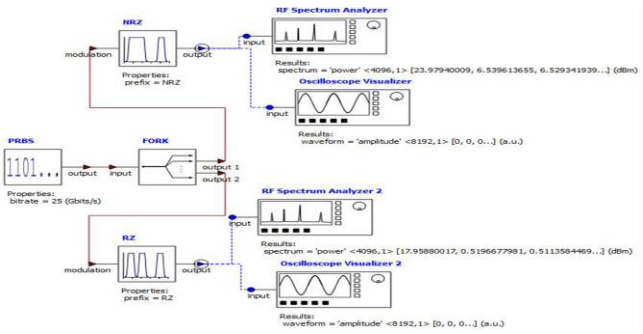


Fig. 1. NRZ & RZ Modulation Technique

Fig. 2. NRZ & RZ Oscilloscope analysis

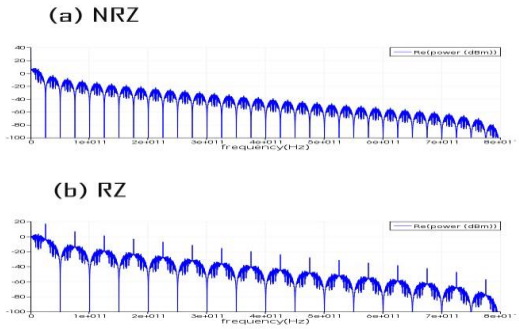


Fig. 3. NRZ & RZ bandwidth spectrum analysis

III. Attenuation in Free Space Optics System

As FSO system uses free space for transmission medium so weather condition is one of the challenges which require consideration [4]. Even the clean state of atmosphere can't be refers as free space because Nitrogen and Oxygen are there. Attenuation cause to signal in atmosphere is called atmospheric attenuation [5]. Beer's Law [5] (1) is used unremarkably to relate atmospheric attenuation. PR is received optical power and Pr is optical power at source. γ (L) is total attenuation [5].

$$\tau(\lambda, L) = \frac{P_R}{P_r} = \exp(-\gamma(\lambda)).L \qquad (1)$$

Rain attenuation is one of the causes for atmospheric attenuation in tropical regions. Rayleigh, Mie and Non selective are there different types of atmospheric scattering. Redirection of light which leads to reduction of received light intensity is called scattering [6]. Non-selective scattering happens when rain drop size is larger than wavelength [7]. Absorption occurs by interaction between molecules and propagation photons in atmosphere [8]. The visibility range is the distance travel by beam till its intensity drop to 5% of its real [9].

IV. Simulation Setup with PIN and APD Photo Detector

Photo detector is a device which converts optical signal to electrical signal. Photo Intrinsic Negative (PIN) and Avalanche Photo Diode (APD) are two photo detectors use in Free space Optic. The overview diagram of system which we design and use for comparison of APD and PIN photodiode is shown in Fig. 4. Simulation was done with 10 GB and 40 GB bit/s on distance of 1 km (1000 meter). NRZ modulation technique was used because its data efficiency is higher than RZ modulation as we compared it in Fig. 2 and Fig. 3. Shot noise and Thermal noise are two noise mechanisms in Photo detector. In Fig.5a we enable Shot noise in Photo detector and in Fig. 5b thermal noise were enabled. Due to high Q-factor we enabled thermal noise in our system. Laser frequency in system was 1550 nm. As Power of system vary according to different weather conditions (1 – 21 db). For tropical areas attenuation for haze and rain can be calculates by considering International visibility code refer from [10].Table 1 shows attenuation giving to the system.

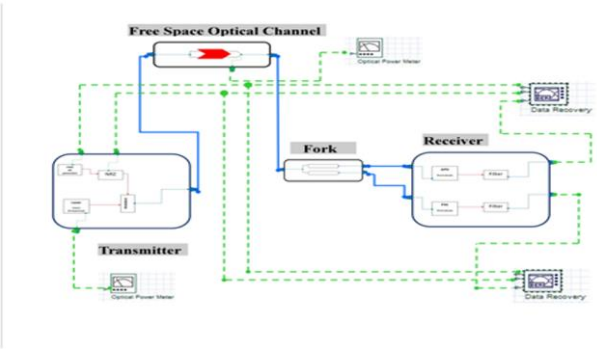


Fig. 4. System Overview Design

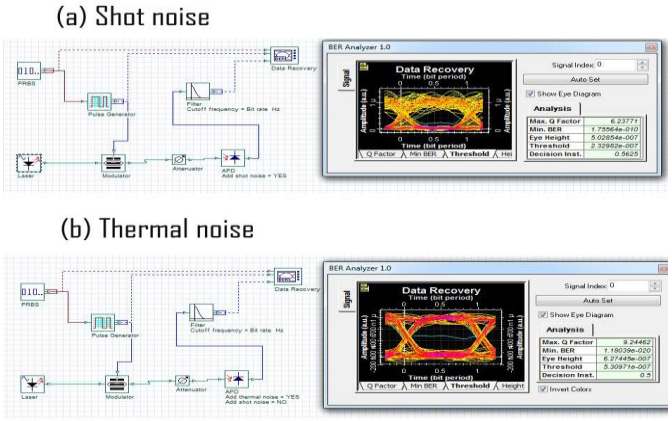


Fig. 5. Shot and Thermal noise Analysis

TABLE. I. WEATHER ATTENUATION

No.	Weather conditions	
	Weather Type	***Attenuation db/Km***
1	Clear	2.5
2	Haze	4.35
3	Light rain	12
4	High rain	20
5	Snow	25

V. PIN & APD PHOTO DETECTOR Q-FACTOR ANALYSIS WITH 10 GB

Figure 6 shows system diagram for 10 GB data over 1 Km Free Space Optical channel in OpticSystem 14.0. Different weather condition attenuation show in Table.1 was given to channel and EYE diagrams to get Q-factor analysis. Fig.7 shows result from APD photo detector and Fig.8 PIN photo detector. Q-Factor of both Photo detectors are analyzed and plotted in graph Fig.9. APD photo detector showed high Q-Factor as compared to PIN photo detector. As the attenuation increases the input power of system was also increased to maintain signal Integrity. Overall power for all weather conditions are plotted in Fig.10.

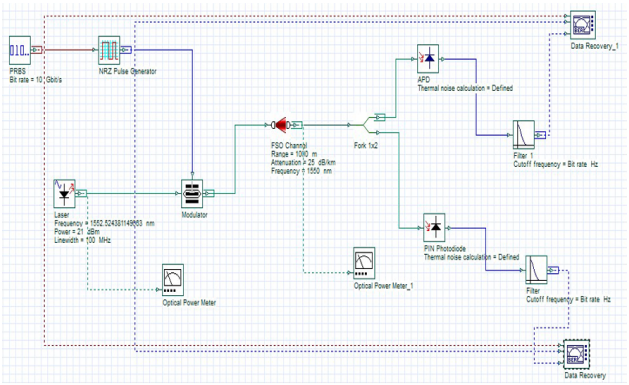


Fig. 6. System Diagram with 10 Gb/s

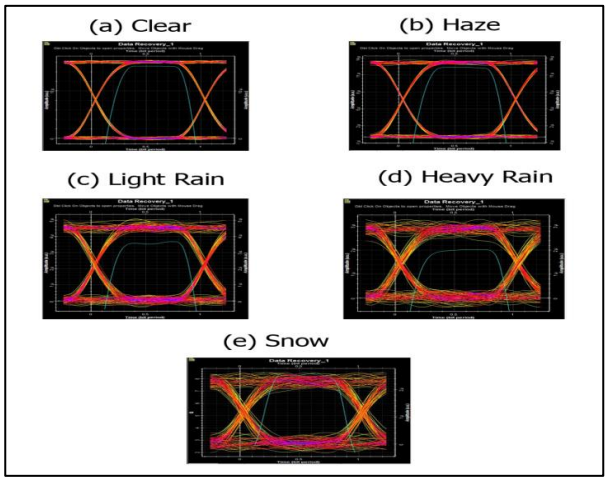


Fig. 7. APD Photo Detector EYE Analysis

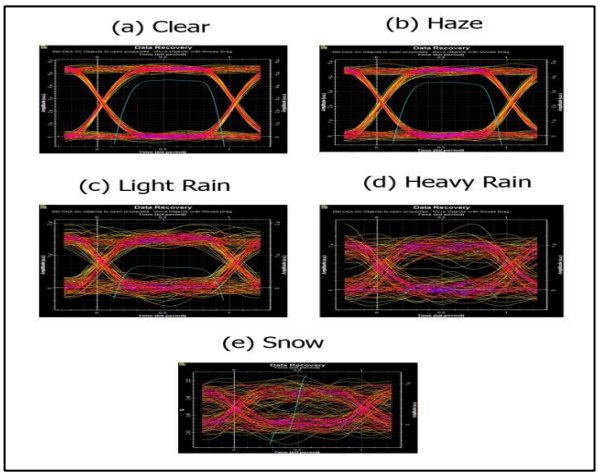


Fig. 8. PIN Photo Detector EYE Analysis

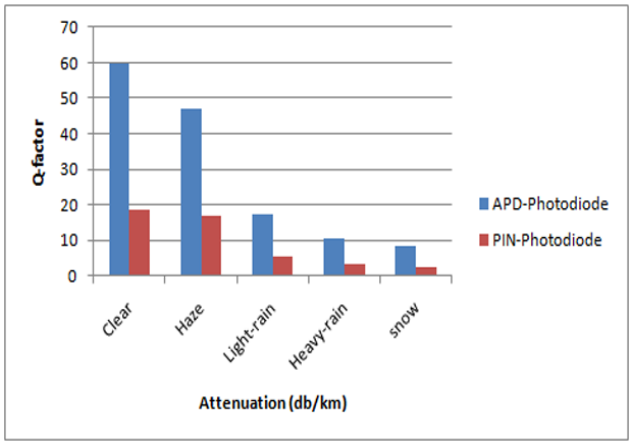


Fig. 9. PIN and APD photo detector Q-factor analysis

VI. PIN & APD PHOTO DETECTOR BIT ERROR RATE (BER) ANALYSIS WITH 40 GB

Same system in Fig.6 was used for 40 GB data transfer over 1 Km and signal Integrity was analyzed by analyzing Bit Error Rate (BER). Fig.11shows plotted graph of BER of APD and PIN photo detector. The whole system power increases with attenuation and increase in the data rate. Fig.12 graph shows over all power analysis for 10 and 40 GB in different weather conditions.

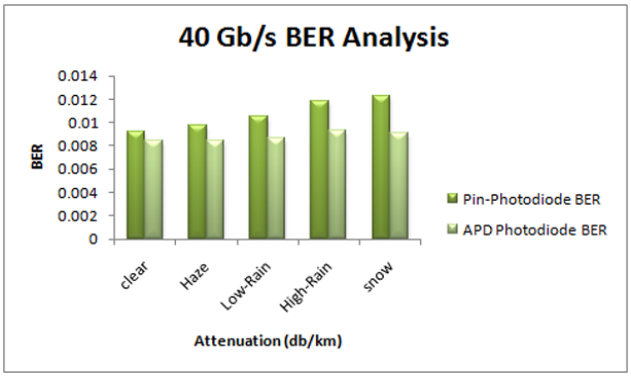


Fig. 10. APD and PIN photo detector BER Analysis

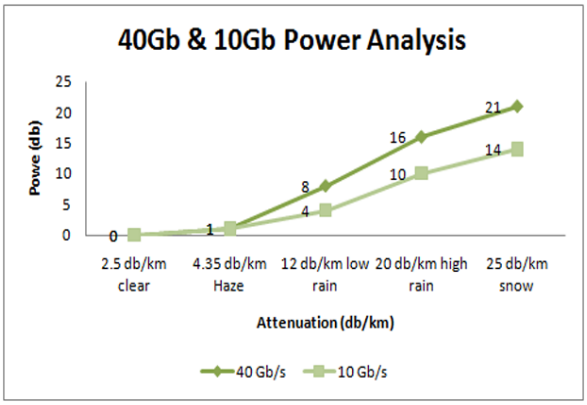


Fig. 11. Power analysis for 40 and 10 Gb system

VII. Conclusion

Signal Integrity and performance comparison of APD and PIN photo detector were evaluated in FSO system. We concluded that APD have better performance than PIN photo detector, so optical receiver with APD photo detector provide better signal Integrity as compared to PIN. As BER can be decrease by increasing optical power so future experiments can be on different NRZ modulation techniques to get better power Integrity in FSO communication.

Acknowledgment

This research is supported by CAS-TWAS and Micro/Nano Electronic System Integration R&D Center China. We thank our colleagues who provided insight and expertise that greatly assisted the research.

References

[1] Arun.K, Advance Free Space Optics (FSO), Volume 186, Springer, CA 2014, pp.09.

[2] Amon.S, and Kopeika.N, The performance limitations of free space optical communication satellite networks due to vibrations- Analog case, Proceedings of the Electrical and Electronic Engineers in Israel, 1996, Nineteenth convention of, November 5-6, 1996, Jerusalem,Isreal,pp.287-290.

[3] Tsang.D, Free Space board to board optical Interconnection, Proceedings of SPIE 1563, Optical Enhancements to Computing Technology, December 1, 1991, University of Lowa, USA,pp.66-71.

[4] Zabidi.S, Islam.M, Khateeb.W, and Naji.A, Investigating of rain attenuation impact on Free Space Optics propagation in tropical region, Proceedings of Mechatronics (ICOM), 2011 4th International Conference On, May 17-19,2011, Kuala Lumpur, Malaysia,pp.1-6.

[5] Robert.K, Propagation handbook for wireless communication system design.Volume 1, CRC Press, Florida July 2003.

[6] Zhuanhong.J, Qingling.Z, Atmospheric Attenuation Analysis in the FSO link, Proceedings of International conference of Communication technology (ICCT'06), Nov 27-30 2006, Guilin, China.pp.1-4.

[7] Maha.A, Simulating atmospheric free space optical propagation, Proceedings of SPIE 4635, Free-Space Laser communication Technologies, April 2002, San Jose, California,pp.192-201.

[8] Popoola.W, Ghassemlooy.Z, Awan.M, and Leitgeb.E, Atmospheric Channel effects on terrestrial free space optical communication links, Proceedings of ECAI 2009 International Conference Electronics, Computers and Artificial Intelligence, July 3-5, Pitesti, Romania.pp.17-23.

[9] Willebrand.H, Ghuman.H, Free space optics: enabling optical connectivity in today's network, edition 1, Sama Publications, Indiana 2001.

[10] ITU-R P.1814, R. (2007-08). "Prediction methods required for the design

Qualitative Study of Existing Research Techniques on Wireless Mesh Network

Naveen T.H
Research Scholar
Visvesvaraya Technological University
Belagavi,
Karnataka,
India

Vasanth G
Prof & Head
Dept. of Computer Science & Engineering
Government Engineering College,
Krishnarajpet,Mandya District,
Karnataka, India

***Abstract*—Wireless Mesh Network (WMN) is one of the significant forms of the wireless mesh network that assists in creating highly interconnected communication node. Since a decade, there have been various studies towards enhancing the performance of WMN which is successful to a large extent. However, with the upcoming technology of pervasive and dynamic networks WMN suffers from various routing issues, Quality-of-Service (QoS) issue, channel allocation, sustainability of routes which makes the theory contradicting when considering for real-world challenges in wireless networks. This paper, therefore, briefs about fundamental information of WMN followed by a discussion of existing research trends and existing research techniques. Finally, the paper also discusses the open research issues after reviewing the existing research techniques.**

Keywords—Access Points; Channel Allocation; Internet Access; Routing Problems; Wireless Mesh Network; QoS

I. INTRODUCTION

In last few years, the energy consumption, as well as cost of the networking devices, is increasing rapidly. Energy is always a matter of concern for any form of wireless network. It is because the optimal residual energy of a node and its higher retention capability can increase the network longevity [1] [2]. In the area of the wireless network, Wireless Mesh Network (WMN) is one of the frequently selected topics of research owing to its increasing number of research problems [3][4][5]. Although WMN seems to have an easier implementation, there are certain sets of common problems which have been addressed by most of the researchers. The first problem in conventional WMN is to perform the selection of a precise radio technique over physical layer [6]. At present, the alternatives of such techniques are Multiple-input Multiple-output (MIMO), Ultra Wide Band (UWB), Code Division Multiple Access (CDMA), etc. [6]. For nodes to work effectively, it is also required to have the faster frequency switching capability. The second problem in WMN is that conventional contention-based methodologies are never enough to enhance the fairness or channel allocations [7]. In order to maintain a mesh topology, it is required for a node to cost-effectively adopt the multiple physical channels, which at present is still an open problem from the viewpoint of channel assignment. WMN is also increasing investigated on 4G and 5G network [8][9][10]. It should be noted that such faster data transmission will require utilizing smart antenna in WMN, which is still an open issue in WMN. It is because the selection of the better version of MAC protocol is still an open problem [11]. Apart from this design and development of routing mechanism in order to accomplish QoS, robustness in communication, reliability, and efficiency is still an open challenge in WMN. The third problem of WMN is poor performance of the transport protocol that still encounters issues in utilizing network resources in order to perform channel allocation fairly. Although, TCP is the most frequently adopted communication protocol over the internet but it cannot be used for WMN. From the application viewpoint, there are challenges too. It is a very challenging task for WMN for offering internet connectivity with retention of QoS [12][13][14][15]. Hence, topological problems along with deployment problems persist over WMN even in recent times. The fourth problem of WMN is scheduling or provisioning problems. The biggest challenge in WMN is that if the number of user or nodes keeps on increasing, it is very difficult to optimize the channel capacity. For this reason, the communication performance highly degrades to a larger extent those results in declination of channel capacity. Although such problems could be possibly overcome by adding new gateway nodes in the adjacent transmission zone, it gives rise to another challenging issue i.e. how to identify the location of repeaters or base station to be fixed? The sole purpose will be obviously to increase the channel capacity and to offer equal QoS to all the connected users. As such forms of a network are connected to each other using internet access; hence it is imperative that dealing with the communication-based requirement will be one of the toughest tasks in WMN. As WMN can be looked upon as an easier network implementation with flexibility to accommodate various numbers of nodes.

Hence, this paper reviews some of the existing techniques to circumvent such problems related to communication and looks for the open research problems in it. This paper describes the routing mechanisms for the mesh network and also the existing research work in it. The Sectional organization is as follows. Section II briefly highlights the fundamental information about WMN followed by the discussion of present research trends in Section III. Discussion of existing research techniques towards varied problems in WMN is carried out in Section IV followed by highlighting of open research issues in Section V. Finally, the conclusion and future work is briefed in Section VI.

II. WIRELESS MESH NETWORK

Basically, a WMN can be defined as an interconnected radio nodes using mesh topology. The formation of the WMN is done by a mesh access point, a mesh user, and gateway. A good example to understand mesh clients are smart phones, personal digital assistant while the mesh access point transmits the traffic towards the gateway. In WMN, it is not necessary for the gateway node to be connected to World Wide Web. In WMN, a mesh cloud is defined as the transmission zones of communication nodes that combine work as a single network. It is feasible for the mesh cloud to have access to such mesh cloud that finally forms a radio network that provides redundancy and is highly trustworthy. Interestingly, failure of a single node doesn't affect the communication process as other nodes find an alternative way to perform communication. WMN is also found to be working with existing IEEE standards e.g. 802.15, 802.11, 802.16, etc. Application of mesh network can be easily found in US military [16], smart energy system [17], satellite phones [18], etc. The WMN is an upcoming innovation that can convey Internet broadband access, remote access, and system availability for network system connectivity among administrators and clients at low expenses. It is correspondence systems that have progressively pulled in Internet Service Providers (ISPs) due to its fast developing and evolution of remote advancements. WMN is a promising innovation in giving high data transfer capacity system scope. WMNs will incredibly help the clients to be constantly online anyplace at whatever time by associating with remote cross section switches [19]-[24].

A. WMN architecture

Designing WMN architecture is an initial move towards giving high-speed Internet connectivity over a particular scope region. WMN uses Wireless Mesh Routers (WMRs) and Mesh Clients (MCs) for transmitting the data packet in a multi-hop pattern, while the network devices e.g. gateway have the lesser amount of adaptability as well as supportability towards the backbone network. To outline more, it is comprised of different forms of the communicating nodes that can interact with each other [22][23]. Middle of the routing nodes manages the signal quality, as well as forward bundles in the interest of different nodes. Additionally, it gives high-transmission capacity Internet get to and offers a minimal effort and adaptable organization. The foundation that backings a WMN are a remote cross section switch system, or Backbone Wireless Mesh Network (BWMN). BWMN gives Internet network to MCs in a multi-bounce style. MCs can get to the Internet through BWMN shaped by Wireless Mesh Routers (WMRs). A common WMN is delineated in Fig.1.There are three sorts of nodes in a WMN: WMN router, WMN gateway, and WMN client. The clients of WMN are the end-client gadgets, for example, portable PCs, PDAs, advanced cells, and so on, that can get to the system for utilizing applications like email, VoIP, diversion, area discovery, and so forth. These gadgets are thought to be versatile; they have constrained force, they may have the directing ability, and might constantly be associated with the system. WMN switches are in the system to course the system movement. They can't end nor start the movement. The switches have the restriction in portability, and they have dependable qualities. Transmission power utilization in cross section switches is low, for multi-jump interchanges system. Moreover, the Medium Access Control (MAC) convention in a cross section switch underpins various channels and numerous interfaces to empower adaptability in a multi-bounce network environment. WMN portals are switches with wired base/the Internet. Since the passages in WMNs have various interfaces to associate with both wired and remote systems, they are costly. Consequently, there is a couple of number of WMN passages in the system. Also, their arrangement significantly affects the execution of the system. Fig.1 shows the WMN architecture with Mesh Router (MR), GWN (Gateway Node), WC (Wireless Client), AP (Access Point), BS (Base Station), and PC (Personal Computer).

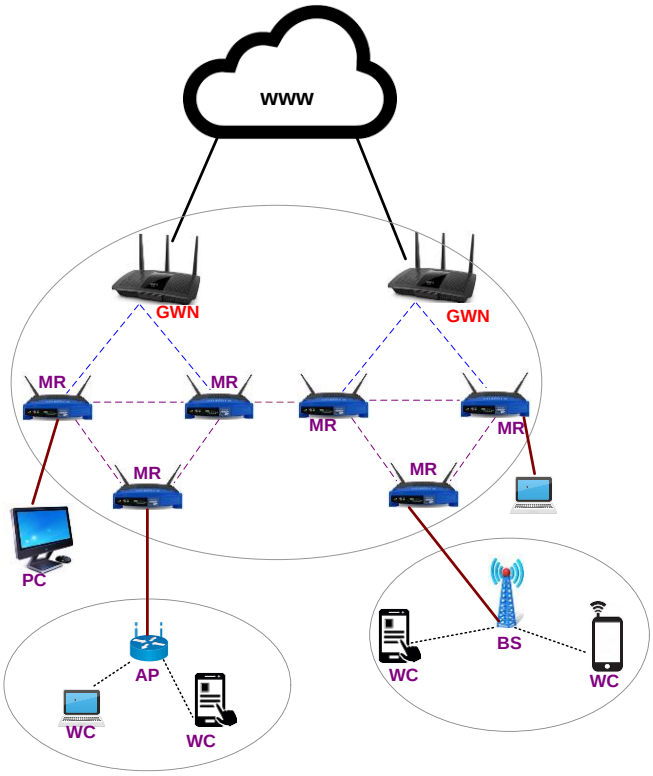


Fig. 1. WMN architecture

B. Classification of WMN

The connectivity based WMNs can be classified as sorts of the different system components, which are either Point to Multi-Point (MPM), Multi-Point (PTM), Point To Point (PTP), or Multi-Point to systems. The complete scientific categorization of this grouping appears in Fig.2. It is believed that PTP forms of the WMN are highly trusted worthy and offer very easier implementation of the wireless network. It basically consists of two communicating nodes (or radio) along with antenna with high gain in order to accomplish high-quality links. Such links are used for applications that demand maximized communication performance with higher speed data transmission. Unfortunately, PTP forms lack scalability and also suffers from lower adaptability. PTM form of network applies star topology to support both single and dual direction transmission. It normally uses the omnidirectional antenna for facilitating uplink transmission, and it uses the antenna with

high gain for supporting downlink transmission. Uses of PTM network are highly suitable for clients requiring high-speed data transmission without much focus on channel capacity. It is also used in backhaul operation. Although PTM networks are scalable to the moderate extent, it lacks reliability as well as adaptability. MTM network is meant for overcoming the flaws of PTM network i.e. to offer the higher degree of adaptability, reliability, and scalability. It is also suitable for large-scale network deployment. In MTM, the communicating devices are inter-connected with various forms of network nodes e.g. switches and routers. The increase in node number also has a positive effect on energy conservation. The utilization of these three forms of the WMN depends on the types of the application and networking demands of the clients.

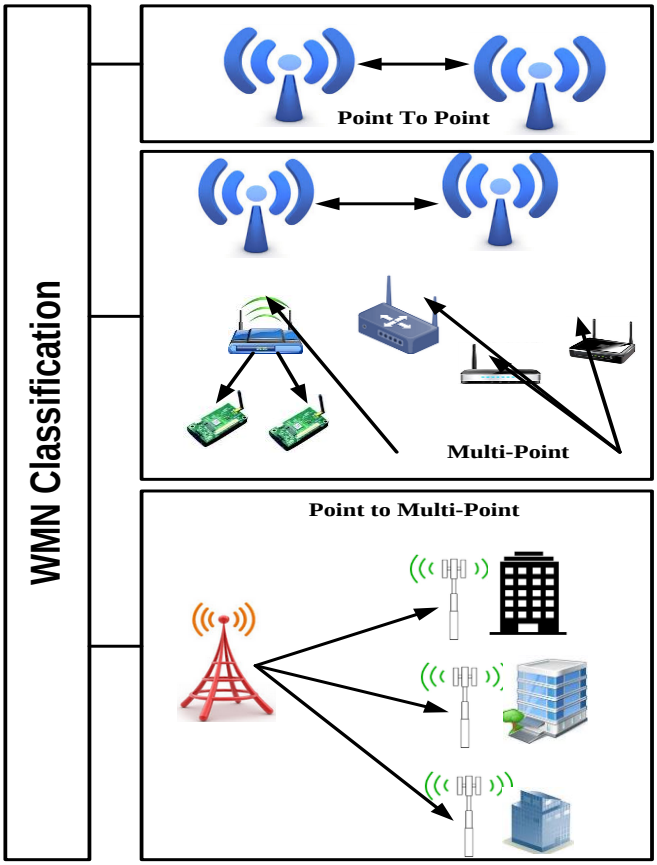


Fig. 2. WMN classification

C. Existing Standards for WMN

Numerous research organizations have continuous exploration ventures on different parts of WMN including energy management, standards, applications, and administrations. Some of the standards for WMN are described below.

- IEEE 802.11s Wi-Fi Mesh: Presently the IEEE 802.11 family is the adequate standard. Developed Service Set (DSS) and Wireless Distribution System are characterized in IEEE 802.11s for applying multi-bounce network methods and giving a convention to auto-arranging ways between WMRs. 802.11s has three fundamental parts viz. i) Network Portal (NP): goes about as a passage to different systems, ii) Network STA (station): goes about as a switch to hand-off edges hop by hop, and iii) Network AP (Access Point): gives handing-off capacities and also the availability administrations for customers.
- IEEE 802.15.1 Bluetooth: Bluetooth is the business name of this standard, and it is practiced on Wireless Personal Area Networks (WPANs). The task group is designed based on Bluetooth concept, and it correctly represents and specifies Media Access Control (MAC) and physical layer (PHY) for different forms of wireless transmission with either stationary or mobile nodes. There are two conceivable cross section topologies in WPAN network systems: full work topology or fractional lattice topology. A full work topology utilizes direct association plan. It implies that every remote node is associated specifically with every other node. Then again, some remote nodes are associated with all others while some are associated just to the remote hubs which hand-off the information.
- IEEE 802.15.4 Zigbee: Zigbee was first developed by Motorola. In Wireless Personal Area Network (WPAN) standard there are two association courses of action, single-bounce, and multi-jump association. Zigbee can bolster network topology by characterizing an organizer. The facilitator is capable of designing the system topology in multi-jump style. It is exceptionally appropriate for Wireless Sensor Mesh Networks (WSMNs).
- IEEE 802.16 WiMAX: This is a remote correspondence standard for Metropolitan Area Networks (MANs) which plans to give Mobile Multi-hop Relay (MHR) usefulness. This will permit sending multi-bounce network topology in WiMAX utilizing some WiMAX base stations to fill in as transfer stations. Setting up a multi-hop network topology gives a cost effective approach. The IEEE 802.16j standard was affirmed on May 2009 as a revision to the IEEE 802.16-2009 standard.

D. Significances of Wireless Mesh Network:

The following are uses of WMNs: -

1) Self-Organizing and Configuring: WMNs are adaptable in system design and not rely on upon the execution and the conventions. Self-mending and self-arranging are the WMNs highlights. This, diminishes the set-up time and support cost. Aside from this, it upgrades system execution. Because of these elements, the system administration suppliers can change, grow, and adjust the system as expected to achieve end clients requests.

2) Low Deployment Cost: Mesh switches are remote and they can benefit in multi-hop transmission. Therefore, utilizing remote switches as a part of substantial territories are less expensive contrasted with single jump switches/access focuses that they have wired associations. Ordinarily, because of wired associations those are more costly, in addition to quicker establishment and upkeep prompts a lower operation cost.

3) Enhanced Reliability: One of the potential features of mesh network is its non-dependability of any specific node. This will mean that in case of any node failures, the communication is still facilitated by other inter-connected nodes. The system also introduced various data exchange mechanism to minimize the bottlenecks in congested zone of the system as well. This additionally permits the activity burdens to be adjusted in the system. The technique also permits various load balancing schemes in order to sustain heavy traffic condition.

4) Scalability: In conventional remote systems, when number of communication nodes increases, the system execution will be influenced to a large extent. Yet, in WMNs, expanding the quantity of communication nodes will build transmission limit for better load adjusting and backup ways to go. For the most part, the nearby bundles (created in customers of cross section switch) run quicker contrasted with packets (produced in two or more hops away) from the neighbors.

5) Interoperability: WMN has a hybrid design which is good with existing benchmarks, for example, WiMAX, Cellular, Wi-Fi, Zigbee, Vehicular, etc. Hence, it is appealing for incremental usage and reuse of existing base station. All innovations specified above, are capable or will be capable soon to arrange a WMN and perform communication with every other. The vast majority of fundamental changes required in systems to empower them speaks with others; can expand the present benchmarks to keep up interoperability [25].

E. Routing conventions for WMN

The routing mechanisms generally derived as proactive, reactive and hybrid type. Proactive methodology performs as established, wired network system. Routers ensure that no less than one way reaches to any goal. Then again, receptive conventions designate the way if just there is a packet that will be sent to the goal. On the other hand, a communication nodes do not have a packet to send to a specific goal; then communication nodes does not ask for a way to this goal. The routing for the WMNs can be taken as four main types like Adhoc based, Controlled flooding based, traffic aware based and opportunistic routing protocol [26][27].

- Ad-hoc Based Routing: Usage of an ad-hoc network is quite higher in WMN. It has good supportability of assisting in routing over dynamic topologies in WMN.
- Controlled Flooding-based Routing: These are used to reduce the control cost. In this, the point is to surge the system unnecessary as loads of association in remote systems happens between close communications nodes. Hence, it is not important to send control packets to inaccessible communication nodes as often as close nodes. Another method for lessening overhead is constraining the quantity of nodes which are in charge of flooding.
- Traffic-Aware based Routing: Traffic-based routing methods consider WMNs general traffic matrix. In this, the ad-hoc on demand distance vector-spreading over tree adjusts AODV from specially appointed systems. In this routing, the gateway requires current way information from each communication nodes in the system to upgrade directing table.

III. Present Research Trends

In order to understand the present research trend, we investigate the manuscript publication towards different forms of wireless network published in the duration of 2010 to till date.

TABLE. I. Summary of the Present Research Trend [CF: Conference, J: Journal, S: Standard, EAA: Early Access Article, B: Books, C: Course]

Keyword	CF	J	S	EAA	B	C
WMN	3641	494	22	12	9	2
WSN	39332	6921	365	60	7	1
WLAN	12650	2192	157	40	5	4
Cellular Network	788	23	0	4	0	0
Adhoc Network	13042	3469	222	34	31	4

From Table 1, it is very much clear that there are very less work being carried out toward WMN as compared to other frequently used wireless network e.g. Wireless Sensor Network (WSN), Wireless Local Area Network (WLAN), cellular network, and adhoc network. Apart from this, we also observed the following trends:

- **Energy Efficiency**: There are approximately 70 journal pertaining to energy problems in WMN, which is quite less as there are various energy-modeling practices in wireless networks.
- **Security**: Security is received less attention with 46 journals whereas there are massive set of security algorithms.
- **Routing**: Routing has received massive attention as there are 214 journal and 1711 conference papers. Hence, there is a high progress made for communication protocols.
- **Channel Assignment**: Studies towards channel assignment has received the lowest research attention as can be seen from only 33journal.
- **Optimization**: Although, there are 76 journals towards optimization techniques in WMN, it is very less in number as there are large numbers of optimization techniques in recent times.

IV. Existing Research Techniques

This section discusses the existing research work towards wireless mesh network. It was noticed that in recent times, the prime emphasis was over the routing schemes in WMN. These schemes act as an enhancement towards the conventional communication protocols. Al-Saadi et al. [28] have addressed the communication problems in heterogeneous WMN with an objective of accomplishing better quality-of-service. The uniqueness of the study is that it has integrated current used 4G standards with IEEE 802.11 in order to incorporate cognitive principles over routing. The simulated study outcome was

found to offer increased transmission capacity with enhanced capability to transmit the packet over longer routes. Study towards enhancing the throughput in WMN was carried out by Ashraf [29]. The authors have addressed the problem of determining a capacity for augmentation purpose in WMN. For this, a mesh network is designed with the single radio with the single channel with a target of achieving maximum traffic flow. The authors have also used mixed-integer linear programming with the greedy approach to finding increased throughput on various network topologies (both single and multiple). Routing techniques also directly affect the energy factor within the nodes and so is channel allocation mechanism. Although, there are various channel allocation techniques in wireless network but very few effective ones for WMN. Hence, Avallone et al. [30] have addressed the problem of integrated allocation of a channel as well as enhancing the routing operation. The authors have used the heuristic-based approach that allows least amount of energy to be allocated to each channel with multi-radio networks. Problems of accomplishing high throughput in WMN was also carried out by Chakraborty et al. [31] using an opportunistic approach. The primary problem addressed in this work is to countermeasure the exposed or latent node issues in WMN. The technique uses block acknowledgment as well as aggregation of data frames. For better outcomes, conventional collision avoidance techniques along with Carrier Sensing technology are used. Fadlullah et al. [32] have used the case study of solar power in WMN for checking out its effectiveness in power harvesting mechanism. A similar study considering problems of energy and throughput was seen in the work of Li et al. [33]. The major problems discussed the authors are instability and inadequacy of power supply owing to various forms of dynamics in WMN. The technique has presented both online and offline evaluation scheme for energy efficiency considering anticipated evaluation of incoming traffic. The position of the access points a significant role in routing process of WMN. Lin et al. [34] have presented a technique where optimization technique of simulated annealing is used over various priority constraints in WMN. Implemented over multiple cases of grids, the evaluation is continued for assessing the fitness function with respect to increasing rounds. Apart from access point position, the interface also affects the routing performance in WMN. One such work was carried out by Mansoori et al. [35] most recently where a different number of interfaces have been used over WMN with different and customized channel capacity to meet the transmission demands of various users in the network. Roh et al. [36] have presented a technique that performs allocation of the channel along with scheduling of routes in order to address the problem associated with controlling the rate. The authors have used Bender's decomposition algorithm to find the simulated outcome with optimized communication performance in a network. Yu et al. [37] have presented a technique that ensures quality of service and enhance the utilization technique of channel capacity.

The author Draves et al. [38] have defined significant study for WMN, radio network routing. The study was intended to achieve the high throughput path among the source and destination and the mechanism which is presented is the test over the 23 number of nodes, and each node are placed with 802.11 wireless cards. The method has obtained the better results in routing than other existing metrics. The combined work of Iannone and Fdida [39] gives a conceptualized overview of mesh distance vector mechanism for the routing of WMN. The mechanism combines routing computation of routing along with the client's path demand for the network. The presented mechanism offers better packet data, exchange, format procedures. The method gives the reduced routing table size, easy management. A Hybrid or combined routing mechanism (Proactive and reactive algorithm) for WMN is presented in Oh [40]. In this, the proactive algorithm will function when the network has low mobility while the reactive algorithm will function when the network becomes vary mobile. Shih et al. [41] provided a decentralized mini slot uplink and downlink scheduling protocol traffic in IEEE 802.16. In this work, the simulation results represent better network throughput. Li et al. [42] illustrated the secure transmission protocol with efficient identity-based encryption in WMN. Author has demonstrated that the session key has the few security properties which chronicle enforceability, privacy and non-renouncement, and the new convention has the greatly improved execution than the other existing techniques.

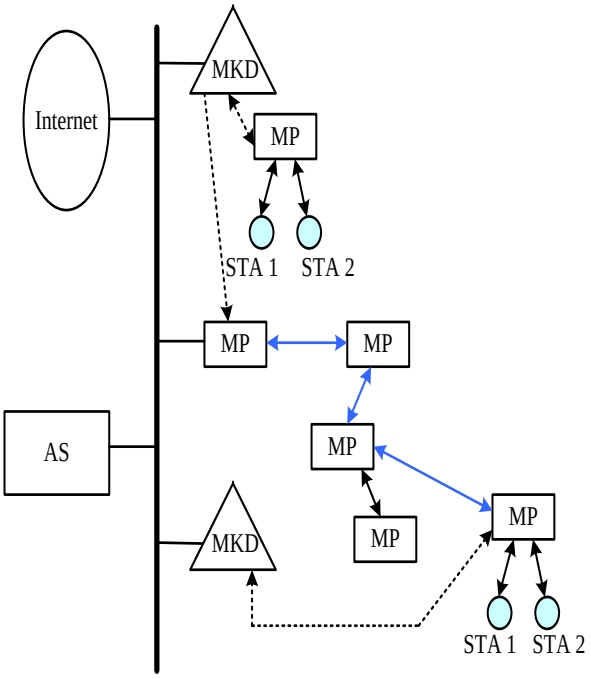


Fig. 3. Wi-Fi Mesh mechanism in Li et al. [42]

The above Fig.3 gives visualization of the Wi-Fi mesh model adopted in Li et al. [42] work. Basically, the technique has attempted to provide security over multiple hop WMN using three different modules i.e. i) mesh point, ii) key distributor, and iii) authentication server. It uses conventional identity-based encryption technique for accomplishing security features in WMN. Mogaibel et al. [43] have presented the on-demand adhoc routing protocol for the purpose of enhancing channel allocation scheme in WMN. The simulation results shows improve performance of multi-radio multichannel WMN. Majority of such forms of implementation was carried out considering simulation parameters e.g. traffic type, simulation time, propagation model (two-ray ground

reflection), number of nodes, packet size, number of radios, traffic type, and number of connection. Boukerche et al. [44] have used Optimized Link State Routing (OLSR) in order to form a large-scale self-organized WMN. The eminent advantage of a self-organized network is its ability to perform network control and management, which reduces both the developmental complexity and the need for maintenance of these networks. The model concludes that it increases throughput and improves the delay and packet delivery of the overall network when compared to the original OLSR protocol.

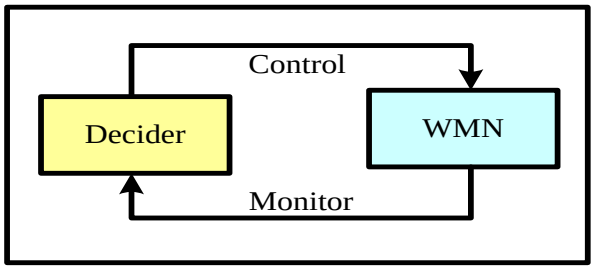


Fig. 4. The flow of [44] work

Fig.4 shows that decider flow consists of the guidelines and connected strategies for the component to be overseen (i.e., the cross section switch). Controls flow is responsible for controlling the properties of presented technique and applying them to the picked directing convention and the monitor flow will collect the data and sends it to the decider stream, Aiming to build the tenets and arrangements as occasions are produced will in the system. Tsai and Chen [45] have described an IEEE 802.11 MAC Protocol over for issues in WMN. Author has discussed his work performance by comparing with other existing research works. Li et al. [46] have presented a technique where the frequently used on-demand algorithm was enhanced to incorporate security feature. The study outcome shows that the presented is more effective in security against identified routing attacks. Le et al. [47] have presented a concept that computes the traffic load in presence of multi-radio over WMN. The analysis of simulation results gives that the network performance was enhanced completely in multi-radio mesh network. Li et al. [48] have illustrated a novel technique that can perform authentication of the wireless mesh network using Kerberos protocol. The conventional Kerberos protocol possess a few constraints in accomplishing clock synchronization and putting away key; in the interim, it is powerless from secret word speculating assault and assaults brought on by malignant programming. In this work, a very unique technique of authentication was presented by the author. By using public-key encryption strategies, the security of the proposed plan is improved. The examination demonstrates that the enhanced authentication technique is fit for remote Mesh system, which can make character confirmation more secure and productive.

Zhao et al. [49] have provided a hybrid technique of constructing communication within WMN. Network topology has been effectively looked into and created as a key answer for enhancing the execution and administrations of remote interchanges. The implementation of the presented technique was carried out over a network with fixed backbone. The authors have presented communication scheme for both Intra and inter WMN. The simulated outcome was assessed using throughput, delay, and mean queue size. In this, the hierarchical mesh network is more than that Point-to-Multipoint (PMP) network. The coverage area of the hierarchy mesh network is larger than that of the PMP network, so it could support more clients and get a higher network throughput. Sun et al. [50] have given a new routing protocol in cognitive wireless mesh networks. An innovative routing protocol associated with cognitive radio named AODV-COG was presented in this work. The simulation results show improved the throughput of the network. Ding et al. [51] have provided the reputation-based Proactive Routing Protocol (PRP) for the Wireless Mesh Backbone (WMB). In this author has designed the adaptive reputation management mechanism and with its simulation results he has concluded that the routing performance was enhance.

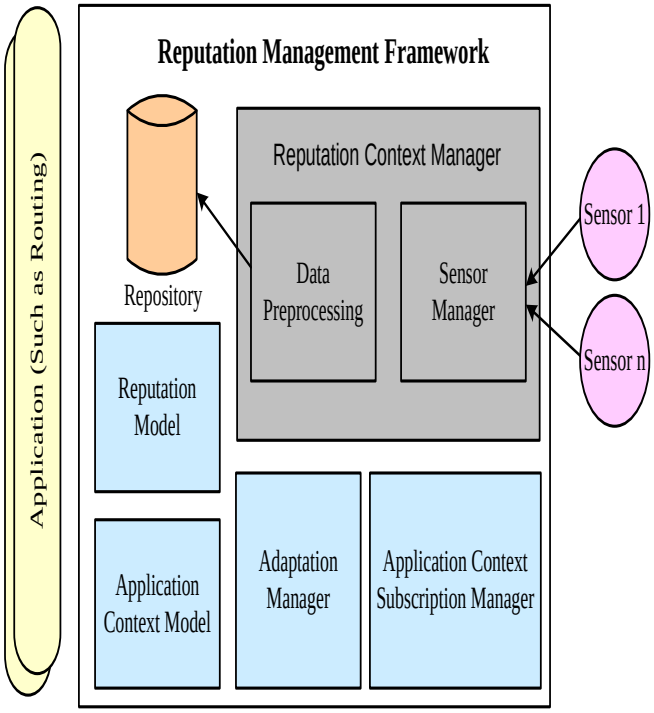


Fig. 5. Ding et al. [51]'s Adaptive Application Systems

Fig.5 represents the author's adaptive application system architecture. The architecture has context-aware applications, external source, and Adaptive Reputation Management Framework (ARMF) which computes, proliferates, stores, notoriety data gathered from heterogeneous connection sources, and makes this data accessible to abnormal state applications. For this situation, they are the routing protocol with various routing measurements. Paschoalino and Madeira [52] has demonstrated the Scalable Link Quality Routing Protocol for Multi-radio WMN. This simple and scalable approach identified the shortest paths and achieved better performance than OLSR, with a low increment in overhead. Jain et al. [53] have presented distributed protocols for WMN to schedule and control transmission rate to have Max-Min Fairness connection. In this, flow control of back pressure is used to minimize the rate at which packets are injected into the

network to flow transmission rate not exceeding its maximum rate delivered to the fair destination rate. Matam et al. [54] have presented a security technique towards link establishment in WMN. The current peer joins Foundation convention archived in IEEE 802.11s standard is not secure and powerless against transfer and wormhole attacks. To address this issue, a proficient strategy utilizing area data is presented in this work. Security of the proposed system is examined utilizing simulation study. Wehbi et al. [55] described a client management protocol for WMN that can offer the better connection for to clients while moving while moving. In the author gives the key mechanism for upcoming WMN. AI-Ghadanfary and Al-Somaidai [56] have presented routing protocols and simulated some of these protocols for multimedia applications. The result of the works says that throughput performance of AODV protocol is better when network nodes increased both for moving or fixed scenarios.

V. Open Research Issues

After reviewing the existing techniques towards WMN, it has been seen that there are various forms of the techniques that has been introduced till date in order to enhance the communication performance of WMN. Still, there are certain pitfalls that have not been found to be addressed and hence they broaden the scope of further research work towards such unaddressed problems. The brief highlights of the open research issues are as follows:

- **Less Focus towards Joint Protocol**: There are lesser studies carried out towards joint protocol usage. Although, certain number of joint routing and energy problems has been addressed, but none of the studies till date has used standard RF circuitry principle of the transmitting node. For this reason, the existing techniques could show superior outcomes from simulation viewpoint but it doesn't guarantee its applicability in real-world implementation.
- **Few Enhancements on Scheduling**: The existing scheduling techniques are more focused on channel resources in order to obtain better quality-of-service. However, there is no focus on energy efficiency towards existing scheduling techniques that possess a big impediment towards nodes working on adverse geographical environment.
- **Less Predictive Approach to Ensure QoS**: Although, there are voluminous studies being carried out towards ensuring reduced delay, higher throughput, etc, but existing techniques doesn't really use any form of predictive schemes based on dynamicity of the traffic. There are also less number of studies where QoS is guaranteed over the uncertain presence of congestion factor. Existing mechanism will require prior information about the traffic load in order to ensure QoS which doesn't happens in real time.
- **Less benchmarked techniques on Routing**: At present, there is no standard or benchmarked modeling of any research work where optimal performance of routing is reported. Studies towards benchmarking will require extensive test-environment to implement the routing over WMN, which is yet to be explored in future.
- **Few Focus on Optimization**: Existing optimization of communication performance in WMN calls for using highly recursive algorithm. This causes higher depletion of node resources as well as occupancy of channel capacity. Moreover, there is no optimization techniques exist that considers network constraint, topology dynamics, and network-based problems (e.g. scattering, fading, interference, etc). Hence, study towards optimization will require a special attention.

VI. Conclusion

This survey paper gives the important aspects of a Wireless Mesh Network along with some existing mechanisms in WMN routing to attain better solution for variable load. The surveys of various researches performed in WMN are discussed from recent IEEE transaction journals. With the recent research gap in existing work, the future study solution is provided which can offer a better idea to have good routing in WMN.

Our future work will be in the direction of accomplishing the open research issues. We will mainly focus on developing an effective scheduling that has potential supportability towards multihop transmission in WMN. The work will also focus on incorporating a predictive principle in order to achieve higher scalability with potential robustness against dynamic traffic condition. Finally, we will also investigate to evolve up with the new non-recursive optimization algorithm. The future solution for better routing in WMN can be attained by below strategic process.

- An algorithm can be designed to mitigate the hidden terminal issues in multi-hop WMN by reducing contention based channel access latencies.
- The optimizing algorithm for enhancing QoS parameters by considering multipath routing, congestion, load balancing.
- Benchmarking of the system by comparing the outcome will be compared along with most significant work of the existing system.

References

[1] Lisa Gansky, The Mesh: Why the Future of Business Is Sharing, Penguin, 2010

[2] Vishram Mishra, Jimson Mathew, Lau Chiew Tong, QoS and Energy Management in Cognitive Radio Network: Case Study Approach, Springer-Technology & Engineering, 2016

[3] Karthika K. C, "Wireless mesh network: A survey," *2016 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*, Chennai, 2016, pp. 1966-1970.

[4] M. Eslami, O. Karimi and T. Khodadadi, "A survey on wireless mesh networks: Architecture, specifications and challenges," *2014 IEEE 5th Control and System Graduate Research Colloquium*, Shah Alam, 2014, pp. 219-222

[5] D. C. Karia, A. Jadiya and R. Kapuskar, "Review of Routing Metrics for Wireless Mesh Networks," *2013 International Conference on Machine Intelligence and Research Advancement*, Katra, 2013, pp. 47-52.

[6] Ian F. Akyildiz, Xudong Wang, Wireless Mesh Networks, John Wiley & Sons, 2009

[7] A. B. M. Alim Al Islam, M. J. Islam, N. Nurain and V. Raghunathan, "Channel Assignment Techniques for Multi-Radio Wireless Mesh

Networks: A Survey," in *IEEE Communications Surveys & Tutorials*, vol. 18, no. 2, pp. 988-1017, Secondquarter 2016.

[8] F. Giust, L. Cominardi and C. J. Bernardos, "Distributed mobility management for future 5G networks: overview and analysis of existing approaches," in *IEEE Communications Magazine*, vol. 53, no. 1, pp. 142-149, January 2015

[9] T. Sharma, K. Ritesh, N. Chauhan and S. Agarwal, "Analogous study of 4G and 5G," *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, New Delhi, 2016, pp. 2137-2140.

[10] Y. L. Ban, C. Li, C. Y. D. Sim, G. Wu and K. L. Wong, "4G/5G Multiple Antennas for Future Multi-Mode Smartphone Applications," in *IEEE Access*, vol. 4, no. , pp. 2981-2988, 2016.

[11] V.C. Gungor1, E. Natalizio2, P. Pace3, and S. Avallone, "Challenges and Issues in Designing Architectures and Protocols for Wireless Mesh Networks", Springer Journal for Wireless Mesh Network, pp.1-27, 2008

[12] S. Kolipaka, B. N. Bhandari and A. Dey, "Joint Admission Control and vertical handoff between WLAN and WIMAX in wireless mesh networks for QoS," *2016 IEEE International Conference on Engineering and Technology (ICETECH)*, Coimbatore, 2016, pp. 1018-1023.

[13] A. V. R. Mayuri and M. V. Subramanyam, "MPGA: QOS adequacy latitude aware cooperative spectrum sensing in Cognitive Wireless Mesh Networks by Meticulous Progression based GA," *2015 Conference on Power, Control, Communication and Computational Technologies for Sustainable Growth (PCCCTSG)*, Kurnool, 2015, pp. 318-325.

[14] N. R. Appini and C. D. V. SubbaRao, "QoS Aware Multicast Framework based on WayPoint routing for Hybrid Wireless Mesh Networks," *2015 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT)*, Coimbatore, 2015, pp. 1-8.

[15] H. Li, J. Zhang, Q. Hong, H. Zheng, Y. Wang and J. Zhang, "QoS-aware channel-width adaptation in wireless mesh networks," *2016 IEEE International Conference on Communications (ICC)*, Kuala Lumpur, 2016, pp. 1-6.

[16] John Edwards, The future of military comms on the battlefield, An Article from Defense System. Accessed from https://defensesystems.com/articles/2012/02/08/cover-story-military-communications-technologies.aspx on 31st Jan, 2017

[17] Y. Xu and W. Wang, "Wireless Mesh Network in Smart Grid: Modeling and Analysis for Time Critical Communications," in *IEEE Transactions on Wireless Communications*, vol. 12, no. 7, pp. 3360-3371, July 2013.

[18] https://www.google.com/patents/US9125041

[19] Raniwala, Ashish, and Tzi-cker Chiueh. "Architecture and algorithms for an IEEE 802.11-based multi-channel wireless mesh network." Proceedings IEEE 24th Annual Joint Conference of the IEEE Computer and Communications Societies.. Vol. 3. IEEE, 2005.

[20] Navda, Vishnu, Anand Kashyap, and Samir R. Das. "Design and evaluation of imesh: an infrastructure-mode wireless mesh network." Sixth IEEE International Symposium on a World of Wireless Mobile and Multimedia Networks. IEEE, 2005.

[21] Akyildiz, Ian F., and Xudong Wang. "A survey on wireless mesh networks." IEEE Communications magazine 43.9 (2005): S23-S30.

[22] Lee, K-D., and Victor CM Leung. "Fair allocation of subcarrier and power in an OFDMA wireless mesh network." IEEE Journal on Selected Areas in Communications 24.11 (2006): 2051-2060.

[23] Niculescu, Dragos, et al. "Performance of VoIP in a 802.11 Wireless Mesh Network." INFOCOM. 2006.

[24] Radunović, Božidar, et al. "Horizon: balancing TCP over multiple paths in wireless mesh network." Proceedings of the 14th ACM international conference on Mobile computing and networking. ACM, 2008.

[25] Akyildiz, Ian F., Xudong Wang, and Weilin Wang. "Wireless mesh networks: a survey." Computer networks 47.4 (2005): 445-487.

[26] Jun, Jangeun, and Mihail L. Sichitiu. "The nominal capacity of wireless mesh networks." IEEE wireless communications 10.5 (2003): 8-14.

[27] Draves, Richard, Jitendra Padhye, and Brian Zill. "Routing in multi-radio, multi-hop wireless mesh networks." Proceedings of the 10th annual international conference on Mobile computing and networking. ACM, 2004.

[28] A. Al-Saadi, R. Setchi, Y. Hicks and S. M. Allen, "Routing Protocol for Heterogeneous Wireless Mesh Networks," in *IEEE Transactions on Vehicular Technology*, vol. 65, no. 12, pp. 9773-9786, Dec. 2016.

[29] U. Ashraf, "Capacity Augmentation in Wireless Mesh Networks," in *IEEE Transactions on Mobile Computing*, vol. 14, no. 7, pp. 1344-1354, July 1 2015.

[30] S. Avallone and A. Banchs, "A Channel Assignment and Routing Algorithm for Energy Harvesting Multiradio Wireless Mesh Networks," in *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 5, pp. 1463-1476, May 2016.

[31] S. Chakraborty, S. Nandi and S. Chattopadhyay, "Alleviating Hidden and Exposed Nodes in High-Throughput Wireless Mesh Networks," in *IEEE Transactions on Wireless Communications*, vol. 15, no. 2, pp. 928-937, Feb. 2016.

[32] Z. M. Fadlullah, T. Nakajo, H. Nishiyama, Y. Owada, K. Hamaguchi and N. Kato, "Field measurement of an implemented solar powered BS-based wireless mesh network," in *IEEE Wireless Communications*, vol. 22, no. 3, pp. 137-143, June 2015.

[33] M. Li, H. Nishiyama, N. Kato, Y. Owada and K. Hamaguchi, "On the Energy-Efficient of Throughput-Based Scheme Using Renewable Energy for Wireless Mesh Networks in Disaster Area," in *IEEE Transactions on Emerging Topics in Computing*, vol. 3, no. 3, pp. 420-431, Sept. 2015.

[34] C. C. Lin, L. Shu and D. J. Deng, "Router Node Placement With Service Priority in Wireless Mesh Networks Using Simulated Annealing With Momentum Terms," in *IEEE Systems Journal*, vol. 10, no. 4, pp. 1402-1411, Dec. 2016.

[35] M. Mansoori, M. Mahdavi and H. Amini Khorasgani, "Performance Analysis of Large Multi-Interface Wireless Mesh Networks with Multi-Different Bandwidth Channel," in *IEEE Transactions on Mobile Computing*, vol. 15, no. 5, pp. 1237-1248, May 1 2016.

[36] H. T. Roh and J. W. Lee, "Channel assignment, link scheduling, routing, and rate control for multi-channel wireless mesh networks with directional antennas," in *Journal of Communications and Networks*, vol. 18, no. 6, pp. 884-891, Dec. 2016.

[37] X. Yu, P. Navaratnam, K. Moessner and H. Cruickshank, "Distributed Resource Reservation in Hybrid MAC With Admission Control for Wireless Mesh Networks," in *IEEE Transactions on Vehicular Technology*, vol. 64, no. 12, pp. 5891-5903, Dec. 2015.

[38] Draves, Richard, Jitendra Padhye, and Brian Zill. "Routing in multi-radio, multi-hop wireless mesh networks." Proceedings of the 10th annual international conference on Mobile computing and networking. ACM, 2004.

[39] Iannone, Luigi, and Serge Fdida. "Meshdv: A distance vector mobility-tolerant routing protocol for wireless mesh networks." IEEE ICPS Workshop on Multi-hop Ad hoc Networks: from theory to reality (REALMAN). 2005.

[40] Oh, Minseok. "A hybrid routing protocol for wireless Mesh Networks." 2008 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting. IEEE, 2008.

[41] Shih, Kuei-Ping, Hung-Chang Chen, and Chi-Tao Chiang. "A Decentralized Minislot Scheduling Protocol (DMSP) for uplink and downlink traffic in IEEE 802.16 wireless mesh networks." 2009 IEEE 20th International Symposium on Personal, Indoor and Mobile Radio Communications. IEEE, 2009.

[42] Li, Yahui, et al. "Efficient security transmission protocol with identity-based encryption in wireless mesh networks." High Performance Computing and Simulation (HPCS), 2010 International Conference on. IEEE, 2010.

[43] Mogaibel, Hassen Abd-Almotaleb, et al. "Channel Reservation Scheme Based on AODV Routing Protocol for Common Traffic in Wireless Mesh Network." Computer and Network Technology (ICCNT), 2010 Second International Conference on. IEEE, 2010.

[44] Boukerche, Azzedine, et al. "A self-x approach for OLSR routing protocol in large-scale wireless mesh networks." IEEE GLOBECOM 2008-2008 IEEE Global Telecommunications Conference. IEEE, 2008.

[45] Tsai, Tzu-Jane, and Ju-Wei Chen. "IEEE 802.11 MAC protocol over wireless mesh networks: problems and perspectives." 19th International Conference on Advanced Information Networking and Applications (AINA'05) Volume 1 (AINA papers). Vol. 2. IEEE, 2005.

[46] Li, Celia, Zhuang Wang, and Cungang Yang. "SEAODV: A Security Enhanced AODV routing protocol for wireless mesh networks." Transactions on computational science XI. Springer Berlin Heidelberg, 2010. 1-16.

[47] Le, Anh-Ngoc, Dong-Won Kum, and You-Ze Cho. "Load-aware routing protocol for multi-radio wireless mesh networks." Communications and Electronics, 2008. ICCE 2008. Second International Conference on. IEEE, 2008.

[48] Li, Min, et al. "A novel identity authentication scheme of wireless mesh network based on improved kerberos protocol." Distributed Computing and Applications to Business, Engineering and Science (DCABES), 2014 13th International Symposium on. IEEE, 2014.

[49] Zhao, Liqiang, et al. "A hybrid routing protocol for hierarchy wireless mesh networks." 2010 6th International Conference on Wireless Communications Networking and Mobile Computing (WiCOM). IEEE, 2010.

[50] Sun, Xuebin, Yurong Zhang, and Chenglin Zhao. "A new routing protocol in cognitive wireless mesh networks." Advanced Intelligence and Awarenss Internet (AIAI 2010), 2010 International Conference on. IET, 2010.

[51] Ding, Qing, et al. "RePro: A reputation-based proactive routing protocol for the wireless mesh backbone." INC, IMS and IDC, 2009. NCM'09. Fifth International Joint Conference on. IEEE, 2009.

[52] Paschoalino, Rachel de C., and Edmundo RM Madeira. "A scalable link quality routing protocol for multi-radio wireless mesh networks." Computer Communications and Networks, 2007. ICCCN 2007. Proceedings of 16th International Conference on. IEEE, 2007.

[53] Jain, Shweta, Samir R. Das, and Himanshu Gupta. "Distributed protocols for scheduling and rate control to achieve max-min fairness in wireless mesh networks." 2007 IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks. IEEE, 2007.

[54] Matam, Rakesh, Somanath Tripathy, and Bhumireddy Swathi. "Provably secure Peer-link establishment protocol for wireless mesh networks." 2014 International Conference on Advances in Computing, Communications and Informatics (ICACCI). 2014.

[55] Wehbi, Bachar, Wissam Mallouli, and Ana Cavalli. "Light client management protocol for wireless mesh networks." 7th International Conference on Mobile Data Management (MDM'06). IEEE, 2006.

[56] Al-Ghadanfary, Karam Anan, and Mohammed Basheer Al-Somaidai. "Simulation of some Routing Protocols in a client wireless mesh network for multimedia applications." Electrical, Communication, Computer, Power, and Control Engineering (ICECCPCE), 2013 International Conference on. IEEE, 2013

A Trust and Reputation Model for Quality Assessment of Online Content

Yousef Elsheikh
Department of Computer Science
Applied Science Private University
Amman 11931, Jordan, PoBox 166

***Abstract*—In recent years, online transactions have become more prevalent than it was. This means that the number of online users to perform such transactions keeps growing, causing an increase in the level of expectations for them. One of those expectations is to enable them to get a better understanding of such transactions before going ahead with it. Consequently, trust and reputation models represent an important milestone to support those users to make their own decisions to facilitate online transactions. Many of the common trust and reputation models used primitive methods to calculate the reputation of online content. These methods are usually inaccurate when there is a divergence in rating. In addition, the lack of predictability through the latter ratings in emerging trends. Others use a probabilistic model or the so-called weighted average, which usually focusing on a single dimension for online user ratings. Even those models that combine multiple dimensions of user ratings are usually not representative on the one hand, and on the other hand are with heterogeneous weights. This paper fills this gap by proposing a model to assess the trust and reputation of online content, relying on three factors namely user behavior, user reliability, and user tendency with homogeneous weights of interest to the user on the Internet. These homogenous weights will be used to measure the reputation of any online content. The proposed model has been validated and compared with some other well-known models, and showed a significant improvement in terms of the Mean Absolute Error (MAE). The proposed model is also good with sparse and dense datasets.**

***Keywords*—Online content; Quality assessment; Trust and reputation model; User behaviour; User reliability; user tendency**

I. Introduction

In recent years, online transactions have become a phenomenon [8]. However, many users are still reluctant to make such online transactions and the reason for this, the inability to assess online content, where the disparity in opinions and feedbacks provided by other users about that online content. Feedbacks are usually offered either in the form of ratings or reviews, as these are among the primary sources to assess the quality of online content [1]. Online content quality is usually assessed through a model of trust and reputation, which in turn collects processes and aggregates user ratings on a particular online content. The main component in such models is how to aggregate ratings. The results of such models is to assess the quality of online content in the form of either a numeric value or stars as is the case for some common models [1].

Recently, many of the trust and reputation models were used to evaluate the different forms of online content. Most of these models are used on a large scale, easy to obtain, for free at most and help in decision-making by the user [4]. However, the accuracy of trust and reputation models are always a source of interest to many users over the Internet because they are often reflect public opinion on a specific online content. The main challenge in the trust and reputation models is how to aggregate the user ratings to assess the quality of online content. The easiest solution is to use primitive methods to achieve a score of trust and reputation regarding specific online content. Despite the simplicity of these methods, they are not effective enough since they do not take into account the quality and reliability of the user or even the popularity of online content to be evaluated. It also cannot predict the trends emerging from recent ratings of the user [8]. Other methods were more mature, such as those used probabilistic and fuzzy logic models. These methods have achieved more accurate results than its predecessors, but it depends largely on the threshold points, which are set by the experts. However, some of the models used the weighted average in order to calculate the score of the reputation of any online content, since the weights may be the reputation of the user, user reliability [2], User leniency [7], rating time [4], or the difference between the current reputation score and the new rating score [11]. This method requires evaluating user rating to assess the quality and trustworthiness of online content, and thus reflect that weight.

This paper focuses mainly on aggregating the ratings using the weighted average method. Most current weighted average models of trust and reputation mostly focus on a single dimension of the weights for the user, for example, in the Lenient-Quality reputation model; the weight of the user is evaluated on the basis of being lenient or not in providing rates [7]. Even models that combine multiple user dimensions they often combine them through a discount function, and significantly associated with thresholds set by the experts [9]. Moreover, some of the weighted average methods do not take into account changes in the weights of the user in terms of reliability of the user, the time variation between the first rating and the latest rating of the same online content, and finally the user experience compared with the experience of other users about specific online content. Also, current methods of weighted average do not take into account user tendencies (positive or negative) during the process of rating online content. Therefore, this paper proposes a new model of trust

and reputation to assess the quality of online content through a combination of three factors and of great importance to the user on the Internet. These factors have been nominated to reflect the user's reputation as a value for the weight. These factors are: 1) user behaviour, 2) user reliability, and 3) user tendency.

II. Related Work

In a review of the literature, trust and reputation models fall into three basic categories namely the weighted average models, fuzzy logic models, and probabilistic models. The Weighted average models are the most commonly used where weights are calculated based on time or data relating to the user. In a study of [5], they used non-linear function of aging, depending on the time-based approach in calculating the weights. In an example of the same approach, the number of previous transactions has been addressed while the times of those previous transactions have not been addressed [10]. All common models mentioned above are not able to adapt to the data when it grows in size. In a study of [8] & [13], they addressed the variation in the user's rating, which was reflected on the user's decision on that item. Moreover, these models often do not take into account the credibility dimension in the user's rating, which reflects negatively on the accuracy of those ratings that are placed.

The other approach of the weighted average models is based on data relating to the user, where the weights are calculated by the reliability of the user, the user's credibility and trustworthiness to the user. In a study of [11], they proposed a model to measure the reliability of the user through the same user ratings. Higher weights were given to the users with ratings that are closer somehow to the average ratings made by the user for a specific online content. Another study of [7] & [12], they addressed and proposed a model that is based on rating user behaviour and tendency in order to measure the leniency of the user. Rating user behaviour or tendency is a value that reflects the extent of user's behaviour or tendency to provide ratings higher or lower than other users. Also, a study of [6] proposed a model of trust and reputation takes user ratings in the account through what is known as a polynomial probabilistic Bayesian probability distribution and Dirichlet distribution. In a study of [2], they proposed a model of trust and reputation for aggregating ratings through mixing in use between the weighted average method and fuzzy logic. User reputation depends on the accuracy of prediction compared with the ratings of other users for various elements in any online content. In a study of [9], they recently addressed the problem of unfair ratings provided by some users by proposing a model of reputation and trust uses fuzzy logic to address that problem.

However, all the above models lacked during the evaluation of the user's reputation to a comprehensive approach that combines three factors making up the model of this paper. In addition, the proposed model in this paper is also good with sparse and dense data sets as opposed to the rest of the models listed above.

III. Proposed Model

The proposed model uses three input factors that have great impact on the user's trust and reputation. These factors are 1) user behaviour, 2) user reliability and 3) user tendency, which are then fused together through the Arithmetic Mean to reflect the user's reputation of any online content as a value for the weight.

The first factor measures the user's behaviour in making the rating. In other words, it is measured when the user usually makes his/her rating. Is he/she the first who assessed online content? Or did he/she make his/her rating after many previous transactions? Rating time is important as it reflects the impact of past transactions on the user's decision. Therefore, time difference between the user rating and first rating received is measured in terms of day unite. Then, these are discounted using age decay function as shown in Equation 1.

$$y_1 = \frac{1}{m}\sum_{j}^{m} \mu^{T_{Cj}-T_{1j}} \quad (1)$$

Where m is the number of online items rated by a user. μ is the discounting variable (in this case we use $\mu = 0.95$). T_{Cj} is the timestamp when the user rated the online item j. T_{1j} is the timestamp of the first rating received for online item j.

The second factor measures the user's reliability. This factor assesses the accuracy of user in providing rating that is very close to the average of rating for the online item under assessment. Equation 2 shows how the distance between the user rating and average of ratings are discounted using discounting function μ^x.

$$y_2 = \frac{1}{m}\sum_{j}^{m} \mu^{|r_j - avg_j|} \quad (2)$$

Where r_j is the rating given by a user for online item j. avg_j is the rating average of online item j.

The third factor assesses the user's tendency in providing positive, neutral or negative ratings. Positive ratings are those that are larger than mid of the rating level, and negative ratings are the opposite while neutral ratings are those with the mid-range value. Rating level is the scale used to score the item, for example in most application the rating scale ranges from 1 to 5. The basic idea of this variable is to find the ratio between errors of either user positive ratings or negative ratings, and total ratings errors. The error is computed by finding the difference between a rating and average of the online item rating. Equation 3 shows how the user tendency variable is calculated. In this first step we classify user ratings into three sets, positive, neutral and negative set. To decide which set should be used, one should assess the rating of the online item under assessment, if the rating belongs to the positive set then we use positive error otherwise we use either negative or neutral set.

$$y_3 = 1 - \frac{\sum_{k=1}^{L}|r_k - avg_k|}{\sum_{j=1}^{m}|r_j - avg_j|} \quad (3)$$

Where L is the number of ratings in the targeted set, avg_k is the average of online item k in the targeted set.

The three factors y_1, y_2 and y_3 are fused together using the Arithmetic Mean as shown in Equation 4. The final online item score is computed as shown in Equation 5.

$$w_i = \frac{y_1 + y_2 + y_3}{3} \quad (4)$$

$$score_j = \sum_{i=1}^{n} w_i \times r_i \quad (5)$$

Where n is the number of user ratings for online item j. w_i is the normalized weight for user i and r_i is the rating provided by user i.

IV. Experiment Design

To validate the proposed trust and reputation model we applied 5-Fold cross validation, which divides the dataset into 5 sets of training and testing data. In each run, 80% of the users are used as training data to build trust and reputation model and 20% are used as testing data to validate the data against the generated score. The errors of validation at each run are recorded using the Mean Absolute Error (MAE) as shown in Equation 6. The MAE (MAE) assesses, for each online item, the closeness of the predicted items scores from the training dataset to the actual ratings in the testing dataset.

$$MAE = \frac{1}{m}\sum_{j=1}^{m}\frac{\sum_{i=1}^{n}(r_i - \bar{r}_j)}{n} \quad (6)$$

Where $\bar{r}_j$ is the predicted score for online item j. m is the number of online items in the testing data. n is the number of ratings for j^{th} online item in the testing data.

To investigate the performance of our proposed model we employ two common stable datasets that are publically available on the internet. These two datasets are 100K and 1M which are taken from large benchmark data repository called Movielens. Both datasets contain ratings for movies. The first dataset (100K) consists of 943 users and 1682 movies, whereas the second dataset consists of 6040 users and 3706 movies as shown in Table 1. Each user in both datasets has rated at least one item online and each item has been rated by at least one user. Both datasets have been widely used in validating trust and reputation models.

TABLE. I. Datasets characteristics

Dataset	No. Of Users	No. Of Movies	No. Of ratings
100K	943	1682	100,000
1M	6040	3706	1,000,209
10M	72,000	10,000	10,000,000

V. Experiment Results

To show the significant improvement, which we obtained in the results, we compared the proposed model with a range of well-known models for trust and reputation in the literature. Comparisons were made between the proposed model with six other models of trust and reputation namely the Average Based Reputation Model, Beta Distribution Based Reputation Model [1], Bayesian Reputation Model [6], Dirichlet Based Reputation Model [5], Fuzzy Logic Rating Based Reputation Model [2] and Lenient-Quality Reputation Model [7]. Table 2 shows for each model, the value for Mean Absolute Error (MAE) obtained on all databases [14].

TABLE. II. Mean Absolute Error Results

Dataset	Proposed Model	Average	BetaDR	Bayesian	Dirichlet	Fuzzy Logic Rating	Lenient-Quality
100K	**0.826**	0.905	0.893	0.911	0.898	0.916	1.024
1M	**0.780**	0.841	0.833	0.844	0.841	0.848	0.962
10M	**0.749**	0.791	0.812	0.790	0.776	0.795	0.917

It is noted above, the results of the proposed model provides better accuracy compared with other models. This shows, through the MAE value, that the proposed model has a lower MAE value than the rest of the models used, which means that it is better in terms of accuracy. Moreover, the above results confirmed that the proposed model can work well over the sparse and dense dataset alike. Sparse dataset is often a problem for many trust and reputation models as they do not work well when dealing with a small number of ratings. Although the Dirichlet model can work with uncertain, small number of ratings, the above results show that the proposed model was significant, compared with Dirichlet model, particularly in the dataset (100K). This is the biggest proof that the proposed model can deal with uncertain ratings in sparse datasets. Unusually, the proposed model shows significant improvement for dense datasets (1M and 10M). This means that the proposed model would generate accurate ratings for dense datasets. This is another proof of the efficiency of the proposed model regarding the dense datasets. The results showed that the proposed model is more reliable than the rest of the models over both spare and dense datasets.

VI. Conclusions

This paper presents a new model of trust and reputation, which uses three factors fused together through the Arithmetic Mean to reflect the user's reputation of any online content as a value for the weight. In other words, this value can be used to calculate the reputation or the quality of any online content alike. The proposed model showed good accuracy in terms of the Mean Absolute Errors (MAE) and addressed several problems for the rest of the models. All this was through sparse and dense datasets. The main limitation in this paper is that it focuses primarily on the aggregation of assessments using the weighted average method, while there are many aggregating methods that can be examined to demonstrate the accuracy of the proposed model.

Acknowledgment

The authors are grateful to the Applied Science Private University, Amman, Jordan, for the financial support granted to cover the publication fee of this research article.

REFERENCES

[1] A. Abdel-Hafez, X. Yue"An accurate rating aggregation method for generating item reputation." Data Science and Advanced Analytics (DSAA), 2015. 36678 2015. IEEE International Conference on. IEEE, 2015.

[2] K. Bharadwaj, M. Y. H. Al-Shamri. "Fuzzy computational models for trust and reputation systems." Electronic commerce research and applications 8.1: 37-47, 2009.

[3] H. F. Maxwell, and J. A. Konstan. "The movielens datasets: History and context." ACM Transactions on Interactive Intelligent Systems (TiiS) 5.4: 19, 2016.

[4] A. Jøsang, I. Roslan, B. Colin. "A survey of trust and reputation systems for online service provision." Decision support systems 43.2: 618-644, 2007.

[5] A. Josang, H. Jochen. "Dirichlet reputation systems." Availability, Reliability and Security. ARES. The Second International Conference on. IEEE, 2007.

[6] W. Andrew, A. Jøsang, J. Indulska. "Filtering out unfair ratings in bayesian reputation systems." Proc. 7th Int. Workshop on Trust in Agent Societies. Vol. 6. 2004.

[7] L. Hady W. Ee-Peng Lim, K. Wang. "Quality and leniency in online collaborative rating systems." ACM Transactions on the Web (TWEB) 6.1: 4, 2012.

[8] L. Christopher, S. Soumya, M. Chiang. "On the volatility of online ratings: An empirical study." Workshop on E-Business. Springer Berlin Heidelberg, 2011.

[9] W.L. Teacy, J. Patel, N. R. Jennings, M. Luck. "Travos: Trust and reputation in the context of inaccurate information sources." Autonomous Agents and Multi-Agent Systems 12.2: 183-198, 2006.

[10] M. Zaki, A. Bouguettaya. "Rateweb: Reputation assessment for trust establishment among web services." The VLDB Journal—The International Journal on Very Large Data Bases 18.4: 885-911, 2009.

[11] R. Tracy, and R. Wilensky. "An algorithm for automated rating of reviewers." Proceedings of the 1st ACM/IEEE-CS joint conference on Digital libraries. ACM, 2001.

[12] F.L. Garcin, L.I. Xia, B. Faltings. "How aggregators influence human rater behavior." Proc. Workshop@ 14th ACM Conference on Electronic Commerce (EC-13). 2013.

[13] M. Zaki, and A. Bouguettaya. "Rater credibility assessment in web services interactions." World Wide Web 12.1: 3-25, 2009.

[14] F. Maxwell Harper, Joseph A. Konstan. The MovieLens Datasets: History and Context. ACM Transactions on Interactive Intelligent Systems 2015, 5, 4, Article 19.

An Electrical Model to U-Slot Patch Antenna with Circular Polarization

Guesmi Chaouki
Department of Physics, FST
Unit of Research in High Frequency Electronic Circuits and Systems

Ghnimi Said
Department of Physics, FST
Unit of Research in High Frequency Electronic Circuits and Systems

Necibi Omrane
Department of Physics, FST
Unit of Research in High Frequency Electronic Circuits and Systems

Gharsallah Ali
Department of Physics, FST
Unit of Research in High Frequency Electronic Circuits and Systems

***Abstract*—The microstrip antenna is one of the best antenna structures, due to its low cost and compact design. In this paper, a coaxial feed circularly polarized square patch antenna is designed using the U-slot. The proposed antenna is suited for the RFID readers in the SHF band. This structure of antenna of FR-4 substrate (dielectric constant = 3.5), is capable to cover the range of frequency of 2.4 to 2.5GHz. The size of patch is 25*25 mm^2. An equivalent electrical model of this antenna was proposed and simulated by the ADS software. The simulated gain is 4.189 dBi and S_{11} bandwidth is about 100 MHz. Analysis and modeling of the proposed antenna was carried out using the CST and HFSS simulator based on the finite element method. The simulation results obtained are presented and discussed.**

***Keywords*—RFID; circularly polarization; U-slot antenna; RFID reader antenna; Electrical model**

I. Introduction

Previously, the circular polarization was created by feeding the antenna to different locations and with a 90 ° phase shift. At that time, the feeding was made directly (without slot) using a coaxial cable or micro-strip line. With the feeding in one place, circular polarization is induced either by making the antenna slightly rectangular (instead of square). Either by cutting two of its corners or by making a diagonal opening in its metallization. These three topologies have been studied by Sharma et al. [1].

The choice in this work stopped on the antenna with truncated corners. In addition to maintaining symmetry at the diagonal, this configuration is easier to conceive since it has a degree of freedom of less than the opening in the metallization. The latter can vary in length and width while the truncation is symmetrical. According to Sharma [1], the antenna with truncated corners provides the lowest axial ratio. But has a slightly smaller bandwidth (axial ratio) than the other topologies [2]-[4].

Many techniques have already been applied to the design of broadband antennas. For example, an insulated slot in a patch, the addition of different slot shapes at the radiating element, such as the L-shaped slot, a T-shaped slot, an H-shaped slot and the fractals slot have also been reported for large bands [5]-[9]. Among these techniques, a U-shaped slot will be used on the patch of the coaxially fed square patch antenna.

These types of antennas also have broad applications in long range and wireless identification or communication systems, such as RFID which is one of the new identification techniques and where the size of the system depends essentially on the size of the antenna [10]. In order to analyze this structure, a new electrical model is developed and compared to a physical patch.

Many calculation methods are adopted to resolve the maxwell equations and then analyse the performance of the antennas. Among these, three of them are broadly used in simulation software:

1) The method of moments (MoM) is used among others in ADS software.

2) The method of finite integral (FIT) is used in the software CST Microwave Studio [11], [12].

3) The Finite Element Method (FEM) is used in the software HFSS [11].

In this work, the method of moments will be adopted to analyze the performance of the electrical model and the results will be compared with those obtained by the Finite Integral and the Finite Element Method.

II. The Approaches to Get the Electrical Model of the Square Patch Antenna

The patch antenna can be modeled simply by a parallel or serial RLC circuit. To calculate its characteristics, the study based on the RLC circuit is the most used. In the next step, the parameters of the RLC circuit in the equivalent electrical model of the patch antenna will be calculated. The parameters of the proposed model are determined using the same solution of Nasimuddin and A. K. Verma [17].

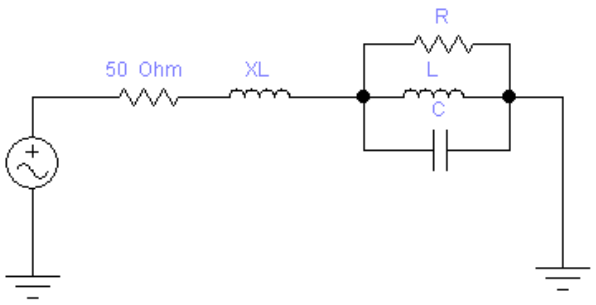


Fig. 1. Electrical model of square patch

The input impedance of a square patch excited by a coaxial cable is given in [18] by equation (1):

$$Z_{in} = R + jX$$

$$Z_{in} = \frac{R}{1+Q_T^2\left(\frac{f}{f_r}-\frac{f_r}{f}\right)^2} + j\left(X_L - \frac{RQ_T\left(\frac{f}{f_r}-\frac{f_r}{f}\right)}{1+Q_T^2\left(\frac{f}{f_r}-\frac{f_r}{f}\right)^2}\right) \quad (1)$$

The resonant resistance R of our parallel circuit RLC is given in [18] by equation (2):

$$R = \frac{Q_T h}{\pi f_r \varepsilon_{dyn} \varepsilon_0 A}\cos^2\left(\frac{\pi X_0}{a}\right) \quad (2)$$

f_r: resonant frequency.

X_0: the distance of the feed point from the edge of the patch.

h: thickness of dielectric.

A : air of the square patch

a : length of the edge

The total Quality factor Q_T is calculating in [18] by equation (3):

$$Q_T = \frac{1}{\frac{1}{Q_R}+\frac{1}{Q_C}+\frac{1}{Q_D}} \quad (3)$$

Q_R is the radiation quality factor, Q_D is the losses in the dielectric and Q_C is losses in conductor.

$$Q_R = \frac{C_0\sqrt{\varepsilon_{dyn}}}{4 f_r h} \quad (4)$$

$$Q_C = \frac{0.786\sqrt{f_r Z_a a h}}{P_a} \quad (5)$$

$$Q_D = \frac{1}{T_g \delta} \quad (6)$$

The impedance of a microstrip line filled with air "Za" is given by equation (7):

$$Z_a(a) = \frac{60\pi}{\sqrt{\varepsilon_r}}\left(\frac{a}{2h}+0.441+0.082\left(\frac{\varepsilon_r-1}{\varepsilon_r^2}\right) + \frac{\varepsilon_r+1}{2\pi\varepsilon_r}\left(1.451+L_n\left(\frac{a}{2h}+0.94\right)\right)\right) \quad (7)$$

$$Z_{a0}(a) = Z_a(a, \varepsilon_r = 1)$$

$$P_a = \frac{2\pi\left(\frac{a}{h}+\frac{\frac{a}{\pi h}}{\frac{a}{2h}+0.94}\right)\left(1+\frac{a}{h}\right)}{\left(\frac{a}{h}+\frac{2}{\pi}L_n\left(2\pi\exp\left(\frac{a}{2h}+0.94\right)\right)\right)^2}; \frac{a}{h} \geq 2 \quad (8)$$

The dynamic permittivity ε_{dyn} is calculated by (9):

$$\varepsilon_{dyn} = \frac{C_{dyn}(\varepsilon)}{C_{dyn}(\varepsilon_0)} \quad (9)$$

$$C_{dyn}(\varepsilon) = \frac{\varepsilon_0\varepsilon_r A}{h y_n y_m} + \frac{1}{2y_n}\left(\frac{\varepsilon_{reff}(a,h,\varepsilon_r)}{C_0 Z(a,h,\varepsilon_r=1)} - \frac{\varepsilon_0\varepsilon_r A}{h}\right)$$

$$y_j = \begin{cases} 1, & j = 0 \\ 2, & j \neq 0 \end{cases} \quad (10)$$

$$Z(a,h,\varepsilon_r=1) = \frac{377}{2\pi}L_n\left(\frac{f\frac{a}{h}}{\frac{a}{h}}+\sqrt{1+\left(\frac{a}{2}\right)^2}\right) \quad (11)$$

$$f\left(\frac{a}{h}\right) = 6+(2\pi-6)\exp\left(-\left(\frac{30.666}{\frac{a}{h}}\right)^{0.758}\right) \quad (12)$$

The formula of C_{dyn} is used to determine the capacity C and to determine the inductance L we use the following equations:

$$w_{res} = 2\pi f_r \quad (13)$$

$$w_{res} = \frac{1}{\sqrt{LC}} \Rightarrow L = \frac{1}{w_{res}^2 C} \quad (14)$$

Equation (15) allows us to calculate the inductive reactance of coax, taking d0 the diameter of the probe:

$$X_L = \frac{377 fh}{C_0} L_n\left(\frac{C_0}{\pi f d_0\sqrt{\varepsilon_0}}\right) \quad (15)$$

III. Antenna Structure and Design

The geometry of the proposed antenna is shown in figure 1. The square patch antenna was truncated to create a circular polarization (CP). The proposed antenna is printed on a FR-4 substrate of relative permittivity ε_r = 3.5 and thickness h= 3.2 mm and fed by a coaxial cable. Many studies have practiced this mode of feeding [13]-[15]. The substrate is stacked with two layers of FR-4 to allow for greater bandwidth, higher gain and efficiency radiation. In order to increase the impedance and the bandwidth of S11, the truncated square patch antenna has been loaded by a U-slot which introduces a capacitance making it possible to eliminate the inductance due to the vertical feed probe.

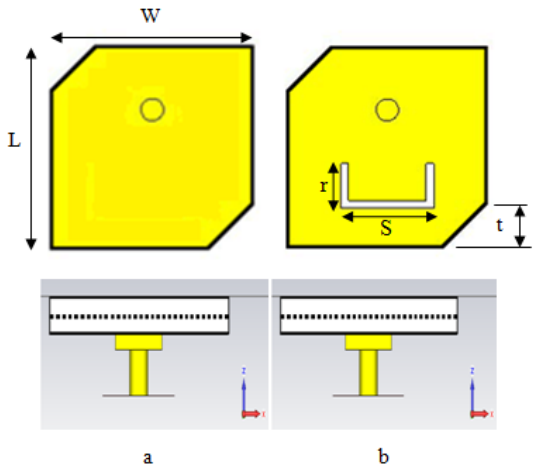


Fig. 2. Design for U-slot Patch Antenna (a): without the slot, (b): with the slot

TABLE. I. Optimized Parameters of the Proposed RFID U-Slot Patch Antenna

Parameters	t	r	s	W	L	ε_r	h
Value(mm)	5	6	12	25	25	3.5	3.2

The truncations made on the square antenna make this element the most difficult to conceive. When taken individually, it can be modeled by a disturbed cavity. The latter will induce a resonant mode perpendicular to that existing without the truncations. To obtain a circular polarization, these modes must have the same amplitude and be out of phase by 90°. The study of a truncated cavity was carried out by Haneishi et al., [16] who set up an equivalent circuit In the case of an antenna fed by a coaxial cable, the resonance frequency of the orthogonal modes (f_{rl} and f_{r2}) is calculated as a function of the surface of the non-truncated antenna (T) and the total truncated surface (Δt) The relation between all these parameters is given by [16]

$$\left|\frac{\Delta_t}{T}\right| Q_0 = \left|\frac{\Delta_t}{T}\right| \frac{f_0}{\Delta_f} = \left|\frac{\Delta_t}{T}\right| \frac{f_0}{f_{r1} - f_{r2}} = \frac{1}{2} \quad (16)$$

Q_o is the quality factor of the cavity and f_0 is the resonance frequency of the undisturbed antenna.

IV. Results and Discussion

Characteristics of the proposed patch antenna were simulated in this section using the CST software. The simulated curves of return loss as a function of the frequency by varying the section "t" of the proposed antenna are shown in figure3. It is noted that the proposed antenna with t = 5 mm gives the best return loss to the antenna desired frequency. Therefore, we fixed t = 5mm and cut a U-shaped slot in the radiating element of the proposed antenna to observe the variations of the return loss. The simulated results are presented in figure4. From this figure it is easy to notice that the desired frequency at 2.45GHz of the proposed antenna is obtained by the slotted structure.

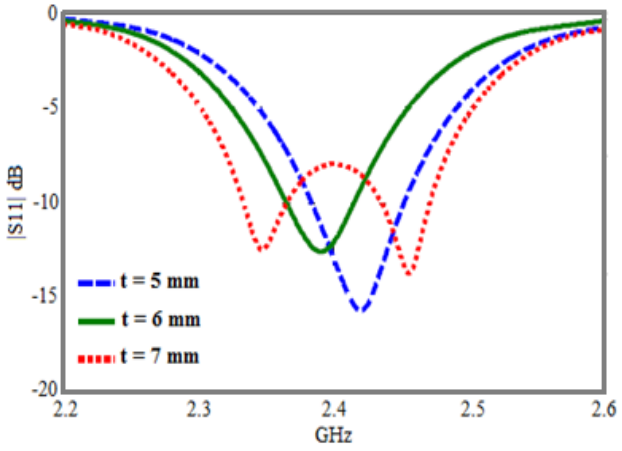


Fig. 3. Simulated return loss of square patch antenna with varieties value of (t)

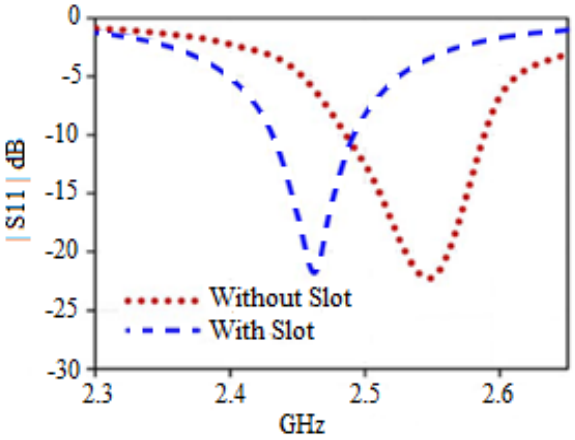


Fig. 4. Simulated return loss for the proposed antenna with and without the slot (with CST)

Figure 5 shows the response of the reflection coefficient of the proposed antenna obtained from the simulation CST and HFSS with respect to the calculated response of the equivalent circuit model.

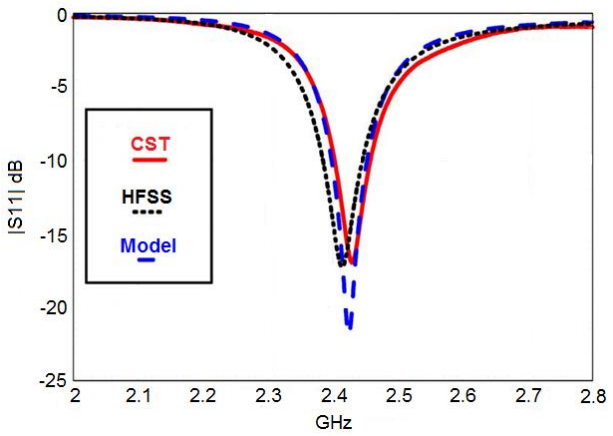


Fig. 5. Simulated return loss for the proposed antenna

Figure 5 shows the reflection coefficient S11 of the proposed antenna obtained from CST and HFSS relative to that obtained from the equivalent circuit using ADS. In Table II the detailed characteristics of these responses are listed. It can be noted from Table II that the requirement of the SHF band 2.45 GHz is satisfied. The results simulated by CST and HFSS agree with that obtained from ADS with some slight differences. This difference is due to the fact that the structure simulator in the CST and HFSS software accounts for all the coupling effects in the simulated antenna physical structure whereas in the equivalent circuit model only the individual elements are taken into account without taking into account the coupling between them.

TABLE. II. COMPARISON OF THE BAND CHARACTERISTICS OBTAINED FROM THE THREE MODELING METHODS

	S11 (dB)	Start freq (GHz)	End freq (GHz)	Center freq (GHz)	BW (GHz)
CST	-18	2.40	2.47	2.45	0.07
HFSS	-18	2.38	2.44	2.41	0.06
ADS	-22	2.39	2.46	2.43	0.07

Figure 6 shows the axial ratio as a function of the frequency. We chose the criterion RA < 3 dB to measure the bandwidth. Note that the bandwidth is also very low. In effect, its value is only 0.30%. This low value comes from the use of truncated corners. Sharma [1] showed that this type of antenna could not provide a large bandwidth at the axial ratio.

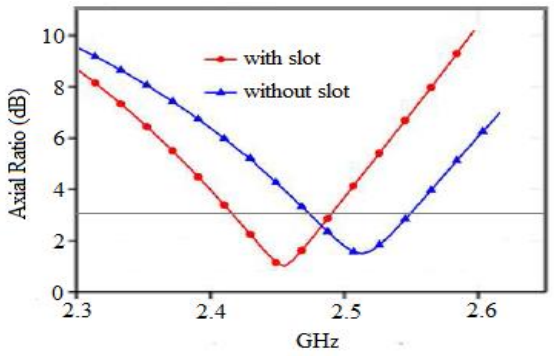


Fig. 6. Simulated axial ratio for the proposed antenna

Figure 7 shows the 3D radiation pattern of the proposed antenna for the resonant frequency 2.45GHz. We note that for this frequency, we have a directional diagram, the efficiency of the radiation which equals 96.7% and the total efficiency is 93.3%. The gain obtained is 4.29 dB. The HPBW value is 92.0 Deg.

The linear and nonlinear gain of theta / phi 0° in polar form are represented respectively by figure 8.

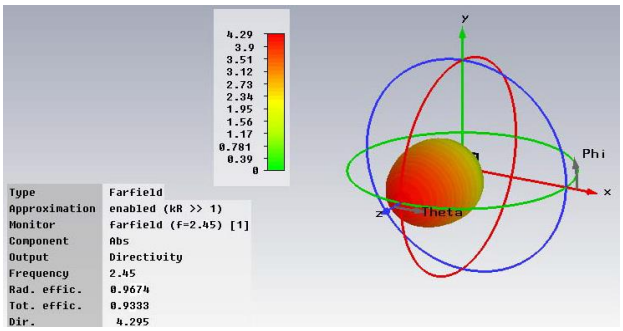
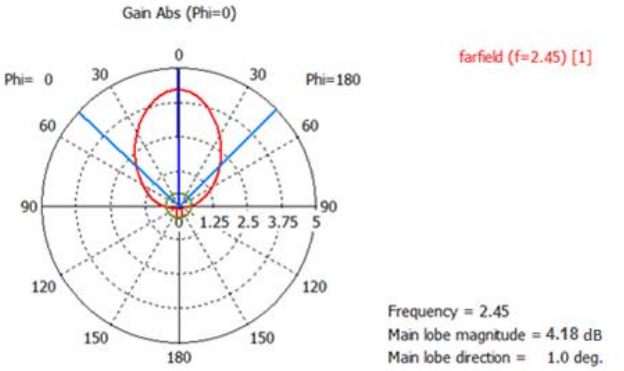
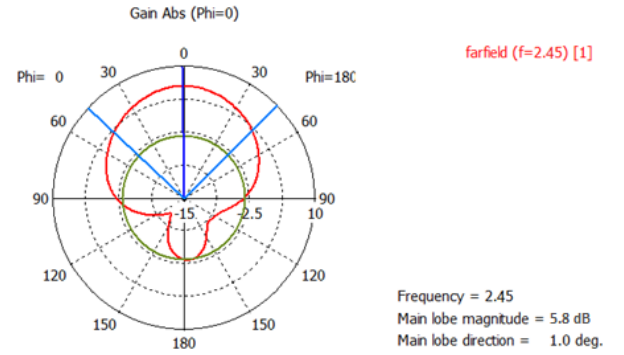


Fig. 7. 3D Farfield for U-slot antenna



(a)

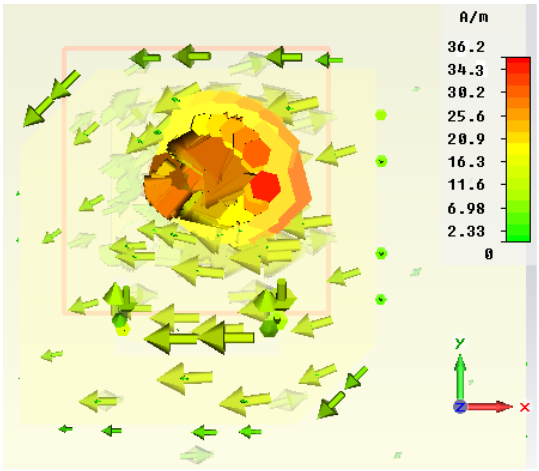


(b)

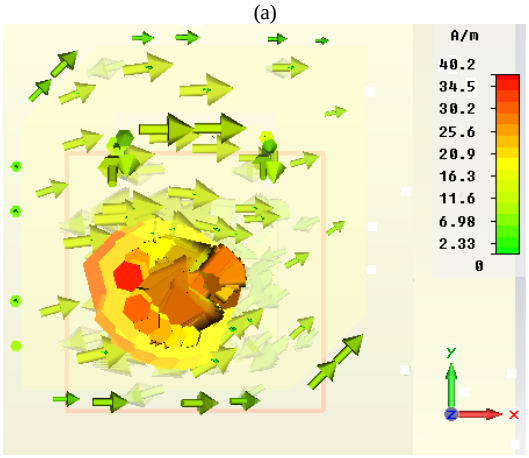
Fig. 8. The polar pattern: (a) Linear and (b) Non-linear

Figure 9 shows the propagation of current from the coaxial feed in the patch. Indeed, by truncating the patch antenna, it generates a circular polarization (CP). Note that when the upper right part and the lower left part of the resulting truncated refers to the right circular polarization, RHCP. The

antenna can work with truncated LHCP at another diagonal axis.



(a)



(b)

Fig. 9. Surface current at U-slot Patch Antenna, (a): LHCP and (b): RHCP

V. CONCLUSIONS

Evaluation and study of antennas is based on their properties and characteristics. The latter vary from one application to another. According to the requirements defined by the application, the antenna design method can be chosen. Thus, each method has its advantages in well-defined cases. For our application, we chose the method of electrical modeling for the analysis of square patch antenna structures referred to RFID readers.

The proposed antenna is circularly polarized using the truncation method. Indeed, the LHCP and RHCP polarization depend on the corners of the diagonal axis which are truncated. This truncated square patch antenna with U-shaped slot has been designed to achieve high bandwidth, high radiation efficiency and resonant frequency equal to 2.35 GHz.

The modeling method that we have applied in this work has several advantages. In fact, it allows to simulate structures in a simple, fast and efficient way. In addition, once the model is built, we can track changes in antenna parameters based on changes in antenna geometry, position and nature of excitation. So we can integrate these antenna structures easily into RFID systems.

REFERENCES

[1] X. Ye, M. He, P. Zhou and H. Sun, "A Compact Single-Feed Circularly Polarized Microstrip Antenna with Symmetric and Wide-Beamwidth Radiation Pattern." International Journal of Antennas and Propagation Volume 1, 2013.

[2] M. F. Jamlos, M. R. Kamarudin, M.A. Jamlos and M. Jusoh, " A novel reconfigurable quadratic antenna for wimax and 4G systems" MICROWAVE AND OPTICAL TECHNOLOGY LETTERS, Pages: 416–421, Volume 54, Issue 2, February 2012

[3] M. Jusoh, M. F. Jamlos and M. R. Kamarudin, "A compact dual bevel planar monopole antenna with lumped element for ultra-high frequency/very high frequency application", MICROWAVE AND OPTICAL TECHNOLOGY LETTERS, Pages: 156–160Volume 54, Issue 1, January 2012

[4] M. Jusoh, M. F. Jamlos, M. R. B. Kamarudin, and M. F. b. A. Malek, "A MIMO antenna design challenges for UWB application, " Progress In Electromagnetics Research B, Vol. 36, 357-371, 2012.

[5] W. LUI, C. CHENG and H. ZHU, "Improved frequency notched ultra wideband slot antenna using square ring resonator", IEEE Trans. Antennas Propag., vol 55, No. 9, pp. 2445–2450, 2007.

[6] A. Farswan, A. K. Gautam, B. K. Kanaujia and K. Rambabu "Design of Koch Fractal Circularly Polarized Antenna for Handheld UHF RFID Reader Applications", IEEE Transactions on Antennas and Propagation , Vol 64, No 2, Feb. 2016.

[7] S. Lakrit, H. Ammor and J. Terhzaz , "Design of H-slot Patch Antenna for Ultra Wideband", European Journal of Scientific Research , Vol 106 No 2, pp:224-228, 2013.

[8] R. Zaker., C.Ghobadi and J.Nourinia, "Bandwidth enhancement of novel compact single and dual band-notched printed monopole antenna with a pair of L-shaped slots", IEEE Trans. Antennas Propag., vol 57, No. 12, pp. 3978–3983, 2009.

[9] N. Prombutr, P. Kirawanich, and P. Akkaraekthalin " Bandwidth Enhancement of UWB Microstrip Antenna with a Modified Ground Plane" International Journal of Microwave Science and Technology, Volume11, 2009 .

[10] K. Finkenzeller, "RFID HANDBOOK", Second Edition, 2003.

[11] Guy A. E. Vandenbosch and Alexander Vasylchenko, "A Practical Guide to 3D Electromagnetic Software Tools, " InTechEUROPE pages 504 – 541,

[12] B. Bieda and P. Słobodzian, "Efficiency of the IE-MoM Approach in the Analysis of Dielectric Bodies Embedded in a Cavity, " Microwave Radar and Wireless Communications (MIKON), 18th International Conference, 2010.

[13] A. Mak, C. R. Rowell, R. D. Murch, and M. Chi-Lun, "Compact multiband planar antenna for 2.4/3.5/5.2/5.8-GHz wireless applications," IEEE Antennas Wireless Propag. Lett., vol. 11, pp. 144–147, 2012.

[14] D. C. Chang, B. H. Zeng, and J. Liu, "CPW-fed circular fractal slot antenna design for dual-band applications," IEEE Trans. Antennas Propag., vol. 56, no. 12, pp. 3630–3637, Dec. 2008.

[15] B.D. Bala, M.K.A. Rahim, , N.A. Murad, M.F. Ismail and H.A. Majid, "Design and Analysis of Metamaterial Antenna Using Triangular Resonator, " Microwave Conference Proceedings (APMC), Asia-Pacific, 2012.

[16] M. HANEISHI and S. YOSHIDA, "A Design Method of Circulary Polarized Rect angular Microstrip Antenna bp One-Point Feed." Electronics and Communications in Japan, 64B, 46-54. 1981.

[17] Nasimuddin and A. K. Verma Amir "Fast and accurate model for analysis of equilateral triangular patch antenna", Journal of Microwaves and Optoelectronics, Vol 3, pp 99-110, April 2004.

[18] S. Malisuwan, M. Charoenwattanaporn and Ut Goenchanart " microstrip antenna Antenna for wireless lan by Appling modified smithchartrepresentation," International journal of the computer, the Internet and management, Vol 11, pp 34-44, 2003.

Modified Hierarchical Method for Task Scheduling in Grid Systems

Ahmad Ali AlZubi
Computer Science Department
King Saud University
Riyadh, Saudi Arabia

***Abstract*—This study aims to increase the productivity of grid systems by an improved scheduling method. A brief overview and analysis of the main scheduling methods in grid systems are presented. A method for increasing efficiency by optimizing the task graph structure considering the grid system node structure is proposed. Task granularity (the ratio between the amount of computation and transferred data) is considered to increase the efficiency of planning. An analysis of the impact on task scheduling efficiency in a grid system is presented. A correspondence of the task graph structure considering the node structure (in which the task is immersed) to the effectiveness of scheduling in a grid system is shown. A modified method for scheduling tasks while considering their granularity is proposed. The relevant algorithm for task scheduling in a grid system is developed. Simulation of the proposed algorithm using the modeling system GridSim is conducted. A comparative analysis between the modified algorithm and the algorithm of the hierarchical scheduler Maui is shown. The general advantages and disadvantages of the proposed algorithm are discussed.**

Keywords—directed acyclic graph (DAG); task granularity; hierarchical method; Maui scheduler; scheduler; scheduling algorithm; task manager; grid; parallelism degree

I. Introduction

Planning and resource allocation in grid systems are crucial tasks due to the heterogeneous structure, large dimensionality and different types of problems encountered [1]. A grid system typically consists of K computed nodes {ri |i=1,2..,K}. Each node ri includes a plurality of Pi={pj | j=1,2,…Ni) processors, the relations between which is given by the loaded Hi=(Bi,Li) graph. A vertex set Bi={bj | j=1,2,…Ni) represents the grid system node processors, and a plurality of ribs Li={lk,j | k,j=1,2,…Ni) of the graph indicates the relationships among the processors. Each vertex bj ∈ Bi has a weight vj equal to the performance of the corresponding CPU pj ∈ Pi. The performance of the complete grid system of the i-th node is equal $V_i = \sum_{j=1}^{N} v_j$. The weight si,j of ribs li,j determines the transmission speed of the communication channel between processors pk and pj. Si is the exchange rate within the i-th grid system node.

In the general case, the task scheduling process in a grid system, which consists of a plurality of computing nodes, is performed as follows: for a grid system consisting of K computing nodes, find a node that provides the optimal solution for the problem in accordance with predetermined criteria.

Depending on the choice of optimization criterion, the problem of finding an optimal node can be formulated as follows:

Find the i-th node of the grid system that provides the minimum time to complete task Ti. Mathematically, this problem can be written as follows:

$$\min_{i=1,K} \{T_i\} \tag{1}$$

$$T_i = \sum_{j=1}^{m} \sum_{l=1}^{P_i} t_{jli} \times X_{jli} + S_i + max\{Tr, Tf_i\} \tag{2}$$

where t_{jli} is the run time of j-th task in the l-th CPU of the i-th grid system node;

S_i is the delivery time of the input data and application results to (from) the i-th grid system node;

Tr is the time when the task is ready to execute in the grid system nodes;

Tf_i is the time then the i-th grid system node is released to perform the task in exclusive mode;

$X_{jli} = 1$ if the j-th task executes in the l-th CPU of the i-th node and $X_{jli} = 0$ otherwise.

Find the i-th node of the grid system with minimal computation cost that performs a given application within a given time (Tz). The mathematical model of this task can be written as follows:

$$\min_{i=1,R} \{C_i\} \tag{3}$$

under condition

$$T_i \le T_z (i = 1, K) \tag{4}$$

In formula (3) Ci is the task execution cost in the i-th node of the grid system. In this case, a subset of nodes is first determined; runtime of these nodes corresponds with restriction (4). A node in which a task is executed at a minimal cost is then selected among this subset (condition (3)).

Find the i-th node of the grid system with the lowest cost that will provide the minimal total execution time of the task. The mathematical model of this task can be written as follows.

Find a grid system node that satisfies conditions (1) and (2) under the restriction

$$C_i \le C_{\min} \tag{5}$$

In this case, initially, in accordance with conditions (1) and (2), the subset of nodes is determined while ensuring that the execution time of the application is minimal. Among them, the node with the lowest cost then is selected in accordance with restriction (5).

II. Methods and Algorithms for Task Scheduling

Three main types of scheduling methods are used in grid systems: centralized, decentralized and hierarchical [2].

In centralized methods, all user tasks are sent to a centralized scheduler. The centralized scheduler forms a unified incoming task queue. The advantage of such methods is their high planning efficiency because the planner has the information of all available resources and the coming challenges. The disadvantage of centralized scheduling is weak scaling. Centralized methods are only suitable for grid systems with a limited number of nodes.

In decentralized methods, the planning function is distributed across all system nodes. Decentralized methods provide better fault tolerance and reliability compared with centralized methods; however, the absence of a meta-scheduler that has information about all tasks and resources reduces the scheduling efficiency.

Hierarchical methods of the task planning process are subdivided into two levels: global and local. The functional components of the task scheduler are associated with two simultaneous types of data flow: information flow of user tasks and control task flow.

At present, task scheduling in grid systems is mainly performed by hierarchical schedulers due to the large number, dimensionality and heterogeneity of tasks. The effectiveness of the hierarchical scheduling method depends on the efficiency of its software implementation, the planning strategies of low-level grid system brokers and the local scheduler.

In [3,4], a review and analysis of the main scheduling methods in grid systems were conducted. Task scheduling in grid systems is an NP-complete problem [5,6], and the solution has different approximate methods and algorithms, such as heuristic algorithms [7], genetic algorithms [8,9,10], algorithms based on stochastic Petri networks [11], ant colony algorithms [12], fuzzy optimization [13], tabu search [14], gravitational emulation local search [15], learning automata [16] and combinations thereof [17,18,19,20]. In the general case, task scheduling is a multi-objective problem in grid systems. Over the last decade, significant research has been carried out in the field of task planning for distributed and parallel systems from the standpoint of minimizing task execution time and calculation cost and optimizing resource utilization [21], security [22] and fault tolerance [23]. In [24], the different scheduling algorithms were summarized based on the grid system structure, showing that the minimal computation value is achieved by a combination of genetic algorithms and other types of algorithms.

Grid systems are used to solve the problems of high-dimensional serial tasks, parallel tasks and parallel–serial tasks. Task sequences are applications that require a single processor for serial operations. Task sequence planning is performed by a single computing unit via algorithms such as Min–Min, Min–Max, and Sufferage [1], which do not provide parallel operations.

Parallel tasks involve the use of multiple processors for the simultaneous execution of operations. The development of computer technology for large-scale problem solving in grid computing is a rapidly developing area and is presented in the form of a workflow of series–parallel tasks with a specific chart of computing synchronization [25]. The computational tasks are represented in the form of a directed acyclic graph (DAG) [26, 27, 28, 29]. The presence of parallel branches in a DAG facilitates the simultaneous use of multiple grid system resources for task execution. In this case, is it crucial to minimize the cost of data transfer among computing grid system nodes. Ref. [30] provides a method for scheduling tasks that considers task granularity [31]—the ratio of computation operations to the volume of transferred data. This increases the efficiency of planning parallel–serial tasks in a grid system. A further increase in the efficiency of scheduling can be achieved by optimizing the structure of the DAG task with the structure of grid system nodes, particularly their granularity. Grid system node granularity is the ratio of node performance to the exchange rate among its components (CPUs).

III. Analysis of the Influence of Granularity on the Effectiveness of Task Planning in Grid Systems

Let us represent the computational task as a DAG: D=(A,E), vertex set A={aj | j=1,2,…M) that represents part of the tasks, and a set of arcs E={ei,j | i,j=1,2,…N) that represents the link between tasks. For each vertex aj ∈ A, its weight wj is given, which is equal to the number of operations performed by the current task. The total number of task operations is $W = \sum_{j=1}^{M} w_j$. Weight qi,j of ribs li,j determines the amount of data transferred among the tasks over the communication channel between CPUs pi and pj. The total amount of data transmitted in solving the task is presented in the form of a DAG: $Q = \sum_{i=1}^{M} \sum_{j=1}^{M} q_{i,j}$

The efficiency of task parallelization depends on the number of calculations and the amount of data transmitted:

$$E_T = \frac{W}{W+Q} \tag{6}$$

Let us represent formula (1) in the form of

$E_T = 1 - \frac{1}{1+W/Q}$,

or

$$E_T = 1 - \frac{1}{1+G_T}, \tag{7}$$

where GT =W/Q is the task granularity.

With increasing task granularity, the effectiveness of its implementation also increases. Reducing the amount of transmitted data required for the problem leads to increased efficiency and granularity. Thus, granularity can be a criterion for the efficiency of parallelization.

The task computation efficiency in the grid system node is determined by the ratio of the task calculation time tT to the exchange time between tasks tC:

$$En= tT / tC.. \quad (8)$$

Substituting the expressions tT=W/V and tC=Q/S into expression (8) yields the following:

$$E_n = \frac{W \times S}{V \times Q}$$

or

$$En =GT / Gn, \quad (9)$$

where Gn =V /S is the granularity of the grid system node.

Formulas (7) and (9) indicate that to achieve the maximum task scheduling efficiency in the grid system, it is necessary to choose the ratio between task granularity and node granularity at which the task is immersed for calculation.

Selecting a node for task immersion will primarily depend on the maximum task granularity at which the condition will be executed (4). The granularity is increased by clustering the DAG task [32]. Thus, adjacent DAG vertices should be combined with a maximum amount of transferred data.

In the case of an absence of grid system computational nodes that allow calculations to be performed in a cluster within a given period of time or a lack of available computational resources, DAG task declustering is performed. In declustering, the number of DAG task vertices increases, but their weights decrease. Thus, the task granularity is reduced by decreases in the weights of DAG vertices, increasing the amount of data transferred between them. This leads to decreased coefficients ET and En. It is appropriate to reduce the task granularity when GT > Gn.

Increasing the number of DAG vertices allows for an increased degree of parallelism of the task and reduces the time of its decision. The following condition must be satisfied:

$$K(t) \geq D(t), \quad (10)$$

where D(t) is the parallelism degree of the task;

K(t) is the number of available CPUs at time moment t.

The parallelism degree of task D(t) is the number of CPUs involved in solving the task at time moment t.

One of the key elements in achieving high performance in task planning is selecting an appropriate ratio between the task granularity and node granularity of the grid system on which it is immersed. Conditions (4) and (10) must be met.

IV. Modified Hierarchical Method for Task Scheduling Considering the Granularity of Tasks and Grid System Nodes

A. Difference between the Modified Hierarchical Method for Task Scheduling and the Base Method of the Maui Scheduler

The Maui hierarchical scheduling algorithm is examined as a base scheduling algorithm for review and modification [33]. The Maui scheduler is an optimal configurable tool that supports multiple resource selection policies and is able to set dynamic priorities, enforce “fair” sharing of resources between users, and facilitate reservation.

The Maui scheduler is one of the most popular and effective grid meta-schedulers and is used in many implementations of grid systems, such as IBM Tivoli [34] and Moab Workload Manager [35].

In the planning process, the Maui scheduler performs the following:

- full list view of nodes in order of the optimal resource search (best fit);
- preliminary calculation of time for solving the task on all nodes.

The Maui scheduler uses task granularity as the primary metric for choosing the node for task immersion, thus eliminating the time-consuming search operation for the optimal resource.

The elimination of the optimal resource search operation also entails the elimination of the time-consuming operation for estimating the task time execution on a resource for each node, thus significantly reducing the planning time.

Instead of searching the list of system resources in the search for a suitable resource, it will browse the resources in the ring as long as the desired resource is not found. Once the desired resource is found, the resource will be implemented for immersion. The subsequent search for a resource starts from a list of resources after previously finding the desired one that has been utilized for immersion. The browsing is executed nonlinearly by ring, which does not lead to a linear increase in algorithm complexity in the case of an increasing number of resources. Additionally, this approach leads to a more balanced loading of the system because all components will be reviewed and will not be permanently assigned for the high-performance resource tasks only. These steps help to improve the efficiency of the scheduling procedure in a grid system.

To select appropriate resources for the task, some of the functions will be transmitted to the resource manager, which has general resource information: the number of CPUs, the performance of a node, and the communication channel capacity of this node. Therefore, in determining the

appropriate resources for the task, the scheduler will request information about the availability of nodes from the manager at the time of the transfer of the current task data. The resource manager returns to the scheduler sub-list of nodes that are available at this moment. The scheduler will determine resource searching with a specific granularity value at a certain stage of the algorithm execution among the list of system resources but only on the resources that will be able to take the tasks immediately and execute it immediately after the allocation of the task to the resource. This approach significantly reduces the decision time regarding whether the resource is suitable for the task. Because of the amount of time during which the response request/receipt are executed from the resource manager, the planning time should not increase significantly relative to the decision time of the base algorithm.

For the resources, let us determine another parameter—resource holding time. Considering this parameter will allow us to avoid another disadvantage of the previous algorithm—the possibility of assigning tasks to a resource that has not yet been freed, which can lead to the formation of local queues to resources. The scheduler calculates this parameter during the task immersion to a resource based on the number of calculations in the task, the capacity of the node channel, the task weight and the node performance. The parameter is stored for the certain resource and provided by the request to a scheduling manager under the condition of free resource availability at the required start time of the task. Based on this parameter, the capacity of the node channel, and the number of calculations in the new task—which will be given to the scheduler—the manager will be able to calculate whether any given resource is available to the point where it will be passed to the data.

The ratio selection between the granularity GT of a task and the granularity Gn of a grid system node on which it is immersed is performed as follows. For the selected task to be executed on the grid system, the search is performed for a node, and the granularity Gn of each node is equal to granularity GT according to condition (4). In this case, the task is immersed on the selected node; otherwise, task granularity GT is corrected depending on its ratio with system node granularity Gn. If node granularity Gn is greater than task granularity GT, then clustering increases GT to a value as close to Gn as possible by condition (4). Then, the immersion on the selected node for its implementation is performed. If node granularity Gn is less than task granularity GT and condition (4) is not executed, then task granularity GT is reduced to meet conditions (4) and (10).

B. Algorithm of the Modified Task Scheduler Function

1. Begin;
2. creation of the task queue;
3. if the task queue is empty, then go to step 19;
4. the selection of the next task;
5. create an available node list at the task downloading time;
6. if the node list is empty, then go to step 5;
7. the selection of node ri c min |GT – Gn| and Ti ≤ TZ /* selection of the node that is most appropriate for criteria ET and En */;
8. if 0.5≤ (GT| Gn) ≤1.5, then go to step 17;
9. if GT < Gn, then go to step 12 /* task granularity smaller than node granularity */;
10. if GT is minimal, then go to step 17 /* further declustering impossible */;
11. decrease of GT, then go to step 8 /* performed by clustering */;
12. if GT maximum, then go to step 17 /* further clustering violates the condition: Ti ≤ TZ */;
13. increase of GT /* performed by clustering */;
14. the calculation of a Ti new value;
15. if Ti ≤ TZ, then go to step 8;
16. decrease of GT /* performed by declustering */;
17. task immersion on node ri;
18. go to step 2;
19. End.

V. Analysis of the Algorithm's Effectiveness for the Process of Scheduling Tasks in a Grid System

A. Simulation of the Task Scheduling Process

The GridSim [36] modeling system, which allows different scheduling policies to be implemented (FCFS, Easy Backfill, Conservative Backfill) is selected as a tool for modeling and analyzing the effectiveness of the proposed algorithm. In the current research, GridSim has been expanded by adding new necessary entities for the simulation of the planning process and execution of workflows in the grid environment. The class diagrams of the implemented modules are shown in Figure 1.

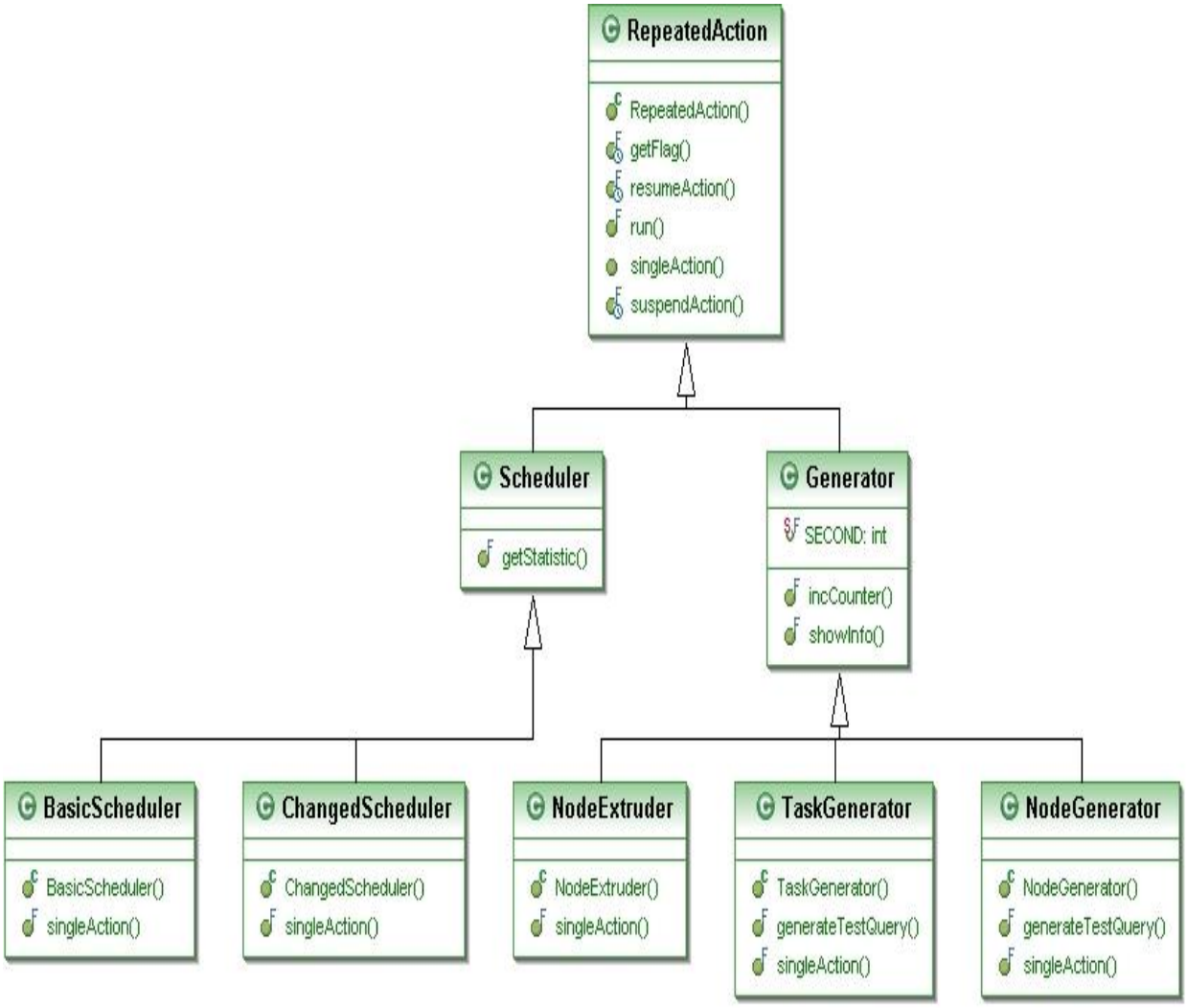


Fig. 1. Class diagram in the modeling system

A modeling system allows for test task immersion in real time. The number of tasks and resources are defined by the user. The result of both scheduling algorithm simulations, with the same input data, are output data, such as average task downtime in the queue, total system boot time, average system node load, average load of communication channels, average scheduler decision time, and system node load chart. The node parameters are the number of CPUs and the performance of the node channel bandwidth of that node.

The task is generated with the granularity, the task weight (the amount of calculations), the estimated runtime for a task at its maximum granularity, and the priority. The task queue is created after generation, sorted by priority, in which the task with the highest priority is at the head of the queue. The scheduler works in real time; all measurements are made in milliseconds.

B. Simulation Results of the Base and Modified Scheduling Algorithm

The modeling system generated loading charts of grid system nodes and communication channels. Using this simulation program, the loading of system nodes were analyzed at different ratios between the task number and grid system nodes. Figures 2 and 3 show the relative loading of the first 25 nodes and the communication channels as a percentage of their maximum values in the solution of 100 tasks on 50 grid system nodes.

A comparison of Figures 2 and 3 illustrates that the loading of the nodes is relatively low with a relatively small difference between the number of tasks and nodes. Furthermore, the modified algorithm provides more balanced and larger loading compared with the baseline algorithm.

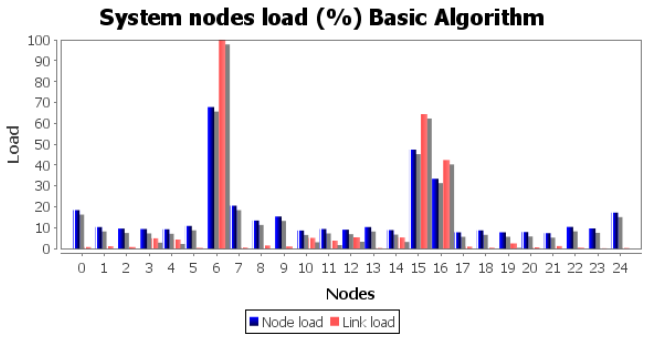


Fig. 2. Relative loading of the first 25 nodes and communication channels as a percentage of their maximum values in the solution of 100 tasks on 50 grid system nodes using the base scheduling algorithm

Figures 4 and 5 show the relative loading of the first 25 nodes and communication channels as a percentage of their maximum values in the solution of 5,000 tasks on 100 grid system nodes.

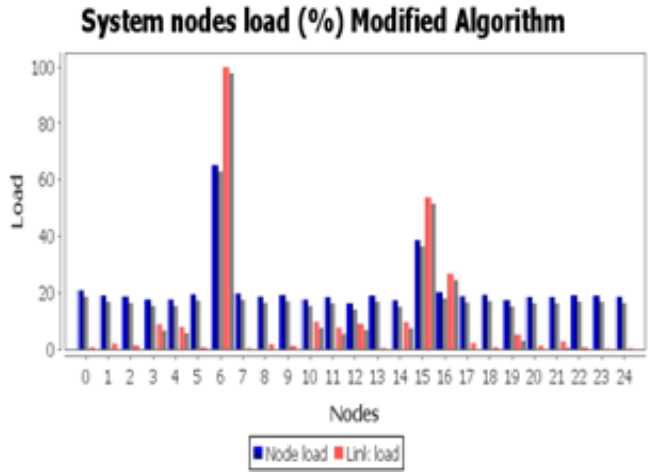


Fig. 3. Relative loading the first 25 nodes and communication channels as a percentage of their maximum values in the solution of 100 tasks on 50 grid system nodes using the modified scheduling algorithm

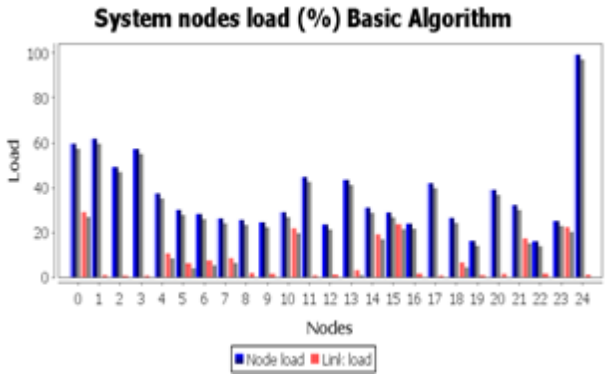


Fig. 4. Relative loading of the first 25 nodes and communication channels as a percentage of their maximum values in the solution of 5,000 tasks on 100 grid system nodes using the base scheduling algorithm

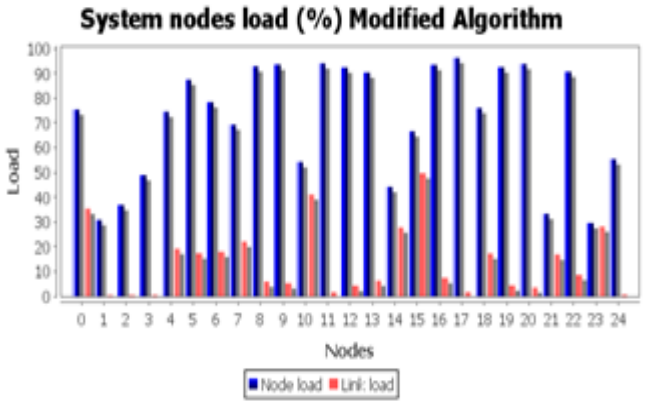


Fig. 5. Relative loading of the first 25 nodes and communication channels as a percentage of their maximum values in the solution of 5,000 tasks on 100 grid system nodes using the modified scheduling algorithm

A comparison of the loading charts (Figures 2–5) with the increase in task queues indicates that the effectiveness of the modified scheduling algorithm is significantly increased due to the higher and more balanced loading of nodes and communication channels.

Experiments were performed for a fixed node number but a variable task number. The base and modified algorithms were modeled, and a histogram was constructed based on the average values for the experiments with tasks as one pair of resources.

Considering how to apply the base or modified scheduling algorithm will influence the average residence time of the task in queue (Figure 6), the total system loading time (Figure 7), and the average scheduler decision time (Figure 8).

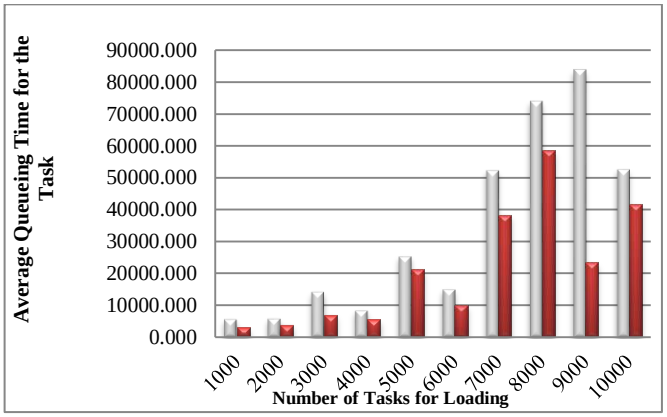


Fig. 6. Average residence time of a task in the queue based on the task number using the base and modified algorithms with a fixed number of resources of 100

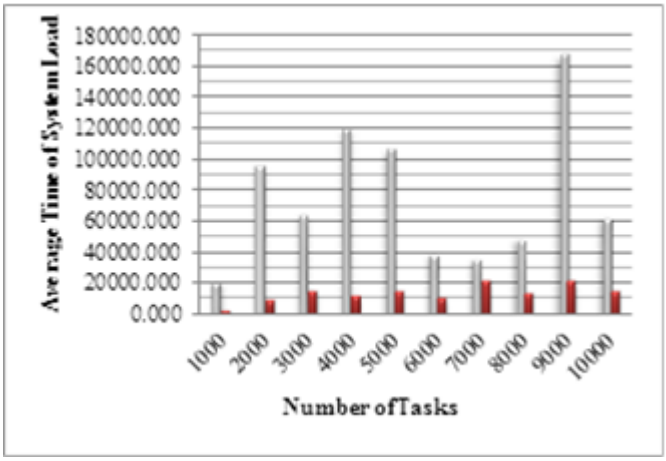


Fig. 7. Total load time of the system with a fixed number of resources of 100

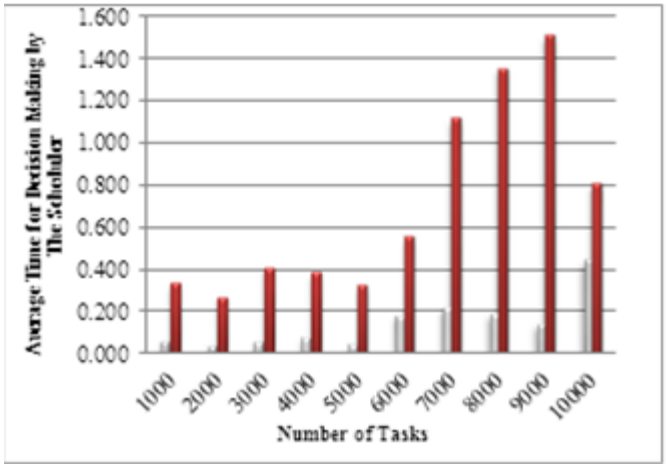


Fig. 8. Average decision-making time by the scheduler with a fixed number or resources of 100

As shown in Figure 6, regardless of the task number, the residence time in the queue is less than that in the modified algorithm by an average of 20%.

As shown in Figure 7, the total system loading time is significantly reduced with the modified algorithm because the modified algorithm selects the optimal ratio between task granularity and system granularity.

The average scheduler decision time regarding the choice of the node for the task immersion using the modified algorithm is essentially independent of the task queue size, unlike the base method. This independence occurs because resource searching continues to loop as long as the desired resource is not found. In the base algorithm, resource searching starts from the beginning each time, which often leads the resource to be linearly dependent on the optimal task number search in the queue.

Figures 9–11 show the simulation results of the base and modified schedulers with a fixed task number and different grid system node numbers. Experiments were performed for a fixed task number but a variable node number. We implemented the model using the base and modified algorithms, and the histograms (Figures 9–11) reflect the average values for the experiments with one pair of resources—tasks.

As shown in Figure 9, the average task residence time in a queue using the modified algorithm with different amounts of resources is approximately 50% less than when using the base algorithm because the decision time using the modified algorithm is less than that of the base algorithm.

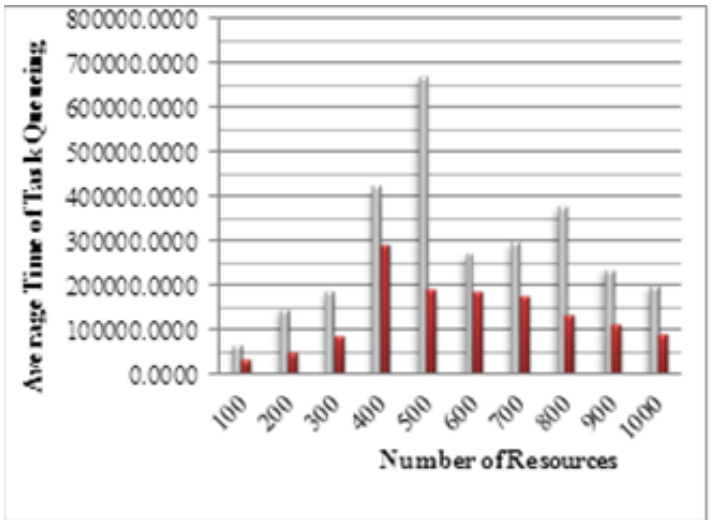


Fig. 9. Dependence of the average task residence time in a queue with a different number of gird system nodes with a fixed number of tasks of 10,000

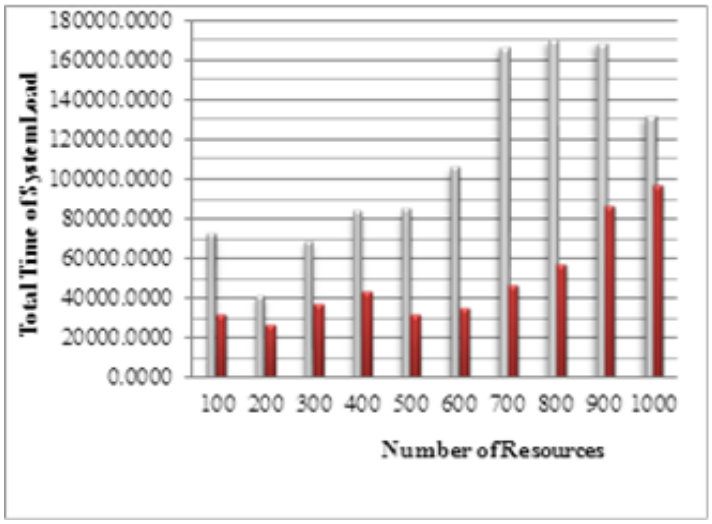


Fig. 10. Total load time of the system with a fixed number of tasks of 10,000

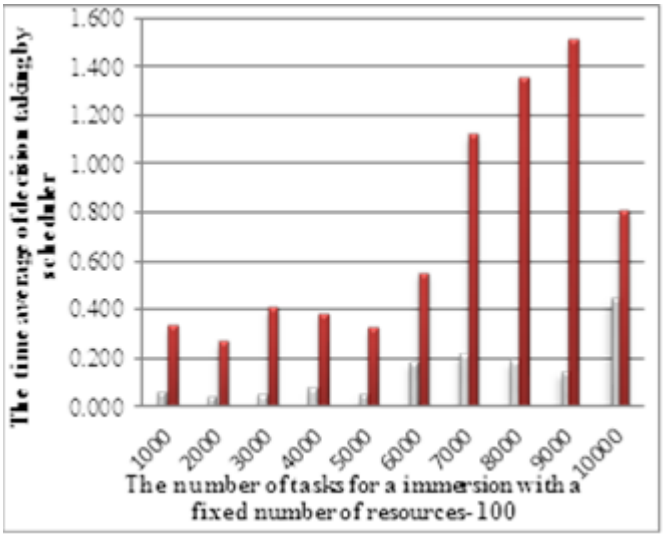


Fig. 11. Average decision-making time by the scheduler with a fixed number of resources of 100

C. Analysis of the Modeling Results

The histogram shows the average task residence time in a queue for an increasing number of tasks and a fixed amount of resources and a fixed number of tasks and an increasing number of resources in proportion to each task and resource pair. Thus, the task downtime in a queue for the base algorithm is larger than for the proposed modified algorithm. When the modified scheduler does not review all resources when searching, it does not calculate the task execution time for each item in the search to achieve the optimal time. The histogram shows the significant time gap associated with the generation of each experience for a different number of resources with different productivity. This time gap can greatly increase the waiting time of task immersion for the resource with the base algorithm.

The histograms show that the system load is reduced when using the modified scheduler because the main load is not in the highest-performing resource and because tasks are evenly distributed across system resources. This reduces the system time and allows many tasks to be executed earlier compared with the base scheduler.

The final histogram shows that using the modified scheduling algorithm increases the decision-making time of the scheduler. Let us analyze why this occurs. The base scheduling algorithm is necessary for a full scan and for calculating the computation time of the task on each resource, which is a time-consuming procedure. In the modified algorithm, a request is sent to the resource manager for the sub-list of resources that will be available at the time of data transfer of the current task. This is the first time value in the total time calculation of decision making. Next, the scheduler waits for a response from the resource manager, which holds the necessary calculations to generate a list; the resource manager then sends a list to the scheduler. After that, the scheduler begins to view the issued list of resources to find the optimal resource for the task. When a resource is found, the scheduler calculates the end of the task on the resource, which also takes time. After these steps, if the base algorithm is being used, the task waits for the most productive resource regardless of how much time the scheduler spends deciding on the most favorable site node. Consequently, the task is idle, and the system is underutilized. This is not observed when the modified scheduling algorithm is used because the system time is considerably lower than that when using the base algorithm even though the decision time is longer.

Table 1 shows the average test results. We generated 10,000 random tasks that were immersed on 1,000 nodes.

TABLE I. RESULTS OF THE ALGORITHMS

Characteristic	Baseline	Modified
Waiting time in the queue	52848	9506
Total working time	17 197 200	4 248 600
Average nodes loading, %	19	45
Average loading of communication channels, %	13	11
Decision time, ms	2,8015	5,2316

VI. CONCLUSION

This paper proposes a modified hierarchical method of task scheduling that increases the efficiency of a grid system by selecting the optimal ratio between task granularity and grid system node granularity on a node on which a given task is immersed. This is accomplished by changing the task granularity via conversion. This ensures a uniform, more balanced load of processors and communication channels between grid system nodes and reduces the residence time of the task in the input task queue. The result is increased productivity in the grid system by an average of 20%.

A further performance increase is related to the possibility of changing the granularity of grid system nodes by changing their structure considering the number of physical communication channels in the processors of a particular computing node and through support for a duplex mode of information transmission in communication channels.

ACKNOWLEDGMENT

This research is supported by King Saud University; Author would express his appreciation to the Deanship of Scientific Research at King Saud University for the provision of funding.

REFERENCES

[1] F. Dong, and G. Selim, “Scheduling Algorithms for Grid Computing: State of the Art and Open Problems (Technical Report),” School of Computing, Queen’s University, Kingston Ontario Rep. 2006–504, 2006.

[2] D. Cokuslu, A. Hameurlain, and K. Erciyes, “Grid resource discovery based on centralized and hierarchical architectures,” Int. J. Infonomics, vol. 3, Jan. 2010.

[3] T. Ma, S. Shi, H. Cao, W. Tian, and J. Wang, “Review on Grid Resource Discovery: Models and Strategies,” IETE Technical Review, Vol. 29, pp. 213-22, 2012.

[4] Mohammed Bakri Bashir, Muhammad Shafie Abd Latiff, Aboamama Atahar Ahmed, Adil Yousif & Manhal Elfadil Eltayeeb, “Content-based Information Retrieval Techniques Based on Grid Computing: A Review”, IETE Technical Review, Vol. 30, pp.223- 232, 2013

[5] H. El-Rewini, T. Lewis, and H. Ali, Task scheduling in parallel and distributed systems. Englewood Cliffs, NJ: Prentice Hall, 1994.

[6] M. L. Pinedo, Scheduling: Theory, Algorithms, and Systems, 5th ed. New York: Springer Verlag, 2008.

[7] F. Xhafa, and A. Abraham, "Computational models and heuristic methods for Grid scheduling problems," Future Generation Computer Systems, vol. 26, pp. 608–621, Apr. 2010.

[8] Y. Zomaya, C. Ward, and B. Macey, "Genetic scheduling for parallel processor systems: comparative studies and performance issues," IEEE Trans. Parallel Distrib. Syst., vol. 10, pp. 795–812, Aug. 1999.

[9] G. Aggarwal, M. Kamboj, C. Singh, and P. Sharma, "A novel resource scheduling algorithm for computational Grid," Int. J. Applied Information Systems (IJAIS), vol. 4, pp. 34–7, Sept. 2012.

[10] R. P. Prado, S. García-Galán, A. J. Yuste, and J. E. M. Expósito, "Genetic fuzzy rule-based scheduling system for grid computing in virtual organizations," Soft Comput., vol. 15, pp. 1255–1271, Jul. 2011.

[11] M. Shojafar, Z. Pooranian, J. H. Abawajy, and M. R. Meybodi, "An efficient scheduling method for Grid systems based on a hierarchical stochastic Petri net," J. Comput. Sci. Eng., vol. 7, pp. 44–52, Mar. 2013.

[12] Y. Yang, G. Wu, J.Chen, and W. Dai, "Multi-objective optimization based on ant colony optimization in grid over optical burst switching networks," Expert Syst. Appl., vol. 37, pp. 1769–1775, Mar. 2010.

[13] C. S. Rao, and D. B. R. Babu, "A fuzzy differential evolution algorithm for Job scheduling on computational grids," Int. J. Computer Trends Technol., vol. 13 , pp. 72–77, Jul. 2014.

[14] B. Ekşioğlu, S. D. Ekşioğlu, and P. Jain, "A tabu search algorithm for the flowshop scheduling problem with changing neighborhoods," Comput. Ind. Eng., vol. 54, pp. 1–11. Feb 2008.

[15] B. Barzegar, A. M. Rahmani, and K. Zamanifar, "Advanced reservation and scheduling in Grid computing systems by gravitational emulation Local search algorithm," Am. J. Scientific Research, no. 18, pp. 62–70, 2011.

[16] J. A. Torkestani, "A new distributed Job scheduling algorithm for Grid systems," Cybernetics Systems, vol. 44, pp. 77–93. 2013.

[17] E. Betzar, A. Xavier, and V. M. Betzar, "Survey on heuristics based resource scheduling in Grid computing," Indian J. Computer Science Engineering (IJCSE), vol. 5, pp. 9–14, Mar. 2014.

[18] G. Guoning T.-L. Huang, and G. Shuai, "Genetic simulated annealing algorithm for task scheduling based on cloud computing environment (Published Conference Proceedings)," in Int. Conf. on Intelligent Computing and Integrated Systems, 2010, pp. 60–63.

[19] Z. Pooranian, A. Harounabadi, M. Shojafar, and J. Mirabedini, "Hybrid PSO for independent Task scheduling in Grid computing to decrease Makespan (Published Conference Proceedings)," in Proc. Int. Conf. on Future Information Technol., Singapore, 2011, pp. 435–9.

[20] Z. Pooranian, A. Harounabadi, M. Shojafar, and N. Hedayat, "New hybrid algorithm for Task scheduling in Grid computing to decrease missed Task", World Academy of Science, Engineering and Technology International Journal of Computer, Electrical, Automation, Control and Information Engineering, Vol:5, 2011, pp. 262–268, 2011.

[21] S. K. Garg, R. Buyya, and H. J. Siegel, "Time and cost trade-off management for scheduling parallel applications on utility grids," Future Generation Computer Systems, vol. 26, pp. 1344–55, Oct. 2010.

[22] M. Khan, "Design and analysis of Security aware scheduling in Grid computing environment" Int. J. Comput. Sci. Inf. Technol. Research vol. 1, pp.42–50, Dec. 2013. Available: www.researchpublish.com

[23] P. Keerthika, and N. Kasthuri, "A hybrid scheduling algorithm with load balancing for computational Grid" Int. J. Advanced Science and Technol., vol. 58, pp. 13–28, 2013.

[24] R. Aron, and I. Chana, "Grid scheduling heuristic methods: state of the Art," Int. J. Computer Information Systems and Industrial Management Applications, vol. 6, pp. 466–73, 2014.

[25] Hirales-Carbajal, A. Tchernykh, R. Yahyapour, J. L. González-García, T. Röblitz, and J. M. Ramírez-Alcaraz, "Multiple workflow scheduling strategies with user run time estimates on a Grid," J. Grid Comput., vol. 10, pp. 325–346, 2012.

[26] Forti, "DAG Scheduling for grid computing systems," Ph.D. Thesis, Dep. Mathematics and Comp. Sci., Univ. Udine, Italy, 2006.

[27] P. Chauhan, and Nitin, "Decentralized Scheduling Algorithm for DAG Based Tasks on P2P Grid," J. Engineering, vol. 2014, pp. 202843, 2014. Available: http://dx.doi.org/10.1155/2014/202843.

[28] F. Pop, C. Dobre, and V. Cristea, "Genetic algorithm for DAG scheduling in Grid environments (Published Conference Proceedings)," in Proc. IEEE 5th Int. Conf. Intelligent Computer Communication and Processing, Cluj-Napoca, 2009, pp. 299–305.

[29] R. Garg, and A. K. Singh, "Adaptive workflow scheduling in grid computing based on dynamic resource availability," Engineering Sci.Technol. Int. J., vol. 18, pp. 256–269, Jun. 2015. Available: http://dx.doi.org/10.1016/j.jestch.2015.01.001.

[30] M. A. Palis, "The granularity metric for fine-Grain real-Time scheduling," IEEE Transactions Comput., vol. 54, pp. 1572–1583, Dec. 2005.

[31] D. Konieczny, J. Kwiatkowski, and G. Skrzypczynski, "Parallel search algorithms for the distributed environments," in Proceedings of the 16th IASTED Int. Conf. Applied Informatics, Garmisch-Partenkirchen, Germany, 1998, pp. 324–327.

[32] Gerasoulis and T. Yang, "A Comparison of Clustering Heuristics for Scheduling Directed Acyclic Graphs on Multiprocessors," J. Parallel Distributed Computing, vol. 16, pp. 276–91, Dec. 1992.

[33] Maui Scheduler™ Administrator's Guide version 3.2, Copyright © 1999-2014, Adaptive Computing Enterprises. Available: http://docs.adaptivecomputing.com/maui/.

[34] Tivoli Workload Scheduler Documentation, Available: http://www-01.ibm.com/software/tivoli/

[35] Moab Workload Manager Documentation, Available: http://www.adaptivecomputing. com/

[36] R. Buyya and M. Murshed, "Gridsim: a toolkit for the modeling and simulation of distributed resource management and scheduling for grid computing," Concurrency and Computation: Practice and Experience, vol. 14, pp. 1175–220, Nov.-Dec. 2002.

Issues and Trends in Satellite Telecommunications

David Hiatt
College of Arts & Sciences
Regent University
Virginia Beach, Virginia, U.S.A.

Young B. Choi
College of Arts & Sciences
Regent University
Virginia Beach, Virginia, U.S.A.

***Abstract*—In this paper we will discuss a bit about satellite telecommunications. A brief introduction and history of satellite telecommunications will be presented. Then a discussion of certain prevalent satellite orbit types will be given, because this is relevant to understanding how certain satellite applications are employed. Various areas of ongoing research in the field of satellite telecommunications, to include bandwidth allocation, satellite constellation design for remote parts of the world, and power generation, among others will be discussed.**

Keywords—satellite communications; telecommunications; satellite orbit types; bandwidth allocation; constellation design; power generation

I. Introduction

It goes without saying that satellites have broad and far reaching applications in many areas. Most of these applications relate to telecommunications, meteorology, military spying, navigation, scientific measurement and ground mapping. Yung-Wey and Tat-Chee remind us satellites have mainly been used in the past for TV and telephone transmissions, "however in recent years, the satellite's role in telecommunication has expanded to provide backbone links to geographically dispersed Local Area Networks (LANs) and Metropolitan Area Networks (MANs)" [1]. The ability to launch a platform into space to carry out certain tasks was a revolutionary idea when it began about 50 years ago and has, in addition to becoming one of the main functions of the space program at the time, it led to advances in communications, military applications, and meteorology among other things. Yung-Wey et. al indicates that as promising as satellites are, "there are several drawbacks such as technological complexity, high costs, and Line-of-Sight (LOS) requirements for reception and transmission" [1]. Even though satellite technology has advanced much in the last 50 years, it continues to advance and progress as new applications are found, speed and data throughput is increased, and ingenious control methods are developed. This paper will seek to give a brief history of satellites and how they are used, and the issues and trends of current satellite telecommunications research.

The remainder of this paper will be organized as follows: Section 2 will give a brief discussion of the history of satellites and their applications so more insight into what is happening can be gained and what the future holds. The following section will give a brief overview of different types of orbits, since this will factor into later discussion of research ongoing with satellites. And then the main body section will discuss several research areas that are ongoing to show where the technology is currently being taken and what the future might hold. The last section will present a conclusion and summary. Through this layout, the need for more research in the area of satellite telecommunications will become clear. There are many areas, as will be presented, where improvements can be made in satellite telecommunications, and improvements to any of these areas will result in advancements in the field. These advancements will be manifested in the way of reduced operating and construction costs as well as improved speed and throughput.

II. Brief History of Satellites

It is well known that the first artificial satellite was Sputnik, launched by the Soviets in 1957. The United States launched their first satellite only a few short months later. According to Slotten, President Kennedy declared that this new technology would be used to create a global communication system using satellites and this became known as the International Telecommunication Satellite Consortium (IntelSat) [2]. Later on, the potential of satellites for telecommunications was first demonstrated when, according to Slotten, "In July 1962, a New York Times correspondent proclaimed that the first live television broadcast from Europe to the United States using the communications satellite Telstar would "rank as one of the magnificent accomplishments in television history."' [2]. Since this time, the United States and the other countries that were part of IntelSat poured money into further telecommunication satellite development. The United States Government Accountability Office states that in the 1970's, the first geostationary satellite was launched and was used for weather observation and forecasting. Since that time, the United States as used two geostationary satellites for this purpose until the present day [3].

III. Types of Orbits

There are three main types of orbits used for satellite telecommunications. LEO, or Low Earth Orbit, is one type of orbit commonly used on satellites today. According to Boriboon and Pongpadpinit, it is a satellite trajectory in which a satellite platform moves very rapidly in a low orbit. A constellation of these satellites will be used because they are so close to the earth that they cannot provide full disk coverage. This closeness is why they move faster than geostationary satellites. But this has an advantage; frequencies can be reused due to the rapid orbits, and thus the capacity of telecommunication use is higher [4]. Pontani points out other advantages of low earth orbit satellites. They are cheaper to build and operate compared to geostationary satellites. They are also cheaper to launch and have lower operating power requirements. Furthermore, they provide much higher resolution images due to being much closer to the earth's

surface. Time delays are obviously smaller as well, which has advantages in telecommunications. For this reason, constellations of LEO satellites are being used more than geostationary satellites these days for telecommunications purposes [5].

Another type of orbit is the geostationary orbit. According to Poole, this is where the satellite orbits at the same angular speed of the earth's rotation, so it stays over the same spot on the equator. This orbital altitude is 35,900 km above the earth's surface. This has advantages of more ground coverage (nearly the entire half of the earth facing the satellite) and continuous dwell time (since the satellite does not move). The drawbacks to this are higher operating costs, lower resolution, and more power needed for transmission due to it being further away.

The last type of orbit to discuss is the highly elliptical orbit. In this setup, according to Poole, the satellite will be at varying distances from the earth and moving at a nonconstant speed. This is because when the satellite travels closer to earth, the speed at which it moves increases due to being pulled by earth's gravity. When it slingshots around and travels away from earth again, it slows down. This highly elliptical orbit enables the satellite to have a long dwell time when it is at apogee (farthest from earth). Thus, by placing numerous satellites in a constellation like this and coordinating their orbits, permanent coverage over a certain area can be achieved without having to use a geostationary or geosynchronous satellite. The following image from Ian Poole at Radio-Electronics.com discusses various types of circular earth orbits and their altitudes.

TABLE. I. TYPES OF SATELLITE ORBITS AND ALTITUTES [6]

SATELLITE ORBIT DEFINITIONS			
ORBIT NAME	ORBIT INITIALS	ORBIT ALTITUDE (KM ABOVE EARTH'S SURFACE)	DETAILS/COMMENTS
Low Earth Orbit	LEO	200-1200	
Medium Earth Orbit	MEO	1200-35790	
Geosynchronous Orbit	GSO	35790	Orbits once a day, but not necessarily in the same direction as the rotation of the Earth – not necessarily stationary
Geostationary Orbit	GEO	35790	Orbits once a day and moves in the same direction as the earth and therefore appears above the same point on the Earth's surface. Can only be above the Equator.
High Earth Orbit	HEO	Above 35790	

IV. ISSUES AND RESEARCH IN SATELLITE TELECOMMUNICATIONS

There is always research going on in the field of satellites, telecommunications, and their applications. One of the areas that has the most research and application is trying to get reliable satellite coverage in remote parts of the world where it is not practical to provide signal via traditional conducted mediums like copper wire or fiber optic cable.

One area of research is in how to get reliable, consistent signal for Internet connectivity into Antarctica. An article by Lee, Wu, and Mortari points out the current problem doing research in Antarctica. That part of the world is a prime area for various kinds of scientific research due to the fact that it is relatively untouched and it is away from population and human activity. Furthermore, the remoteness of Antarctica, coupled with the extremely cold temperatures cause conventional network connectivity to be unavailable. With the important research going on there, it is important to have this connectivity. There are small ad hoc wireless networks that permit connectivity throughout research stations and to the nearby deployed sensors, but not to the outside world. For connectivity to the outside world, people rely on satellites. But the only satellites currently providing coverage there are geostationary satellites and certain iridium low earth orbit (LEO) satellites. The geostationary ones orbit at the equator and thus provide spotty coverage along the coasts of Antarctica, leaving the interior and polar region uncovered. The LEO satellite are optimized for maritime use and not properly positioned to cover most of Antarctica. Thus, the coverage is spotty and not constant due to satellite movement and other conditions. The need for more reliable coverage is clear [7]. Lee, et. al designed a satellite constellation using three satellites that had them arranged such that they were closest to the earth over the northern hemisphere and much farther from the earth over the South Polar region. This resulted in much slower movement over Antarctica and as a result, greater dwell time over that region, which is what they wanted [7]. Lee and his team tested the coverage by evaluating the amount of time that the satellites were over the areas of interest producing acceptable connectivity in the required bandwidth [7]. Using an analysis of 5 selected stations such that one is representing the South Pole and the other four are in the four quadrants, Lee et. al showed that all 5 stations had coverage to at least one of the 3 satellites 100% of the time in a day, and for over 50% of the time each station could communicate with two of the satellites. This was a very promising result with only a limited number of satellites [7].

Another area of research in the area of telecommunications satellites concerns the generation of electricity on them in a cost effective way that does not make the satellites too heavy. The reason weight is a concern is because it costs more money to launch a satellite in a rocket the heavier it is. In a paper about this topic, Geneste discusses a proposed way to generate electricity on a satellite that is an alternative to the traditional photovoltaic cells. These are the cells traditionally known as solar panels, and have been the go-to power source for satellites for decades. According to Geneste, the costs are typically high because space users are rather limited in what they can use to power satellites beyond a standard primary battery. He discusses ongoing research in a thermal acoustic engine. This is a version of a standard Stirling Engine. A Stirling Engine would not be appropriate for a satellite because satellite designers typically do not like to have moving parts on a satellite. A thermal acoustic engine would be a design that has a sound wave play the part of the piston. As he explains it,

"the linear alternator on the left generates a primary sound which is thermally amplified and feeds the linear alternator on the right which creates electricity. A portion of the electricity is injected again in the system for further functioning" [8]. Geneste goes on to point out that currently these types of engines are so heavy that any savings in power generation or design are cancelled by the excessive weight, which translates into extra cost to launch the satellite into space. Therefore research is being done into ways to make thermal acoustic engines that weigh less so they could be more cost effective than photovoltaic cells [8]. He states that this would "be a dramatic breakthrough in the upcoming years for electricity generation onboard space crafts in general allowing much greater available powers, more reliable power subsystems and in the end much cheaper devices" [8].

There is another area of concern that is receiving a considerable amount of attention regarding research in the realm of hybrid satellite communications (hybrid communications just means that Internet data is being sent by a combination of both terrestrial backbone networks and satellite links serving as a backbone too). This area of concern deals with the packet headers. Yung-Wey and Tat-Chee remind us "to ensure that data are sent to their destination over the hybrid satellite-wireless networks, several layers of encapsulations are applied to the packet" [1]. With all the different types of layers the packet goes through, several layers of encapsulation are applied, inducing UDP, TCP/IP, RTP headers, and others, sometimes the headers are far bigger than the payload itself. This is especially true with Voice over IP packets, which have small payloads that can be 25% of the total packet size including headers. Yung-Wey, et. al, propose that to save space and limited satellite transmission resources, they should use some kind of compression that makes use of the fact that many of the header fields that are repeated at each layer are constant, such as the source and destination information, etc. Two existing compression technologies that accomplish this are called Robust Header Compression and Payload Header Suppression [1]. He explains that in Robust Header Compression, the redundant information is not needed to traverse the satellite links. The information in the header is assigned a code so that lets the receiving device can reconstruct the original header information to complete the frame again [1]. The following table, as posted from Yung-Wey, et. al., shows the reduction in byte size of the header information in the tested packets.

TABLE. II. REDUCTION IN BYTE SIZES OF THE HEADER INFORMATION IN THE TESTED PACKETS [1]

	Eth+uncompressed IP/UDP/RTP	Eth-HC+compressed IP/UDP/RTP
SS-BS	60	24
BS-RCST	81	24
RCST-satGW	95	21
satGW-AR	81	24
AR-CN	60	-

IP/UDP/RTP – Internet protocol/User datagram protocol/Real-time transport protocol; Eth-HC – Ethernet header compression; SS – Subscriber station; BS – Base stations; RCST – Return channel satellite terminal; AR – Access router

So it was shown in Yung-Wey, et. al that hybrid header compression did indeed contribute to a higher quality of service in Voice over IP connections through hybrid and WiMAX networks. Furthermore, small payload packets benefit a lot with the much smaller header overhead and this had the effect of reducing mouth to ear delay and jitter in the transmissions. This paper indicated that these tests were primarily done over WiMAX networks but can be expanded to other hybrid wireless terrestrial technologies as well [1].

Other research is ongoing in the areas related to satellite telecommunications. Security of the transmissions is one common problem with satellite communications. According to Chang and Cheng, the transmissions passing through the air from the ground to the satellite or from the satellite to the ground are at risk from a variety of attacks such as man in the middle or sniffing. These signals can be intercepted or impersonated. It is possible for a hacker to gain access to information they should not have [9]. The authors point out that a well-designed satellite communication system will have some kind of authentication scheme so that the satellite and ground station will be able to authenticate each other before a session key is negotiated. Research is ongoing on this area to increase the security and reduce the ability for these keys to become compromised in various types of sniffing or man in the middle attacks [9].

It goes without saying that data transmitted between satellites and the ground stations are wireless. Research by Zhang, et. al shows that "the low frequency bands crowding and the increase of broadband services diffusion has created the premise for the development of communications in the millimeter-wave (mm-wave) bands (Q-V bands 35-75 GHz, W-band 75-110 GHz). Due to wider bandwidths and higher frequencies, wireless deliver is expected to provide multigigabit data transmission" [10]. Therefore, they point out that the transmissions on the ground to and from the ground station can only be carried by optical lines as only that can handle the high speed data transmission rates. This introduces complexity in the ground station because expensive and complex equipment is needed to fully integrate the fiber optic signals and the high speed wireless signals to and from the satellites. A whole new set of demodulation equipment is needed for each satellite downlink, which also means more fiber optic lines to the central office from the ground station. Zhang, et. al proposes using optical polarization multiplexing to simplify and lower the cost of the systems operating in the ground station. This is a principle in which the light waves can be broken up by polarization multiplexing into the different satellite signals. Then all the satellite downlinks can be multiplexed onto one fiber optic line to the central office, reducing cost in both fiber optic lines and demodulation equipment, which may be redundant. That way, at the central office the combined signal can be deconstructed again using a polarization beam splitter [10].

V. CONCLUSION

It is clear to see that satellites have a whole host of applications, many of which are in the telecommunications field. It is not as simple as launching a platform containing antennas into space and sending transmissions through it. The huge amount of data and voice transmitted today demands that constant improvements be made in security and data throughput. Furthermore, it is essential that cheaper operating

and construction costs are realized because these satellites are going to wear out from the huge demands placed on them. Any efficiency that can be gained in data throughput, power generation, and launch cost will have positive implications for the future. One can never rest in this field; research must be constantly striving to improve cost of ownership, security, and data transmission rates, as well as developing newer and better protocols for how bandwidth is managed. Huge strides have already been made just since the Internet became mainstream.

REFERENCES

[1] Yung-Wey, C., & Tat-Chee, W. (2013). Comparative Study on Hybrid Header Compression over Satellite-Wireless Networks. *IETE Technical Review(Medknow Publications & Media Pvt. Ltd.), 30(6).* 461-472. doi:10.4103/0256-4602.125663.

[2] Slotten, H. R. (2015). International Governance, Organizational Standards, and the First Global Satellite Communication System. Journal Of Policy History, 27(3), 521-549. doi:10.1017/S0898030615000214

[3] United States Government Accountability Office. (2014). GEOSTATIONARY WEATHER SATELLITES: PROGRESS MADE, BUT WEAKNESSES IN SCHEDULING, CONTINGENCY PLANNING, AND COMMUNICATING WITH USERS NEED TO BE ADDRESSED. Journal of Current Issues in Media & Telecommunications, 6(3), 253-295.

[4] Boriboon, A., & Pongpadpinit, S. (2016). Optimized routing protocol for broadband hybrid satellite constellation communication IP network system. EURASIP Journal On Wireless Communications & Networking, 2016(1), 1-11. doi:10.1186/s13638-016-0616-2.

[5] Pontani, M. (2015). LOW EARTH ORBIT SATELLITE CONSTELLATIONS FOR LOCAL TELECOMMUNICATION AND MONITORING SERVICES. Journal Of Current Issues In Media & Telecommunications, 7(3), 299-322.

[6] Poole, Ian. Satellite Orbit Types and Definitions. Radio-Electronics.com: Resources and Analysis for Electronics Engineers. Retrieved from: http://www.radio-electronics.com/info/satellite/satellite-orbits/satellites-orbit-definitions.php

[7] Lee, S., Wu, Y., & Mortari, D. (2015). Satellite constellation design for telecommunication in Antarctica. International Journal of Satellite Communications and Networking.

[8] Geneste, J. (2015). A New Architecture for Electricity Generation Obboard Telecommunications Satellites. Magnetohydrodynamics (0024-998X), 51(3), 629-635.

[9] Chang, C., Cheng, T., & Wu, H. (2014). An authentication and key agreement protocol for satellite communications. International Journal Of Communication Systems, 27(10), 1994-2006. doi:10.1002/dac.2448.

[10] Zhang, Z., Yu, J., Chi, N., Xiao, J., Xu, Y., & Fang, Y. (2015). A novel architecture of satellite-ground communication system at W-band based on RF transparent demodulation technique. Microwave & Optical Technology Letters, 57(2), 409-414. doi:10.1002/mop.28862

A New Model of Information Systems Efficiency based on Key Performance Indicator (KPI)

Ahmad AbdulQadir AlRababah
Faculty of Computing and Information Technology in Rabigh,
King Abdulaziz University,
Rabigh 21911, Kingdom of Saudi Arabia

***Abstract*—Any company concerning with information technology considers Automated performance management processes as a key component of its operations as it enables the company get a clear long -term assessment about the performance of employees as well as operating units of the company. One technique that the company could use to evaluate the present performance of both employees and operating units is the utilization of KPI -based management information system. The current study seeks to provide a new model of information system efficiency based on key performance indicator and the extent to which such approach helps the company evaluate the performance within the company. In addition to recognize the requirements and criteria needed to establish an effective system of performance measurement , the axioms that may influence the designing of the Model KPIs and the approaches that We need to be contracted with a fixed set of key performance indicators in order to facilitate hiring.**

Keywords*—*Information Systems Management; System performance; Key performance Indicators; Data warehouse; Information Systems Integration

I. Introduction

For modern enterprise management - process complexity is an important constraint of calibration solutions that needed to be taken, based on a very short time, a large number of financial analyses and other information. Modern manager should not only know how to solve problems quickly, but also forecast them on proper time, and help enterprises find new opportunities and achieve good prospects for the new development [5].

Indicators or key performance indicators (KPIs) in business environment are mostly quantitative information. It illustrates structures and processes of a company. Recently, KPIs are considered very important for planning and controlling over supporting information, creating transparency and supporting decision makers of the management[1].

The article deals with the development of information systems (IS), which allows anyone to solve complex problems effectively in the world of business. At the core of the control system, efficiency is considered as a formal method for evaluating the performance staff members and operating units within the company. Key performance indicators, - KPI can be seen as a supporter that estimates misleading business process (BP) performed by staff members and operating units.

Modern business environment where information has become the most important resource requires new approaches to asses performances of organizations .comparing to traditional performance measurement system which evolved just financial and accounting indicators, KPI. KPI is considered as one of the newest approaches that are used to financial and non- financial to reveal how successful companies have accomplished long -term goals. To constitute an effective system of performance measurement, it is very crucial to define and standardize all processes within the organization [8].

Many companies have utilized inappropriate measures; therefore such measures should not be called performance indicators (KPIs). In fact, most of the organizations that are considered to have true control system use key performance indicators and why there are accountants, business leaders and consultants have experience and sufficient information about Key Performance Indicators.

II. Materials And Methods

There are four kinds of key performance indicators:

- (KRIs) , That can say it has achieved critical success factor of perspective or purposefully
- (RIs): Here we can say that you have done.
- (PIs) , He can say it what should be done .
- KPIs, Who can say that it must be done to increase the high degree of performance

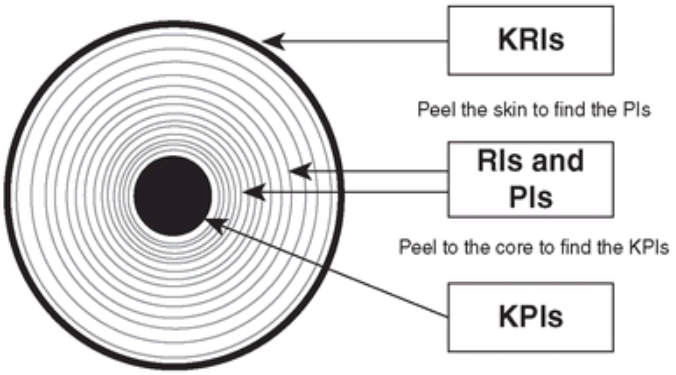


Fig. 1. Four kinds of key performance indicators

We will define key performance indicators by organizations, measures are being carried out by these organizations to review progress against goals and we will

King AbdulAziz University- Kingdom of Saudi Arabia

dividing these goals, which are to achieve progress by departments and individuals are reviewed these goals and achieve them and apply them at regular intervals by organizations [2].

- Input: Input consists of individuals and materials used to carry out the task and implement all the means to get the output
- Activities: are the methods used to input considered so as to generate outputs and outcomes at the end of the process
- Output: The output is the final product of goods and services produced
- Outcomes: Results are considered that affect the specific outcomes that should be the end of the strategic goals of the institutions and the results of performance that are supplied with plans
- Impacts: These exact results, which are carried out such as the creation of jobs for young people or reduce the need for organizations considered in order to prove how to use the available resources so as to obtain information on the precise results.

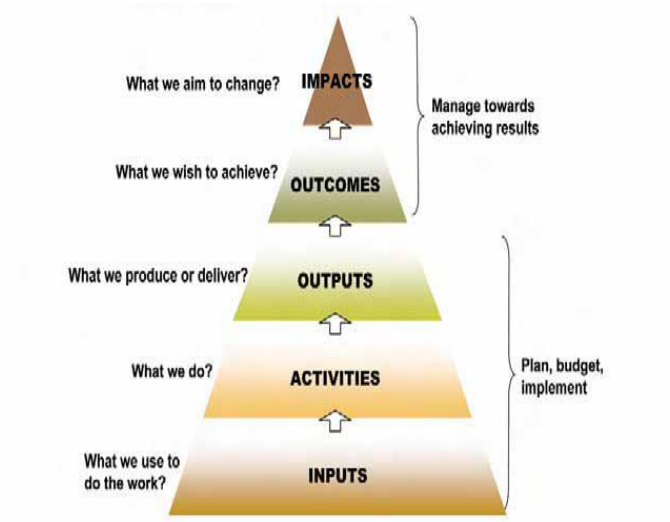


Fig. 2. Key performance information concepts

The researcher sees that a complete management information system should be in order to planning skills of the different variables and is considered a form Key Performance Indicators is a model designed for a short time but there is a goal of all this process, but elusive and the reason for this is to use a small number of Mini easy, simple and fast way, but in a comprehensive manner for the operational needs of managers, but are to achieve the interests the public is a form key performance indicators can be a cornerstone of the information management system [4, 8] .

Here we will measure the design of the main performance on basic intuitions which are relevant to measure performance and to be developed without being influence the basic model indicators, performance has to be measured against goals fixed by the Courts;

First, must be linked to performance indicators targets performance and must be recorded important notes, but not all the courts that is reminiscent of the goals however that the courts will be affected generally be through the resolution of conflicts in accordance with the law through a fair, fast and efficient process and are key performance indicators in the form of focus on punctuality and be effective also in terms of cost, or what is known as case management or goal of the court system This is not to focus on the objectives or results of the process are not important to the contrary, as he can use the terminology to reduce the concentration and confirms that the performance respect to objectives through derived from the court system and plans procedural structural characteristics and see all the people in a bid to reduce the time and cost required to achieve targets a modest sense to try to do the same thing with the objectives of a fair process .

Second, it should be setting goals that in the measurement of the courts must rely on performance standards with respect to the general objectives in terms of cost and in terms of time and this has many rollover courts and performance standards to overcome the frustration and feelings of control used in therapy with less than measurement standards performance and this example to take advantage of the time and this is a completely failed to take into account that the process should be time-consuming to defend the distinctive Qaeda between the time that it takes for this process and that the delay is not justified and must be similar observations in terms of cost and that some effective

Third, it must be performance results follow more measures as the main performance of the courts are in accordance with the operation and the achievement of performance targets and is expected to indicators of performance are the work of the present and future of the court and must be of a physical link or are in accordance with the progress standards and should not call for them and must The issues to be a new beginning will generic terms that are required ideally to create a corrective action show

Finally, must be the main targets a few simple indicators in order to show the central idea, however, it should be doubled simply must Court, which maintained a session of the drought flood statistics show that after years of safety most data archaeological courts and are now producing full books on the tables and charts so as to improve the stomach reporting good numbers and considers this a tie with a significant space in the management of the operations day after day, however, it should be the objectives of the main work will be very easy to provide the maximum size of useful data with a minimum of information are in the form of loss easily must be The final message is simple [3, 5].

In March 1999, has been put forward a discussion paper to provide projects of key performance indicators in an attempt to control the judiciary and how the president administration for each of the courts of New South Wales were accepted several comments have been model correct in the light of this report, it has fought hard for each the false comments and how we can avoid that general and private performance indicators: The key performance indicators to measure the performance, which

connects all the users of the service or are measured service aspects that are relevant to the user:

- General key performance indicators measured aspects of performance that are linked to the users of the service is not aimed at a specific class by itself.
- Private Key Performance Indicators is aimed at the user who performed a specific service designated bodies of nationals relevant to those users of the service.

The evolution of key performance indicators: These factors will be displayed as a series of stages may apply some of them and others do not occur at any step in this process is to identify these factors through the collection and analysis after a comprehensive review and must be set up key performance indicators

1) Introduce the public and use for measurement: To be necessary to determine the measurement objectives in order to identify and develop indicators appropriate performance and will observe whether the goal of the measurement is the comparison, both internally and in order to improve the quality or other institutions, which aims to influence the process of selection of key performance indicators if it is to prepare performance indicators to measure performance should be selected key performance indicators widely internationally and there are many high-quality areas such as efficiency security before the start of performance measurement process because it is important to identify the domain that you are measuring the performance to him and which are in turn dependent on the public for the quality of the product evaluated in order to determine balance between key performance indicators

2) Investigation a balance in measurement: Large numbers of the diversity of stakeholders in the field of health and sponsorships are needed to several areas to meet the information needs and the implementation of a number of ways to assist in the identification of a set of key performance indicators,

In the beginning of the panel, which was known by Kaplan and Norton, a dashboard and balanced, which proposes four views indicators specific goals to provide an opinion on the system

- Services expected of a member of that can measure the needs of the beneficiary and wait service
- The method of administration, which measure the basic mode of action that are identified as significant and effective task
- It is measuring an idea to improve communication systems and the ability of the public

The evaluation of the efficient use of resources to achieve the objectives of the arrangement and used in the economy and efficiency

*3) Indicator definition***:** That kind are provided, such as this form of collapse that must be included when applying key performance indicators and model access to the key elements that must be taken when determining indicators main performance, however template is typically to retrieve the requirements that improve indicators main performance in separate services

*4) Consult with stakeholders and advisory group***:** This type there must be consultation with all stakeholders along the preparation of the data and the process of consultation with each other to help identify their needs at the same time to accept the main performance indicators also facilitates consultation on data elements and help them to know the data [6].

The consultation with the decision-makers help identify their own needs and that can be relied upon at any later time and be consultation with the service that will help determine their information requirements and to obtain data that can be available and can facilitate dialogues and analysis of individuals in facilitating the identification providers core competencies and to participate the service users that will help determine data requirements and the proposed information raises a lot of privacy and confidentiality consists [7].

Information system provides an overview of the process KPI´s in a defined time period (month, quarter, year), including the possibility of obtaining review of the process, organizational units, employee and business partner. The system limit the reviewing of information in accordance with the authorization of a system user, through the personalization of content. The system use data from an existing integrated information system (ERP) and other records necessary to calculate the KPI´s of business processes. The system is implemented as a SOA solution, in accordance with the SOA methodology and the use of Silver light technology. A Web portal is created for the interaction between user and system. The user accesses the system, the system performs its identification, records user's activity, and then takes the appropriate data from the business processes records based on them calculates KPI´s for the corresponding processes. Finally, the system displays the process performance to the user. All the basic and alternative steps for interactive work are defined.

Detailed instructions for measuring KPI´s of supply, which are applied in the private methods individually for each defined KPI´s [2],[7]. Silverlight Web portal was developed in Visual Studio 2010 SP1 environment, using C# programming language for programming "code-behind" classes and Silverlight technologies. Applied is the Form-based user login to the portal is implemented using the Login control. Finally, Silverlight Web portal allows the user different views of graphs and tables with KPI´s ratings of supply and sales processes. By calling Prikazi (DISPLAY) the service Performance is called, which calculates and returns the process performance data to Web portal, in accordance with defined parameters, and results are presented in tables and graphs. Both displays are updating by selecting Information or Ratings.

III. RESULTS

The current study has shown that establishing a model of key process performance that characterized with standardized criteria to measure the effectiveness of operations and success

of employees with accompany is possible. The study also has shown that automated collection as well as processing of the needed data related to company could provide accurate, full, and up to date information about how to keep records concerning the process implementation, mainly if these records are governed by appropriate procedures. In addition, the study has shown that short-term performance measurement process could help analyze data on time, and efficiently resolve inconsistencies with the goals of process improvement, products / services and overall company results. Finally, the study has proved that applying solutions presented by this paper could help link IT resources with business goals of the organization, help the organization to build communication with customers and suppliers and internal links of organizational units, allow more accurate, complete information, make crucial quality decisions, and at the same time support key business processes through the increased availability of information which significantly influence increasing the total effectiveness of the company.

IV. Discussion

To evaluate the performance of companies are using key performance indicators. They help comparing the present performance of the company with the previous one, Standards that measure the industry and even individual industries and accordingly are trying any system of logistic to improve and guide the decisions are made by the standards, which are the rating to later In fact, this idea is considered brilliant and clear about the factors that drive logistics system development and it could be available for a range of adequate targets planning.

V. Conclusion and Future Recommendation

Four reviews of the obtained processes performance are possible: review of KPI´s during the specified time period and by months, review of processes ratings during the specified time period and in parallel by months, including the possibility of obtaining these views for individual organizational units, employee and partner, who were involved in implementation of business processes. It is also possible to view KPI´s only one process or the review of only one particular KPI´s.

Acknowledgement

This work was supported by the Deanship of Scientific Research (DSR) at King Abdulaziz University, Jeddah, Saudi Arabia, under Grant No. (830-57-D1436). The author, therefore, gratefully acknowledge the technical and financial support from the DSR.

References

[1] Meier, Horst, et al. "Key performance indicators for assessing the planning and delivery of industrial services." Procedia Cirp 11 (2013): 99-104.

[2] FinPa New Media 2009, Key Performance Indicators, FinPa New Media, Melbourne, viewed 24 February 2009, <http://swinburne.projects.finpa.com.au/toolbox10/releases/final/toolbox/resources/res4040/res4040.htm>.

[3] Ballantine J, Brignall S, Modell S. Performance measurement and management in public health services: a comparison of UK and Swedish practice. Management Accounting Research. 1998; 9 pp.71-4.

[4] Kaplan RS, Norton DP. The Balanced Scorecard - Measures that Drive Performance. Harvard Business Review. 1992; 70(1): pp.71-9.

[5] Audit Commission. On Target: the practice of performance indicators. Audit Commission for Local Authorities and the National Health Service in England and Wales; 2000. Available online from: http://www.auditcommission. gov.uk/Products/NATIONAL-REPORT/266D51B7-0C33-4b4b-9832-7484511275E6/archive_mptarget.pdf.

[6] AIHW. A guide to data development. Australian Institute of Health and Welfare (AIHW); Report No.: AIHW Cat. no. HWI 94. 2007.

[7] WHO. Mental Health Information Systems (Mental Health Policy and Service Guidance Package). World Health Organisation; 2005.

[8] Arora, Amishi, and Sukhbir Kaur. "Performance Assessment Model for Management Educators Based on KRA/KPI." International Conference on Technology and Business Management March vol 23.2015.

GIS Utilization for Delivering a Time Condition Products

Noha I. Sharaf
Information System Dept
Faculty of Computers and Information Science
Mansoura, Egypt

Bahaa T. Shabana
Computer Sciences Dept
Misr Higher Institute for Commerce & Computer
Mansoura, Egypt

Hazem M. El-Bakry
Information System Dept
Faculty of Computers and Information Science
Mansoura, Egypt

***Abstract*—As population is increasing rapidly all over the world, the need for delivering products is being more difficult especially for conditional products (products with life time). Many Customers require conditional products to be delivered to their locations. Distribution center may have multi depots (multi store branches) instead of one depot. Every depot has limited number of vehicles to minimize cost. Capacities of these vehicles are based on two dimensions (weight and volume). Geographic information system (GIS) is used for localizing customers' destinations. Then OD Cost Matrix is used to assign every customer destination to the least cost depot to be served from it. Finally Network analyst is used to solve the vehicle routing problem generating final route directions for every vehicle and calculating the best time for lunch break of drivers automatically. This case study is applied on Mansoura city in Egypt.**

***Keywords*—conditional products; distribution centers (depots); capacity of vehicle; vehicle routing problems (VRP); geographic information system (GIS); OD Cost Matrix; Network analyst**

I. INTRODUCTION

As demands on conditional products are increasing very rapidly all over the world, the transportation of deliveries is being a hard process. Traditional ways used in delivering conditional products depending on experience of the drivers. In many situations, it is not efficient way as long as many customers requesting many products from the same depot. These products have a life time and must be delivered within it. Having many factors in consideration, like capacity of limited vehicles which is based on two dimensions (weight and volume), calculating the time of break for every driver, life time of products, sequence of delivering products matching minimum cost, travel distance and travel time without violating time constraints. To minimize travel distance and total time, vehicles should visit the customer's location only one time. Vehicles start and end on same depot. Customers are clustered according to geographical zone. Assumed, all depots are huge enough to have all requested products. To minimize cost, products with small life time should be delivered faster than others so as not to spoil. A driver working hours is set to be 6 hours per day. Working more than 6 hours is calculated as overtime. Total cost is calculated by cost of the drivers' working hours, overtime hours and fuel cost consumptions in the transportation. GIS tools help in solving the problem of delivering conditional products. AS GIS use spatial data [1] (geographical data) and attribute data [2] especially in road networks [3]. GIS proved its efficiency in many applications such as transportation (GIS-T) [4] , finding best route, traffic and road congestion problems [5] intelligent transportation systems (ITS) , scheduling and routing school buses [6], Travelling sales man (TSP), Vehicle routing problems (VRP) as it is considered an extension of TSP [7] and supply chain management. In this paper, GIS tools and network analyst are used to solve problem of delivering conditional products within its life time to customers located in Mansoura city in Egypt and determining the suitable time for drivers break automatically.

II. LITERATURE REVIEW

Surekha P and S.Sumathi used Genetic Algorithm (GA) to solve the problem of Multi-Depot Vehicle Routing Problem (MDVRP). Then, they used Clarke and Wright saving Method in routing using MATLAB R2008b software. In this paper, OD cost matrix and Network Analyst VRP is used to solve the problem of transporting deliveries from the least cost depot to customers' destinations without violating lifetime constraints of products using ArcMap 9.3 software.

Rong-Chang Chen, Chih-Hui Shieh, Kai-Ting Chan, Shin-Yi Chiu, Jyun-You Fan, Yu-Ting Chang and Nuo-Jhen Ma introduced systematic approach for solving the problem of delivering Service for Bento Industry based on three-stage approach. They used GIS for locating customers' locations. Then, they used K-means algorithm for clustering. Finally, they used Genetic algorithm (GA) to get the shortest route with shortest travel distance. In this paper, Network Analyst is used instead of GA and OD cost Matrix is used to assign requested orders to be delivered from least cost depot to minimize total cost. In addition to, the system is determining the break time of drivers based on products assigned to their schedule.

Hari Shankar, Gangesh Mani, Kamal Pandey used Tabu search algorithm with GIS to solve the multi depot vehicle routing problem with time window (MDCVRPTW) in capital city Dehradun of Uttrakhand state. They considered many parameters with predefined static break time. In this paper, break time of every driver is calculated automatically according to his schedule without violating products' lifetime and vehicles constraints.

M.Abousaeidi, R.Fauzi and R.Muhamad used GIS to find the fastest delivery route not the shortest route as the shortest route may not be the optimal route. GIS based on VRP was used by V.K.Purwar, Varun Singh and R.C. Vaishya to solve

milk distribution problem in Allahabad city. In this paper, GIS used to find the route delivering all conditional products within lifetime with minimum total cost by using OD Cost Matrix.

In this paper, GIS tools are used to solve the problem of Multi depot vehicle routing problem (MDVRP) for conditional products with life time in Mansoura city in Egypt. Minimizing, total cost and travel distance during visiting all the required regions (Customers' destinations). Travelling to every customer's destination from the nearest least cost depot. Generating routes with directions and calculating the proper time for lunch break of every driver automatically based on his schedule.

III. Problem Definition

Main problem is delivering all requested conditional products to customers without violating constraints with minimum total cost and calculating drivers' lunch break automatically. Using Systematic approach based on GIS tools (ArcGIS software) instead of traditional ways to transmit requested orders to customers' locations from the nearest least cost depot. Not the depot where received the customers' request. As customer can make request from any depot. But it must be served from the nearest depot with the least cost. This case study is applied in Mansoura city in Egypt.

A. Proposed Constraints

Customers make orders requesting conditional products with varying lifetime. Products must be delivered within its life time to be valid to be used. Every Customer's order cannot be composed into different vehicles [8].

To minimize total cost, depots have limited number of vehicles. Customer's location should be visited only one time and served from the nearest depot even though the requested order was from any other depot. Capacities of vehicles transmitting orders are based on two dimensions weight and volume and have assumed to be 6000[9], [10]. Vehicles start and end on same depot. Having restrictions over visiting routes according to their geographical places (route Zones). Number of depots and vehicles are limited (assumed having 2 multi-depots with multi vehicles). All depots are huge enough to store all types of products. Customer's demand is being served from the least cost depot to minimize travel cost and travel distance [11]. Drivers working hours cost are assumed to be 12 EGP per hour and over time is calculated after working more than 6 hours and is assumed to be 18 EGP per hour. Availability time of vehicles, drivers and depots are proposed to start from 8:00 AM to 5:00 PM. Drivers must have a dynamic payable period for lunch break instead of static time break based on different policies of different companies. Starting time and ending time of lunch break is calculated automatically. Managing total cost of transportation to be minimized including factors of driving cost per hour, overtime driving cost per hour, fuel consumption cost which is calculated by the travel distance cost per Mile. Including gained revenue from delivering orders. Route directions of vehicles are changed automatically according to occurred barriers [12].

B. Expected Result

Giving every driver the best route directions including time of his lunch break calculated automatically without violating any constraints from the proposed constraints with minimum total cost in distance and time as much as possible.

C. Data Collection

- Streets collected from world street map as a layer into ArcMap 9.3[13].
- Geo-database designed on ArcCatalog.
- Data projection applied on world street map layer with Projected coordinate system WGS _1984 _UTM_Zone_36N

D. Database in details

This case study is applied on random region in Mansoura city in Egypt as shown in Fig 1. Assumed, having, 13 customer requesting 6 different conditional products from two center depots. Mansoura city is not huge enough but has heavy congestion which makes delivering conditional products hard process to be achieved optimally.

Fig. 1. Random selected area in Mansoura city in Egypt

Geo-database contains dataset with feature classes of streets as polyline type and depots as point type [14]. Properties of streets and center depots feature classes are represented in table I and table II consequently. Used Network Dataset is shown in Fig 2.

TABLE I. PROPERTIES OF STREETS

Field Name	Data Type
OBJECTID	Object ID
SHAPE	Geometry
NAME	Text
TYPE	Text
Oneway	Text
FT_Minutes	Double
TF_Minutes	Double
SHAPE_Length	Double
houseFromL	Short Integer
houseFromR	Short Integer
houseToL	Short Integer
houseToR	Short Integer
FT_Speed	Double
TF_Speed	Double

Fig. 2. Network Dataset of the selected region

TABLE II. PROPERTIES OF DEPOTS

Field Name	Data Type
OBJECTID	Object ID
SHAPE	Geometry
StoreName	Text
StartTime	Date
CloseTime	Date

E. Steps of solving the problem

Step1: Designing geo-database in ArcCatalog.

Step2: Collecting or writing assumed data from customers (customers' demands).

Step3: For all customers

3.1 Check if customer's demand was not from nearest depot

3.2 Use OD cost Matrix on network analyst in ArcMap to get nearest least cost depot [15].

Step4: Adding all proposed constraints.

Step5: Using network analyst vehicle routing problem to solve the problem [16]

Step6: Output result (generating routes with directions and calculating proper time of dynamic launch break).

Step7: If there is barrier occurred in resulting route directions (street)

7.1 Repeat steps 5, 6

Else print directions windows of all routes

IV. SYSTEM'S INPUT DATA

Assumed proposed input data properties are shown in table III, table IV consequently. Attributes of the customers' orders table are customer name, customer address, product Name, depot address which represents address of depot that received the customer demand, order time which is time of requesting the product, Product LifeTime which is the end time of life time of the requested product, total orders which is the total number of requested orders, revenue per unit, total revenue, unit weight, unit volume, total weight, total volume, total quantities as is represented on two dimensions (weight and volume), maximum transmission time in minutes and specialty name as some products need a foodkeeper to save its temperature so as not to spoil like an ice-cream.

Attributes of drivers' lunch break table are RouteName as break is calculated separately based on every route, serviceTime as it is varying between companies based on their internal policies. Starting time and ending time are set to be NULL to be calculated by the system based on conditional products of each route.

TABLE III. PROPERTIES OF CUSTOMERS' ORDERS

Field Name	Data Type
ID	Object ID
CustomerName	Text
CustomerAddress	Text
ProductName	Text
DepositAddress	Text
OrderDate	DateTime
Product LifeTime	DateTime
TotalOrders	Short Integer
UnitRevenue	Double
TotalRevenue	Double
UnitWeight	Double
UnitVolume	Double
TotalWeight	Double
TotalVolume	Double
TotalQuantities	Double
MaxTransitTime	DateTime
SpecialtyName	Text

TABLE IV. PROPERTIES OF DRIVERS' LUNCH BREAK

Field Name	Data Type
RouteName	Text
ServiceTime	DateTime
StartTime	DateTime
EndTime	DateTime

V. SYSTEM'S OUTPUT

For every customer, the nearest depot should be specified to get requested products from it and deliver them to customers. OD Cost Matrix in ArcMap is used to determine the least cost depot from all customers' locations as shown in Fig 3.

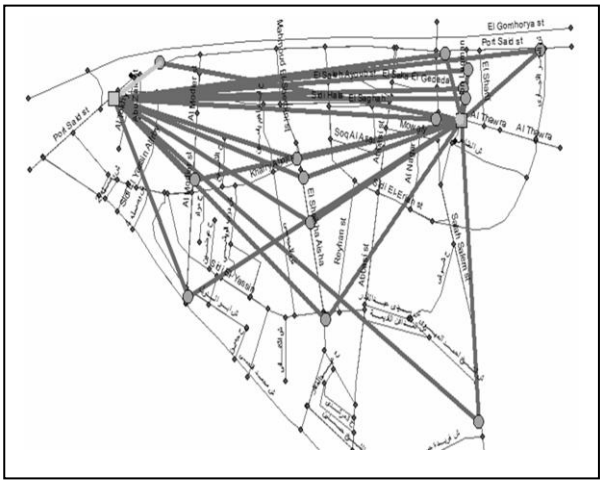


Fig. 3. Result of OD Cost Matrix

As assumed, having 2 depots (left and right) and received 14 orders. The Output result is 28 lines. Every customer's order is represented in two rows. Each row representing Rank destination, total minutes (cost of traveling along the network) from customer destination to every depot (the left and right depot). The least cost depot has rank 1 as shown in Fig 4. For example Travel cost from Nancy destination to left depot is 0.65 minutes, and travel cost from Nancy destination to right depot is 7.2 minutes. So, rank value 1 is given to left depot (left store) and Nancy order is collected from the left depot (the least cost depot) and so on.

Attributes of Lines

ObjectID	Shape	Name	OriginID	DestinationID	DestinationRank	Total_Minutes
57	Polyline	Nancy (Home) - Left strore	17	18	1	0.655376
58	Polyline	Nancy (Home) - Right store	17	19	2	7.258508
59	Polyline	Hend (Home) - Right store	18	19	1	1.139772
60	Polyline	Hend (Home) - Left strore	18	18	2	5.720807
61	Polyline	Heba (Home) - Left strore	19	18	1	1.865804
62	Polyline	Heba (Home) - Right store	19	19	2	6.361728
63	Polyline	Samy (Home) - Right store	20	19	1	0.338936
64	Polyline	Samy (Home) - Left strore	20	18	2	7.026742
65	Polyline	Basem (Home) - Left strore	21	18	1	2.018723
66	Polyline	Basem (Home) - Right store	21	19	2	6.208809
67	Polyline	Manar (Home) - Right store	22	19	1	1.361076
68	Polyline	Manar (Home) - Left strore	22	18	2	7.274
69	Polyline	Nawal (Home) - Right store	23	19	1	0.207879
70	Polyline	Nawal (Home) - Left strore	23	18	2	7.068458
71	Polyline	Ali (Home) - Left strore	24	18	1	2.251741

Record: 1 Show: All Selected Records (2 out of 28 Selected) Options

Fig. 4. Result of OD Cost Matrix

The Output of OD Cost Matrix is used as an input into network analyst VRP. After using the vehicle routing problem in network analyst in ArcMap 9.3. Every order is assigned to only one vehicle to be delivered to the customer. Routes with directions of vehicles are also determined without violating all proposed constraints.

Route properties are shown in Fig 5. Capacity of vehicles is proposed to be 6000 for weight and 6000 for volume. Cost per unit variable is set to 12 EGP per hour. By dividing 12/60 equals 0.2. Cost per unit time equals 0.2 .Over time starts after working 6 hours so it is set to be 360 (6hours * 60 minute). Cost per unit over time is set to be 18 EGP per hour. By dividing 18/60 equals 0.3. After solving the problem with no barrier, calculated fields of routes are shown in Fig 6 and Fig 7 consequently without time violation.

Properties

Attribute	Value
StartDepotServiceTime	<Null>
EndDepotServiceTime	<Null>
EarliestStartTime	11/18/2015 8:00:00 AM
LatestStartTime	11/18/2015 10:00:00 AM
Capacities	6000 6000
FixedCost	<Null>
CostPerUnitTime	0.2
CostPerUnitDistance	1
OvertimeStartTime	360
CostPerUnitOvertime	0.3
MaxOrderCount	30
MaxTotalTime	<Null>
MaxTotalTravelTime	<Null>
MaxTotalDistance	<Null>
SpecialtyNames	foodKeeper
AssignmentRule	Include

OK Cancel

Fig. 5. Properties of routes

Properties

Attribute	Value
OrderCount	14
TotalCost	37.753895
RegularTimeCost	28.532966
OvertimeCost	0
DistanceCost	9.220928
TotalTime	142.664832
TotalOrderServiceTime	35
TotalBreakServiceTime	30
TotalTravelTime	25.7415
TotalDistance	9.220928
StartTime	11/18/2015 9:16:21 AM
EndTime	11/18/2015 11:39:01 AM
TotalWaitTime	51.923333
TotalViolationTime	0
RenewalCount	0
TotalRenewalServiceTime	0

OK Cancel

Fig. 6. Calculated fields of Left Route

Properties

Attribute	Value
ViolatedConstraints	<Null>
OrderCount	8
TotalCost	19.98059
RegularTimeCost	16.698578
OvertimeCost	0
DistanceCost	3.282012
TotalTime	83.492889
TotalOrderServiceTime	20
TotalBreakServiceTime	20
TotalTravelTime	9.346903
TotalDistance	3.282012
StartTime	11/18/2015 10:00:00 AM
EndTime	11/18/2015 11:23:30 AM
TotalWaitTime	34.145986
TotalViolationTime	0
RenewalCount	0
TotalRenewalServiceTime	0

OK Cancel

Fig. 7. Calculated fields of Left Route

Break time of every vehicle is calculated dynamically and appears in directions window of Left Route and Right Route as shown in Fig 8 and Fig 9 consequently. Break time of each vehicle is calculated automatically according to requested orders in this route. As seen break time of the right route differs from the break time of the left route. Break time of the left route starts on 9:37AM and ends on 10:07AM. Break time of the right route starts on 10:35 AM and ends on 10:55 AM. Period of the break on left route is 30 minutes varying from period of the beak on right route which is 20 minutes.

If there is an accident occurred on sidi Hala st, calculated fields and directions of left route are changed automatically as shown in Fig 10, Fig 11.

This means that after adding barriers into system like an accident or fire such as the proposed accident on Sidi Hala st, system regenerates new route directions with alternative paths and recalculated properties again.

The break time on left route is changed automatically to meet all constraints (changed from 9:37 Am and 10:07 Am to be 9:30 Am and 10:00 Am).

As seen total cost field is increased from 37.75 to 39.64. Also total time is increased from 142.66 to 149.07. The starting time and ending time is changed. But still there are no violations in time constraints. Directions window of the left route changed and Sidi Hala st was not visited. Resulting route paths before and after barrier is shown in Fig 12.

Fig. 8. Directions of Left Route with dynamic breaktime

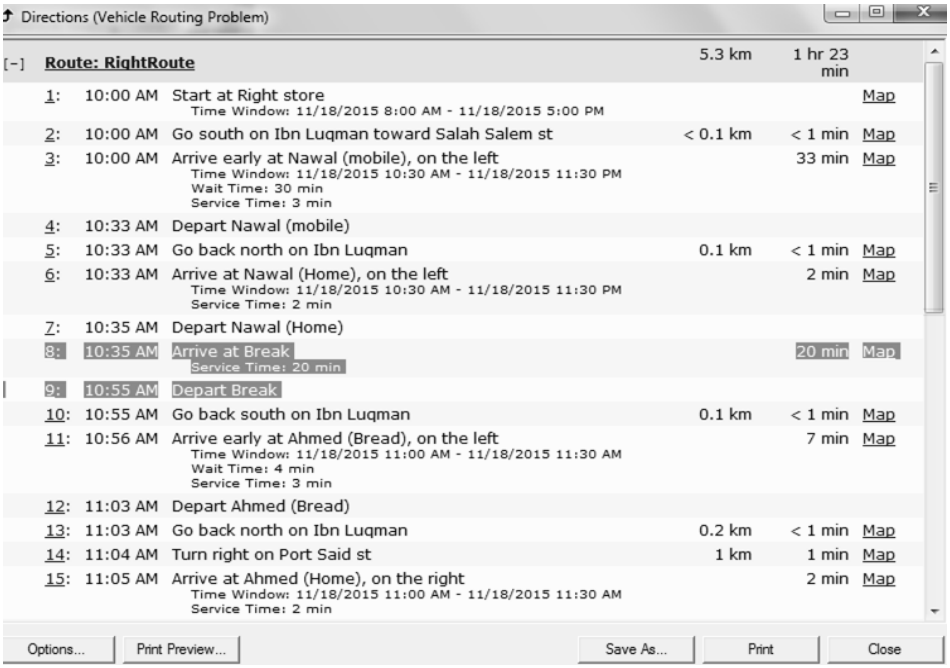


Fig. 9. Directions of Right Route with dynamic breaktime

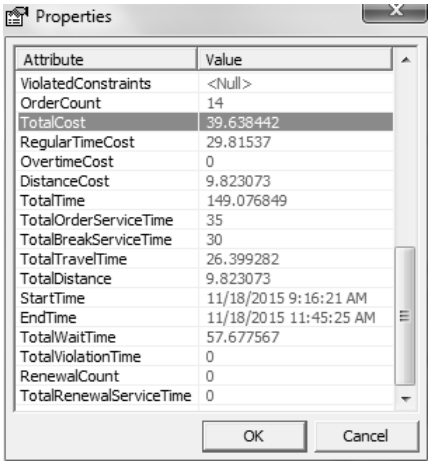


Fig. 10. Calculated fields of Left Route after solving with barriers

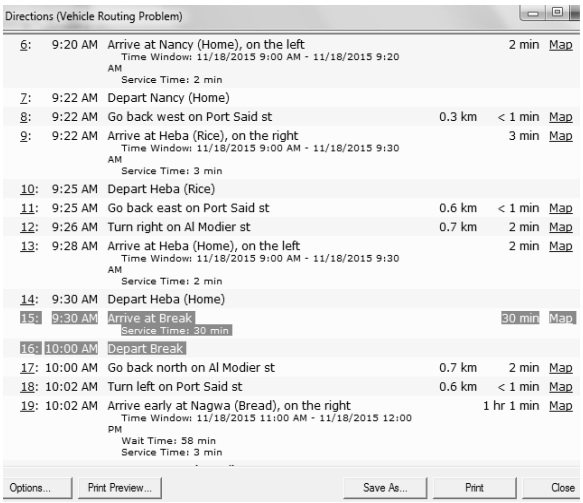


Fig. 11. Left Route before and after barrier

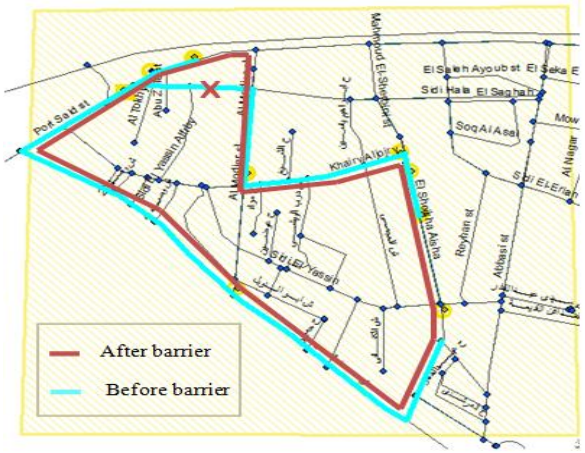


Fig. 12. Directions of Left Route after with barrier

VI. Conclusion

The problem of delivering conditional products in Mansoura city in Egypt is solved based on GIS. Using proposed approach, best route is determined easily as it can be printed and given to drivers to inform them about their schedule. OD cost Matrix is used for optimizing the solution as it calculates the nearest least cost depot to customer's destination to transmit requested orders from that depot to minimize travel time and travel cost. Lunch Break time of drivers are calculated automatically according to conditional products that are assigned to be delivered on every route. When barrier occurred on a street, system resolves the problem again and regenerates alternative route directions without time violations. The break time automatically recalculated and may be changed.

References

[1] Esri. (2012, July) esri. [Online]. http://www.esri.com/library/bestpractices/what-is-gis.pdf

[2] A.Prakash, Geographical Information System, An overview. India: Indian Institute of Information Technology.

[3] P.Keenan, "Modelling vehicle routing in GIS," springer, 2008.

[4] Jean-Paul Rodrigue Shih-Lung Shaw. people.hofstra.edu. [Online]. https://people.hofstra.edu/geotrans/eng/methods/ch1m4en.html

[5] M.E.A.El-Mikkawy, B.T.Shabana A.M.Riad, "Real Time Route for Dynamic Road Congestions," International Journal of Computer Science Issues, 2012.

[6] A.Pandey, K.Pandey, M.Saim V.Shukla, "V.SGIS-BASED SOLUTION OF SCHEDULING AND ROUTING SCHOOL BUSES- A THEORETICAL APPROACH," International conference on technologies for sustainability-Engineering Information technology, 2015.

[7] Yew Soon Ong, Chen Kim Heng, Puay Siew Tan, and Nengsheng Allan Zhang G.Kim, "City Vehicle Routing Problem (City VRP): A Review, IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS," IEEE, 2015.

[8] Michael H. Cole Y.Zhong, "Y.Zhong, Michael H. Cole, "A Simple Approach to Linehaul-Backhaul Problems: A Guided Local Search Approach for the Vehicle Routing Problem".

[9] E.Neftalí Escobar Gómez, F.Taracena Sanz, "THE VEHICLE ROUTING PROBLEM WITH LIMITED VEHICLE CAPACITIES," International Journal for Traffic and Transport Engineering, 2013.

[10] J.Yves Potvin, M.Gendreau J.Francois Cote, "The Vehicle Routing Problem with stochastic Two-Dimensional Items ," CIRRELT, 2013.

[11] S.Krichen S.Faiz, Geographical Information Systems and spatial Optimization., 2013.

[12] Esri. (2016) arcgis. [Online]. https://desktop.arcgis.com/en/desktop/latest/guide-books/extensions/network-analyst/vehicle-routing-problem.htm

[13] Esri. (2009, Dec.) arcgis. [Online]. http://www.arcgis.com/home/item.html?id=3b93337983e9436f8db950e38a8629af

[14] A.Vienneau, J.Bailey, J.Banning and S.Woo M.Harlow, "ArcGIS 9 – Using ArcCatalog," Esri, 1999.

[15] Esri. OD Cost Matrix. [Online]. http://webhelp.esri.com/arcgisdesktop/9.3/index.cfm?TopicName=Creating_an_OD_cost_matrix

[16] Esri. Types of Network analyst. [Online]. http://webhelp.esri.com/arcgisdesktop/9.3/index.cfm?TopicName=Types_of_network_analyses

[17] G.Mani, K.Pandey H.Shankar, "GIS Based Solution of Multi-Depot Capacitated Vehicle Routing Problem with Time Window Using Tabu Search Algorithm," International Journal of Traffic and Transportation Engineering, 2014.

[18] R.Fauzi, R.Muhamad M.Abousaeidi, "Geographic Information System (GIS) modeling approach to determine the fastest delivery routes," Saudi Journal of Biological Sciences, 2015.

[19] Chih-Hui Shieh, Kai-Ting Chan, Shin-Yi Chiu, Jyun-You Fan, Yu-Ting Chang, Nuo-Jhen Ma Rong-Chang Chen, "A Systematic Approach to Order Fulfillment of On-demand Delivery Service for Bento Industry," ELSEVIER, 2013.

[20] S.Sumathi Surekha P, "Solution To Multi-Depot Vehicle Routing Problem Using Genetic Algorithms," World Applied Programming journal, August 2011.

[21] Varun Singh, R.C. Vaishya V.K.Purwar, "GEOSPATIAL ANALYSIS OF REGIONAL MILK DISTRIBUTION FOR ALLAHABAD CITY," The International Daily journal, 2015.

Techniques used to Improve Spatial Visualization Skills of Students in Engineering Graphics Course: A Survey

Asmaa Saeed Alqahtani
Department of Computer Science
Najran University
Najran, Saudi Arabia

Lamya Foaud Daghestani
Department of Computer Science
King Abdulaziz University
Jeddah, Saudi Arabia

Lamiaa Fattouh Ibrahim
Department of Computer Science
King Abdulaziz University
Jeddah, Saudi Arabia

***Abstract*—Spatial visualization skills are crucial in engineering fields and are required to support the spatial abilities of engineering students. Instructors in engineering colleges indicated that freshmen students faced difficulties when visualizing models in engineering graphics. Students cannot correctly understand and process visual object and mental images of the engineering models. Traditional tools using textbooks, physical models, and modeling techniques is not sufficient for improving the spatial visualization skills of engineering students. This paper is a survey of all techniques used to learn freshmen students in engineering graphics and improve their spatial visualization skills. Also, it presents the method of evaluation the spatial visualization skills. After describing techniques and presented the literature review, this work presents a comparison between methodologies and techniques used in previous studies. Finally, we summarize a road of the map for the techniques and strategies to improve the spatial visualization skills for freshmen engineering students.**

Keywords*—3D graphics; virtual reality; spatial visualization skills; mental rotation skill; engineering graphics

I. Introduction

The most important skill in engineering fields is Spatial Visualization Skills (SVS) [1]. SVS defined as "the ability to generate, retain, retrieve and transform well-structured visual images" [2]. Spatial ability is the ability of the human brain to produce, keep, retrieve, and transform the 3D models, virtual images, and objects. Spatial ability related to the cognitive load of memory. If students improve their spatial skills, then there is no overloaded in the cognitive load [3].

To improve SVS, we need to improve Mental Rotation Skill (MRS) [4]. MRS is the ability to mentally transform 3D objects and virtual images by rotating it in the space (in mind). MRS need a cognitive process to allow a person to mentally (in space) transform by rotating 3D objects and virtual images [2]. MRS is classified of a type of spatial skills. Most of the studies conduct that the spatial skills are essential for the engineering students. Such as, in [1] conducts that the spatial visualization is related to the engineering fields, MRS is related to SVS to build 3D models, and the spatial skill is required for the computer graphics in engineering [1].

Instructors in engineering colleges lock solving the difficulties faced the freshmen students in studying engineering graphics course when they tried to visualize the models [5]. Students cannot easily understand and processing the visual objects and mental images [6]. Traditional tools using textbooks, physical models, and modeling techniques is not enough to graduate engineering students with high SVS, and they fail to pass the graphic courses with high grade [7].

The freshmen engineering students need to have high MRS and ability to mentally visualize the 3D models like it in reality [7]. The MRS improved by training students to use virtual models and interacting with it by using their eyes and hands [8], [9]. The training before studying the courses will allow students to improve their MRS and then study better with a little of difficulties. Engineering students also need to improve their SVS, and hence improving the working of the memory of human brain. Improving the working memory leads to high cognitive load in the brain [9].

The next section discusses strategies of MRS. Section 3 represents the techniques used to improve SVS and MRS. After that, Section 4 discusses the methods of evaluation SVS and MRS. Section 5, the literature review. Section 6 compares the papers worked in the field of enhancing the SVS. Section 7 the adaptation systems. Section 8 represents the gender differences in MRS. Section 9 summarize the road map of this survey. The paper's conclusion presented in section 10.

II. Strategies for Learning Mental Rotation Skill

There are three strategies for learning MRS in a 3D environment. The three cognitive styles are a holistic mental rotation, analytical mental rotation, and the combination of two previous strategies that called combined strategy [10].

A. Holistic mental rotation

The holistic mental rotation is rotating a 3D object mentally without considering for any features of the object such as color, coordinate axes, shadow, and size [10]. The holistic mental rotation is rotated for a 3D object without any steps for rotating it. The males are tending to use the holistic mental rotation and solving the mental rotation tests using this style by using the right hemisphere which is responsible for the holistic mental rotation [10]. Parallel style "tend to encode visual images globally as a single perceptual unit, which they process holistically" [11].

B. The analytical mental rotation

The analytical strategy based on rotating the object mentally step by step. The features of objects are considered to rotate the object mentally. The analytical mental rotation is used by females to solve the spatial visualization tests and tend to use the left hemisphere which is responsible for the analytical mental rotation [10]. Sequential style "tend to encode and process images analytically, part by part, using spatial relations to arrange and analyze the components" [11].

C. Combined mental rotation

The third type of mental rotation cognitive styles is a combined of holistic and analytic styles. The combined style allows learners to rotate a 3D object mentally by both the left and right hemispheres when rotating the object mentally. In this style, the user can mentally rotate the object with holistic processing and analytical processing based on some features of the object. The later studies by Li [12], [13], [14], [10], [15] noticed and concluded that the females are tending and preferring to use the combined strategy or analytical strategy.

III. Techniques Used to Improve Spatial Visualization Skills

The following sections describe the different techniques used to improve SVS.

A. Software tutorials

There were two electronic tutorials used for first-year students in engineering college to teach the engineering subjects in electronic methods rather than just based on textbooks. The two tutorials are Visual Reasoning Tutorial and Orthographic Projection Tutor. The Visual Reasoning Tutorial allows students to construct a solid model from two orthographic projections. The Visual Reasoning Tutorial has a glass box with the model inside it. The students need sweeping operations to construct that model [16].

B. Flash courseware

Flash courseware is an electronic course that allows interactive 3D elements to add in an electronic course designed by Flash with text and images. It constructed by using ActionScript and library of Sandy to build 3D elements animated using Flash [17].

C. E-Learning module

There is a developed E-Learning module based on "Computer Aided Interactive Learning of Engineering Graphics." This module aimed to utilize the computer abilities for computing the attributes of the geometric model such as plans, sections of solids, and perspective projections. It consists of many pages and exercises as it a traditional book, but it presented on a computer screen with a little of animation by the degree of angles for 2D orthographic and planes [18].

D. EBook, multi-touch screen technology

Electronic book or e-book is a digital tablet. The digital tablet produced by many companies and brand such as Windows, Apple, and Samsung. The digital tablet contains operating systems such as iOS or Android systems. These e-books (digital tablets) has multi-touch screen features. It allows the user to touch screen as it a smartphone. There are applications on it used to educate engineering students and improve their SVS [19].

E. Virtual reality

The virtual reality (VR) is a newly emerging computer interface characterized by high degrees of immersion, believability, and interaction, with the goal of making the user believe, that he is actually within the computer-generated environment [9]. According to [20], VR is " technology that allows us to create environments where we can interact with any object in real time, and that has been widely used for training and learning purposes." Integration some of the technologies, such as computer and graphics, can generate the technology of the VR. So, the VR can outline as it is a progressing computer interface to allow the person to be immersed within a simulated environment generated by a computer.

F. Augmented reality

Augmented Reality (AR) is one type of VR according to the feeling of immersion. The different types of the VR systems take place according to different using of technological supplies. Those various suppliers represented in various displayed hardware and interaction devices. "Virtual reality systems are classified according to the level of immersion they provide, ranging from semi-immersive virtual reality to fully immersive virtual reality to augmented reality (AR)" [9].

The non-immersive system often called desktop VR (without any input devices). It is based on the monitor screens as it is a window to the virtual world without an additional device. The AR is one type of the VR systems.

G. Portable document format

Portable document format 3D (PDF3D) is PDF format embedded 3D models. The PDF3D can read the 3D models. The 3D models embedded in PDF can manipulate by the user [21].

H. Web3D

The Web3D is the web contain 3D models as it VR and AR environments. The web3D designed using software such as AutoCAD, 3D website (browsers), X3DOM, Maya, 3D Studio, and Blender. Also, it can use a programming language to create 3D models such as virtual reality modeling language (VRML) [22].

I. The SketchUp software

The SketchUp produced by Trimble used to build a system allows students to put a paper contained 3D model designed by SketchUp in front of the camera, then the model constructed as it in reality. This system has a plugin such as AR [23].

J. Colored 3D models:

The Coloured 3D models used instead of monochrome models. It is a 3D printed model on paper and presented on a computer screen as static 3D models. The user wears blue glasses to see the shadow of 3D models on the screen of computer [24].

K. Training website:

Training website contained 3D standard models based on holistic mental rotation strategy. The website contained many levels [7].

IV. Evaluation Methods of Spatial Visualization Skills

A standard test of SVS is needed for each student before and after applying proposed techniques to measure improvement and enhancement of their SVS. There are many measures to test the MRS. The measures (standard tests) are Mental Rotations Test (MRT), Purdue Spatial Visualization Test (PSVT), Purdue Spatial Visualization Test: Rotations (PSVT: R), Online surveys, and Concurrent STEM course grades [25], [26].

The basic test is MRT. It founded in 1971 by Shepard and Metzler. A mental rotation test is a measurement tool for MRS and SVS. However, it used for general STEM fields. MRT is the most test has difficulties in calculating the result of students when they tested to measure their improved mental rotation. MRT required to calculate not result of the test, but also how much response time for each question with calculating error rate [27], [26].

One of the standard tests is special for academic using and the engineering field. The test is PSVT: R. We are going to use this test in our research. The PSVT: R is exceedingly used in different fields and programs of engineering in colleges and universities. The PSVT: R used in engineering learning and education because it based on isometric drawings for spatial visualization [2]. Moreover, PSVT: R used in engineering colleges because it "can be used as an assessment tool for diagnosing and improving students' spatial visualization skills in any engineering course that requires basic understandings of visual representation of objects" [28].

PSVT: R considered MRT, but it is developed in 1977 by Bonder and Guay with more facilities. MRT considered at most of the rotation skill only and all questions of the test based on ten cubes face-to-face as it is a rigid arm-like structure as shown in Fig. 1. The student decided if two objects are matched or dis-matched, or can choose one of three models which one rotated as the original model [29]. Moreover, the samples of questions are not available and need to use a computer to test it due to the needing to calculate the response time. However, PSVT: R consider both rotation and spatial skill with visualization ability.

The PSVT test consists of three parts that are "development" to see how the student can visualize the folding of 3D objects. The second part is "rotation" to see how the student can mentally rotate the 3D objects. The third part is "visualization" to see how the student can visualize 3D objects from on point view through glass cube [29]. The PSVT test is suitable for a student from 15 age and older especially the engineering students. The following Fig. 2, 3, and 4 show the PSVT test examples.

Fig. 2 shows the first part of PSVT that is development test. One object is not folding and just drown in 2D. Then the student needs to fold and visualize the structure of 3D objects. After that, the student selects one answer among five answers. In the example of Fig. 2, the correct answer is only B. Each part consists of 12 questions.

Fig. 3 shows the second part of PSVT: R that is rotation test. One object rotated in particular directions. The original and rotated 3D objects shown as an example for the student. Then another 3D object is displayed and ask the student to rotate it mentally as in the direction of example, and then select only one rotated 3D object among five answers. The correct answer in this example is only D.

Fig. 4 shows the third part of PSVT that is view test. A 3D object is drawn and asks the student to imagine it as in a glass cube, and he/she is around the cube. Then, the student has to move around the cube until he/she see the black dot between student and 3D object. After that, the student visualizes the structure of the subject matter according to his seeing of an object in the direction of the black dot. Select one answer among five 3D objects. The correct answer in the example of Fig. 4 is the only E.

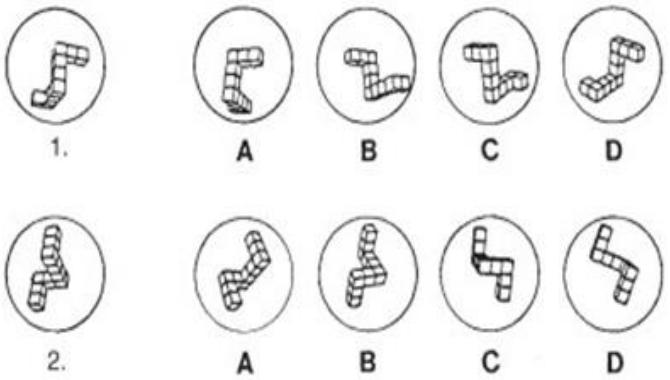


Fig. 1. Sample problem MRT

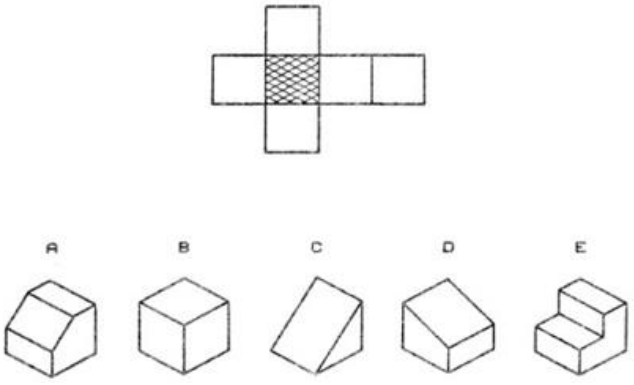


Fig. 2. PSVT test (development part), [30]

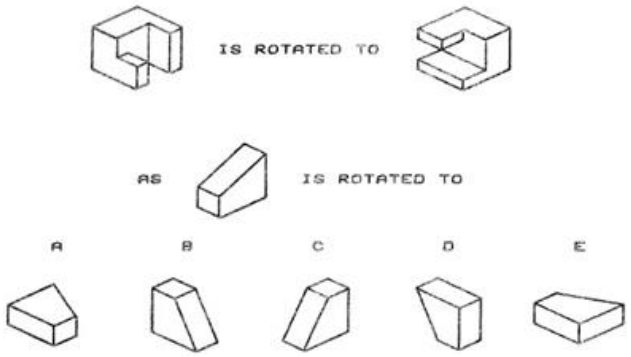


Fig. 3. PSVT:R test (rotation part), [30]

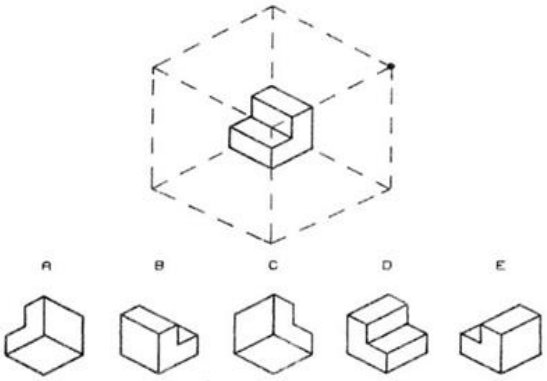


Fig. 4. PSVT test (view part), [30]

The part of the rotation is as shown in Fig. 3 that is PSVT: R. It is considered one of the standard tests of mental rotation [31]. In general, PSTV there are 36 questions, with 6 questions answered before all three parts (2 answered question as examples for each part).

In the PSVT: R, There are 30 questions with solving examples and the students have to answer it during 20 minutes by Guay in (1976) [28]. The PSVT: R available in the ETS test collection and has since been widely used by researchers in engineering and technology fields.

In 2009, before release the Revised PSVT: R, Smith [32] suggests adding the axes of coordinates to the questions of PSVT: R. Adding the axes is to represent the orientation of objects in space to minimize the effect of spatial visualization in the exam. Moreover, adding the coordinates is to because the students take twenty minutes to solve twenty question whether it thirty questions as Barnoff concluded in his research [27]. An example for PSVT: R with axes shows in Fig. 5.

V. Literature Review

In this section, we present the previous studies used to facilitate the education of engineering students and improve their SVS and MRS. All the following studies used one or many of techniques, mental rotation strategies, a method of evaluations mentioned in section 3 and 4.

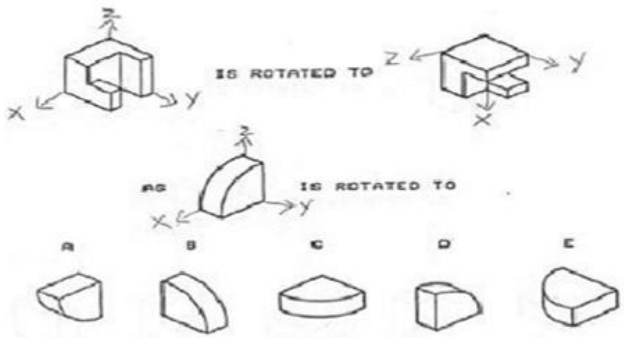


Fig. 5. Sample of PSVT: R with labels for axes

Esparragoza in [16], proposed to use two tutorials in many effective tools for the freshman students in Engineering College to teach and learn the engineering subjects in better methods rather than just based on textbooks. The two software tutorials are Visual Reasoning Tutorial and Orthographic Projection Tutor.

Visual Reasoning Tutorial allows students to construct a solid model from two orthographic projections. The Visual Reasoning Tutorial has a glass box with the model inside it. The students need sweeping operations to construct that model.

The first tutorial is Visual Reasoning Tutorial consist of many missing view problems to help students to complete the process of visual thinking and visual analysis. There is more than one solution possible for the same problem. The students can check their solutions by using assistant constructed in this tutorial. There are many sweeping operations in this tutorial with edges forming the face of a solid model. The students can have direct feedback after each sweeping and constructing operations.

The second tutorial is Orthographic Projection Tutor. It is a software also designed for the freshman engineering students to start solving problems at the highest level. Each problem in this tutorial has many exercises about the concepts of orthographic projections. The Orthographic Projection Tutor made the concepts and issues related and linked based on the definitions and concepts in orthographic projection by using a learning network.

The using of physical models in engineering graphics replaced in this research by those two tutorials Visual Reasoning Tutorial and Orthographic Projection Tutor.

The tutorials constructed by using AutoCAD and SolidWorks software to enhance the visualization skills of 3D presented in engineering graphs for the students.

On the other hand, the developed module was not enough to improve the visualization skills in engineering graphics. The proposed E-Learning module was like traditional books contain pictures and texts.

Manseur in [33], discussed the concepts of the VR and using of its tools in science in general and especially in

engineering education to encourage the using of visualization. The author presented software tools with many examples of using and developing the VR in sciences and engineering. The VRML discussed in this research with many languages that can integrate with it. To develop the using of VRML, new software tools integrated into a graphical programming language such as Visual C++, Visual Basic, Java, Labview or other. These tools with programming language have Graphical User Interface (GUI) to allow the students to interact with the VR model.

In engineering education, the usage of 3D visualization will enhance the teaching and learning of excellent modeling and simulation methods. Spazz3D is one programming software package that used effectively by integrating it with VRML. It has GUI to seek the input from students and introduce the result in efficient methods.

Using of VRML will assist the students to understand the six degrees of 3D space with a mathematical representation of a solid model. A small program called POS and RPY Animator created. The program allows the students to choose a solid model, entering the position and the value of orientation, and then having the virtual displays of the model. The students can move the object into its corresponding location.

Using VRML with graphical programming language allows the animation of 3D objects either by the user action or automatically. The animation in VRML will support the visualization of moving systems.

However, the program created in this research is just reflecting the vectors and the orientation of the 3D model in space. It lacks for the real visualization for the plans and the orthographic projection of a model.

In [17], Hong and Mie proposed a new method to allow interactive 3D elements to be added in Flash courseware. The authors construct this method by using ActionScript and library of Sandy and construct a 3D view of Sandy APIs and Flash functions management. The using of the new method implemented with engineering graphics courseware.

The proposed method applied to engineering graphics. The example presented to show the teaching of the intersection curve using Flash courseware and to how the intersection curve will be when the two cylinders intersect. It allows students to translate the solid object in x, y, and Z-axes by using the keyboard. Also, it allows students to rotate the camera around the axis.

This research presented a method to add interactive 3D elements into Flash courseware. SandyAPIs used to create the view of a 3D object in space. However, the proposed method needs to support to visualize the objects as real. It did not present the plans and projections of objects. Students still need to see, understand, and realize the model in space using VR.

In [19], Torre et al. proposed a functional prototype of the interactive multi-touch eBook with built-in digital capabilities drawing aimed at engineering graphics and visualization courses. The prototype targeted to enhance the concepts of engineering graphics and concepts of visualizations. The proposed multi-touch book consists of many exercises, textual elements, and rich media content to allow students to digitally sketching in the electronic environment like the traditional environments that include paper, pencil, and eraser.

The proposed prototype combines both eBook and interactive multi-touch screen features. The multi-touch digital book allows students to visualize and interact 3D models contained in the eBook. The concepts are about orthographic projections and multi-view drawing. It included interactive sketching exercises. The study applied to iPad only using Apple's iBooks. The prototype in a tablet is as widgets. The students can draw automatically on the screen as its real paper, and submit the solved exercises to the instructors immediately using e-mail. The widget that proposed consisted of many elements such as photo gallery, sketching environment, 3D models, and many exercises, videos to learn about covering orthographic, projection and multi-view drawings, and email to allow students to send exercises and assignments to the instructors and to get the feedback.

To evaluate the proposed prototype, a comparative study made between two groups of students. The first group study using traditional tools, while the second group studies the course in one semester using the eBook multi-touch screen. Also, the authors made questionnaire about the prototype. The results showed positive reaction and approval.

On the other hand, the prototype specified for iPad from Apple's Company only. The students need to see and visualize the models as it in the real world with visual height and width to rotate its plans by students themselves as its touchable models.

In [21], Martin et al. aimed to develop educational material based on AR formats and several virtual. The authors also proposed to recognize how students behave while using teaching materials based on AR formats and several virtual, and checking if they are useful materials to improve their spatial skills and capabilities.

The work presented three different technologies that are AR, VR, and PDF3D. The three techniques applied to three groups of students in the same classroom and the same level. One of these technologies applied to all three groups and then studied to find out the spatial ability progress of the student, and discovering the impact of the tool used in the acquisition of graphic design knowledge. There is a fourth group (control group) that will not use any of three technologies. They will use the traditional methodology of teaching and learn the concepts and models in engineering graphics that based on the textbook and real models only.

The AR technology uses either direct or indirect view of a physical environment of the real world and uses a set of devices to combine virtual information and material information that already exists. This study uses distinct AR from VR. The exercise in this material based on AR constructed using BuildAR Pro AR application. BuildAR Pro AR allows to create scenes consists of the set of images or marks that modify a 3D object. After creating the scene, a webcam on the computer recognizes an image that related to the 3D model and showed it integrated into the real world.

The portable document format included in PRO-X version. This material has many features that are an open standard,

multiplatform, extendable, reliable and secure, sophistication for information integrity, search capability, accessibility, and interactive.

All three technologies have the same goal that the student gets to know the piece instead of a real physical model and allow the student to get the information needed to sketch the piece and create a workshop contour plan.

The authors present a pilot study to compare the results of improvement in spatial ability need for freshman engineering students. The research also showed a survey about satisfaction and motivation of the proposed methodology using the three technologies. The study did in the graphic design laboratory.

Based on the results of motivations using a survey, the students reached a good level of development in the learning the visual skills. However, the students made inappropriate use of the strategies of the proposed technologies. They still need to the existence of instructor to guide them.

Another study used AR technique by Gutierrez [34] suggested another solution using mobile augmented reality (MAR) instead of using desktop VR by AR. Gutierrez and his colleagues designed an application called DiedricAR based on MAR. The application aimed to use for descriptive geometry which based on (graphical language engineering).

The application designed to support the autonomously and ubiquitous learning. The DiedricAR system used by students to learn how to solve dihedral exercises (drawing exercises) and showing 3D models based on its 2D orthographic in the workbook (augmented book). The workbook is the traditional book contains exercises and pages, but with the possibility to capture the exercise statement and showing it a solution with 3D models on the screen of smart devices.

The DiedericAR consists of a workbook, models, and markers. The models are 3D models displayed on the screen of smart devices based on the orthographic projection exercises in the book.

The system of smart devices applied to the DiedrcAR application Android and iOS. However, the system tested only on Android system that is "Samsung Galaxy Tab 10.1; Samsung Galaxy SII; Samsung GT-P- 1000; Samsung Galaxy Ace; LG Optimus L3; Samsung Galaxy 3 GT i-5800".

The DiedricAR system used by 20 students selected randomly. This group of a student called group B, while another group (group A) was not using the system. The Gropu A take the course with traditional materials only.

Two mental tests used for both groups that are: PSVT: R and Differential Aptitude Test (DAT-5: SR): spatial visualization. Also, a questionnaire introduced to group B to assess them satisfy about the DiedricAR system. The grade of tests shows that group B improved by 17% more than another group (group A).

On the other hand, the system still needs to support the tactile module. The students need to visualize the 3D models and their orthographic projection in virtual space as it is in the real world. The using of smart devices required using the camera and the textbook. Moreover, there are no need to use the handle drawing nowadays; especially we live in the technologies and technical world.

Another research study in 2015 by Gutierreza et al. [35] aimed to create an AR based on the didactic content of engineering graphics course. The AR application was instead of traditional material such as traditional textbook, using paper and pencil, and modeling using SketchUp software, and online multimedia web-based exercises.

The designing of AR application constructed by using AR, USB Camera (QuickCam Pro 9000), AR-Dehaes, and computer vision techniques. The AR application has many levels and exercises in the AR book.

The AR application has software that displays 3D models on a computer screen by capturing it from the paper of the book. By capturing the model, it's constructed as it 3D model on a computer screen. Moreover, the levels of AR application have videos to explain the orthographic concepts and freehand sketching. The notebook of exercises contained on AR application. The exercises such as find wrong in 2D orthographic views, draw one view of orthographic projections based on existing two views of the same model and draw 3D models (perspective) based on the orthographic projections.

The AR application designed for engineering students at the La Uaguna University in Spain.

The experiment applied to the training group of students. The students tested two tests that are (DAT-5: SR Level 2) and (MRT). Also, the researchers introduce a survey for the students to evaluate the AR application (usability and satisfaction). The results of mental tests and questionnaire conclude that the training students have high marks and their spatial ability improved through the training system.

On the other hand, the system focus on training the students how to draw 3D perspective with its projection views, while there are a lot of computer programs can do it easily and with less effort and time of works. Likewise, the students cannot scan and view the 3D models as it in real by rotating, translate, flying and zooming. The students still need to learn from the tactile model of learning (by the

In [7], the study used VR and one strategy of mental rotation strategies. This study used holistic mental rotation for both genders. That is, there is no any features for the 3D models. The rotation of 3D models was based on choosing a number of angels. It was website training contained many levels.

The software and tools used for building the website were HTML, SQL database, Amazon server, and JavaScript. The hardware were laptops and desktop computers.

The evaluation of this study used Revised PSVT: R test. The outcomes of this study were improving in students score of PSVT: R test. The "increase for males in the experimental group was 1.72 times greater than that of the comparison group (experimental group (N=75), and the comparison group (N=134))" and "The increase for females in the experimental group was 2.45 times greater than that of the comparison group. (experimental group (N=19), and the comparison group (N=29))".

On the other hand, the strategy of this training is supporting for the males more than females. The 3D models rotated in axes directions only and based on precise angles degree. The researcher used all models of PSVT: R in training. It should be two sets of models (training sets and testing sets).

Another study to solve the problem of spatial visualization in engineering graphics is done by Olmedoa et al. [22] in 2015. Olmedoa and his colleagues assumed the necessary of interacting with 3D models to obtain good results. Olmedoa suggested building web3D for university-level students. To build web3D, the researchers aimed to use VR/AR, and they implemented it.

The web3D contains many 3D models. It designed by many software and tools such as CAD, VRML, Catia, 3D website (browsers), X3DOM, Maya, 3D Studio, and Blender to form web3D for engineering students. By those tools, the researchers built VR/AR environment. Many universities connect to this web3D such as the University of the Basque Country. It allows students to interact the models by rotating it and moving it only without zooming the models.

The web3D system (AR/VR system) used with other traditional techniques used in education, such as Blackboard, PowerPoint presentation, and textbooks. The system is used only in the labs of universities.

After using the web3D by students, a survey introduced about the advantages and disadvantages of VR system based on web3D. The results were as the following:

- 70% and more of the students agreed with the advantages of the web3D system. It has improved the student's spatial ability rather than imagining a 3D model when it's constant (printed) on paper.
- 70% and more were asking to have the system on their personal laptops and PCs, and in classrooms, not only in labs through university's connections.
- 50% of students was not agreed. They have opined that the system has disadvantages because the system does not offer permanently and not installed on their personal computers. They need to use and practice it at home. Moreover, they take the time to wait for the connection between university and web3D.
- Also, the 50% students complain that the models cannot zoom for more viewing and scanning.

Katsioloudis et al. in [24], published research in 2016 about how to develop the spatial ability of engineering students using the impact of colored 3D models (by using colored 3D models). They assumed based on past researches that the information has color will improve the cognitive load and help students to get more information without losing it easily. So, the researchers suggest using colored 3D models instead of monochrome models (black or white models). However, the using of more color will negative effect spatial ability and cognitive load.

This study aimed to use colored 3D models for three control groups of students that are: the first group used 3D printed model without color. The second group used the same printed 3D models, but also wearing blue glasses. The third group used PC contains 3D blue models with its shadow on the screen.

The spatial visualization ability was measured using two data collection instruments that are MCT (Mind-Consciousness-Thought) and the creation of section view drawing. Then the results analyzed using ANOVA to find the mean and significant differences for each group.

The results and the average of each group are very similar without significant differences. The researchers after concluding that there are no differences assumed that the spatial ability would improve only for students who have low ability. Moreover, the researchers believe that "the population used (engineering technology students) did not demonstrate a statistically significant difference in spatial abilities from the addition of the color because spatial skills were well developed in this population" [24].

On the other hand, the research lacked to use the VR to create models and allow students to visualize it, rotate it, and scan the models as in the real. There is a need to add a third channel of learning that is (tactile module) by using the hands. The previous research was based only on the visual module (using eyes only). There is a need to use all three channels of learning based on the VR model for Cognitive learning [9]. We assumed to use that model with affecting of brightness and shadow of color 3D models.

Another prominent researcher is Sorby. She has been doing many types of research since 1999 till 2016 to improve the SVS of engineering students and comparing between the females' and males' skills when they are doing the mental rotation tests. In her researches [36], [37], [1], [38], [39], she was trying to change the content of engineering courses. Also, changing the methods of teaching and using the modeling programs such as CAD programs to improve the SVS of students. The researcher did not yet introduce a training system specific for improving the MRS and independent of the contents of the course. However, the researcher concluded that there is a difference in MRS depends on the gender.

VI. Comparison between all Papers Worked in the Field of Enhancing SVS

This section compares all previous studies based on many criteria such as if the paper used technique based on any strategy of MRS, whether paper used random 3D models or models from exercises of course or standard models used in measurement tools as PSVT. Moreover, the comparison shows whether paper proposed system as it course or system training, and what gender targeted for its study.

We can see in Table 1 that only one paper used mental rotation strategy. The black box means that paper satisfies the property. However, it was the holistic mental rotation and this strategy not suitable for females. In the second column, no paper based on any adaptation system to make the training fit for every student's level and skill of learning. In the next section, we present more detail about the adaptation systems.

TABLE I. COMPARISON BETWEEN PAPERS PRESENTED IN LITERATURE REVIEW

Ref.	Technique used in papers	Based on strategy of MRS	Adaptation system	Type of 3D model			Type of system		For improving SVS or MRS?
				Selected randomly	Course models	Standard models	course	Training system	
[16]	Software tutorials			■			■		SVS
[18]	E-learning module				■		■		SVS
[19]	eBook, multi-touch screen technology				■		■		SVS
[21]	VR, AR, and PDF3D			■				■	SVS
[33], [34]	AR only				■		■		SVS
[35]	SketchUp with AR				■		■		SVS
[7]	VR, Training website	■				■		■	MRS
[22]	Web 3D			■				■	SVS
[24]	Colored 3D			■				■	SVS
[17]	Flash Courseware				■		■		SVS

VII. ADAPTATION SYSTEM FOR LEARNING

Each learner has different abilities and skills in learning. Some of the learners get the information and understand it in short time based on its prior knowledge, abilities, and learning skills. On the other hand, some students learn with a long time to process the received information based on many factors such as age, prior knowledge, abilities, and learning skills.

According to that differentiation in getting and processing information, there is an electronic system can change according to the level of each learner. It can be adapted based on each student level of training and learning. The system is called Adaptive Hypermedia System (AHS) [40]. The AHS offers more attention for the individual learning [41].

The AHS is based on some Artificial Intelligence (AI) technologies [42].The AHS is an electronic system to support adaptive (dynamic) content for each learner [40]. The AHS used in the educational area to offer appropriate content and information for each student, allow each learner to train based on his level, increase the satisfaction of learner, and improve the efficiency (learn in a short time) and effectiveness (assessment results) of learning [43].

The AHS used on the web and in the educational area [44]. This system "build a model of the individual user and use it to adapt the content or/and the hyper-structure of the pages in a hypermedia environment" [45]. There are four categories of AHS. The categories are [46] (Adaptive Interaction, Adaptive Course Delivery, Content Discovery and Assembly, and Adaptive Collaboration Support).

There is only one category used in the educational area and allow the content (material) of the system to be changed for each learner. The category is (Adaptive Course Delivery) which use many techniques applied in learning and educational environments. The content adapted for each student based on the information and lever of each one. This category "intended to tailor a course (or, in some cases, a series of courses) to the individual learner." Adaptive Course Delivery used to adapt between contents of courses and student requirements or level [46]. Therefore, and based on the features of this category, we used it in building our system of training.

The AHS used in learning environments has three basic models. The models are domain model (DM), user (learner) model (UM), and adaptation model (AM) [40], [46], [45], [42].

Domain model (DM): the domain model (application model) is the course produced to the learner in AHS of e-learning environments. The DM represent a description of how the content information is built and represent the relationship between the course elements (pages and information).

User model (UM): the UM represent the information of users such as result and levels of training, prior knowledge, goals, and navigation history of each learner. The UM updated based on the history of knowledge and navigation. DM represents the user (learner) knowledge and history of navigation. The UM represent the relation between the student and the DM by keeping track of navigation of the learner among the content of the system.

Adaptation model (AM): the AM represent what can be adapted, when and how it is adapted. The AM consists of adaptation rules to fit between content information and navigation of user. The AM used as a guide for the learner to navigate among the pages and links on the web. The adaptation is performed based on UM and DM by the adaptation rules

which presented in AM. The AM consist a set of generic and specific rules. The general adaptation rules are used to represent variables that are concepts and concepts relationship. However, the specific adaptation rule is concert concepts (not variables) and must be defined by the author.

VIII. Gender Differences

There are differences in SVS especially MRS between females' and males' skills. Many researches by Li [12], [13], [14], [10], [15] concluded that males outperform females in a most of spatial tasks, particularly when they involve mental rotation tests. "Males tend to outperform women on spatial reasoning tests significantly. Differences have been attributed to a multitude of factors including biological, social and cultural, and educational factors and are believed to contribute to the fact that males outnumber females in science and mathematics fields" [15].

Another studies by Sorby [47], [48], [1], [49], [38] and studies by Yoon and Maeda [50], [51], [52], [2], [53] and [54] concluded that there is a gender significant difference in MRS and when they solve mental rotation tasks. Males outperform females on mental rotation tasks due to the differentiation in using mental rotation strategy that males using holistic strategy while females using analytical strategy. Other researchers [55], [56] conclude that the females SVS needed to be improved according to their strategy (analytical style) to overcome the difference between them and males' score of MRS. That is, many of females tend to use their both hemispheres (combined style) and many of females using the left hemisphere (sequential style) that is an analytical strategy when they solve the mental rotation tests. The analytical strategy for mentally rotating a 3D object is based on rotating the object with step by step way and based on features of the object such as all three axes rotated with the object, colors, shapes, and sizes.

IX. Summary of Road MAP for Design Better System for Training on MRS

We can summarize that the freshmen engineering students need to improve their SVS by many strategies and not only by using interactive 3D models. The guidelines to build a system training are the following:

- Training on mental rotation using VR technology to allow students to learn on models as it in reality.
- Using holistic strategy for training male gender and combined strategy for female gender is the most appropriate mental rotation strategies.
- Using adaptation systems to adjust the training exercises to fit the level of each student.
- Do training at the beginning of the semester to overcome the difficulties while studying engineering courses.
- The 3D models have to be standard models such as models produced by Purdue in PSVT.
- It is better to be target female gender because it has lower spatial abilities than males.

X. Conclusion

Engineering students need to have high MRS as it ensures the high SVS that enables students to pass and understand the engineering graphics courses. The freshmen students in engineering fields are facing difficulties to how to imagine any 3D model. The male outperforms females in MRS. So, the first-year students women engineering students need to enhance their spatial abilities.

In this paper, we illustrate the technique used to improve the SVS in general, and MRS mainly. We represented the previous studies tried to improve the SVS, and we represented the related strategies and method for that improving. A comparative study concluded for the solutions proposed for enhancing the spatial skills of engineering students.

A road map for improving the MRS for female engineering is finally presented based on the discussed solution. For the future work, we advise building system training enhancing the MRS for freshmen females' students. The training system must be gathering the advantages of previous solutions and based on the suggestions of road map design we can build a better system for training on MRS.

References

[1] S. A. Sorby, "Developing 3-D spatial visualization skills," Engineering Design Graphics Journal, vol. 63, no. 2, 2009.

[2] Y. Maeda and S. Y. Yoon, "A meta-analysis on gender differences in mental rotation ability measured by the Purdue spatial visualization tests: Visualization of rotations (PSVT: R)," Educational Psychology Review, vol. 25, no. 1, pp. 69–94, 2013.

[3] G. J. Grant, "Impact of dynamic graphics on mental rotation of 3D objects with undergraduate students of varying levels of spatial ability," 2016.

[4] J. V. Ernst, D. Lane, and A. C. Clark, "Pictorial Visual Rotation Ability of Engineering Design Graphics Students," Engineering Design Graphics Journal, vol. 79, no. 1, pp. 1–13, 2015.

[5] E. S. Uria and M. G. Mugika, Methodology for Part Visualization Problem Solving-the Importance of the Process. INTECH Open Access Publisher, 2010.

[6] C. Özgen, "STRATEGIES AND DIFFICULTIES IN SOLVING SPATIAL VISUALIZATION PROBLEMS: A CASE STUDY WITH ADULTS," 2012.

[7] P. Cole, "Measuring the Effectiveness of Software-Based Training to Improve the Spatial Visualization Skills of Students in STEM Disciplines in Higher Education Institutions," 2016.

[8] J. A. Jimoh, "Comparative Effects of Two and Three Dimensional Techniques of Autocad on Spatial Ability, Interest and Achievement of National Diploma Students in Engineering Graphics," 2010.

[9] L. Daghestani, "The Design, Implementation and Evaluation of a Desktop Virtual Reality for Teaching Numeracy Concepts via Virtual Manipulatives," 2013.

[10] Y. Li, S. Wu, J. Zhu, and M. W. O'Boyle, "Sex and ability differences in neural activation for disembedding figures: An EEG investigation," Learning and Individual Differences, vol. 35, pp. 142–146, 2014.

[11] O. Blazhenkova and M. Kozhevnikov, "The new object-spatial-verbal cognitive style model: Theory and measurement," Applied cognitive psychology, vol. 23, no. 5, pp. 638–663, 2009.

[12] Y. Li and M. W. O'Boyle, "How sex, native language, and college major relate to the cognitive strategies used during 3-D mental rotation," The Psychological Record, vol. 58, no. 2, p. 287, 2008.

[13] Y. Li and M. W. O'Boyle, "Differences in mental rotation strategies for native speakers of Chinese and English and how they vary as a function of sex and college major," The Psychological Record, vol. 61, no. 1, p. 2, 2011.

[14] Y. Li and M. O'Boyle, "How sex and college major relate to mental rotation accuracy and preferred strategy: an electroencephalographic (EEG) investigation," The Psychological Record, vol. 63, no. 1, p. 27, 2013.

[15] Y. Li and M. O'Boyle, "Mental Rotation Strategies Used by Males/Females and Physical/Social Science Majors: An EEG Study."

[16] I. Esparragoza, "Enhancing Visualization Skills in Freshman Engineering Students," in Proceedings from the 59th Annual Meeting and Conference of the ASEE Engineering Design Graphics Division, 2004, pp. 21–23.

[17] Li-hong Luo and T. Xia-mei, "A Method of Using Interactive 3D Element in Flash Courseware," in E-Learning, E-Business, Enterprise Information Systems, and E-Government, 2009. EEEE'09. International Conference on, 2009, pp. 65–67.

[18] M. N. Raol and V. Vaithyanathan, "Computer Aided Interactive Learning of Engineering Graphics-An E-Learning Module," in Computing, Communication and Networking, 2008. ICCCn 2008. International Conference on, 2008, pp. 1–4.

[19] J. de la Torre, J. L. Saorin, M. Contero, and J. Dorribo-Camba, "Interactive sketching in multi-touch digital books. A prototype for technical graphics," in Frontiers in Education Conference, 2013 IEEE, 2013, pp. 190–194.

[20] A. Rodriguez, B. Rey, M. Clemente, M. Wrzesien, and M. Alcañiz, "Assessing brain activations associated with emotional regulation during virtual reality mood induction procedures," Expert Systems with Applications, vol. 42, no. 3, pp. 1699–1709, 2015.

[21] M. G. Dominguez, J. Martin-Gutierrez, C. R. Gonzalez, and C. M. M. Corredeaguas, "Methodologies and tools to improve spatial ability," Procedia-Social and Behavioral Sciences, vol. 51, pp. 736–744, 2012.

[22] H. Olmedo, K. Olalde, and B. Garcia, "MotoStudent and the Web3D," Procedia Computer Science, vol. 75, pp. 84–94, 2015.

[23] N. A. A. Gonzalez, "How to Include Augmented Reality in Descriptive Geometry Teaching," Procedia Computer Science, vol. 75, pp. 250–256, 2015.

[24] J. Petros, M. Jones, M. R. Moustafa, and V. Jovanovic, "Application of Color on 3D Dynamic Visualizations for Engineering Technology Students and Effects on Spatial Visualization Ability: A Quasi-Experimental Study," 2016.

[25] J. Martin-Gutierrez, R. E. N. Trujillo, and M. M. Acosta-Gonzalez, "Augmented Reality Application Assistant for Spatial Ability Training. HMD vs Computer Screen Use Study," Procedia-Social and Behavioral Sciences, vol. 93, pp. 49–53, 2013.

[26] D. I. Miller and D. F. Halpern, "Can spatial training improve long-term outcomes for gifted STEM undergraduates?," Learning and Individual Differences, vol. 26, pp. 141–152, 2013.

[27] T. J. Branoff, "Spatial Visualization Measurement: A Modification of the Purdue Spatial Visualization Test-Visualization of Rotations.," Engineering Design Graphics Journal, vol. 64, no. 2, pp. 14–22, 2000.

[28] S. Islam, "Assessment of Spatial Visualization Skills in Freshman Seminar."

[29] M. Harle and M. Towns, "A review of spatial ability literature, its connection to chemistry, and implications for instruction," Journal of Chemical Education, vol. 88, no. 3, pp. 351–360, 2010.

[30] R. Guay, Purdue spatial visualization test. Purdue University, 1976.

[31] S. A. Sorby and N. Veurink, "Long-term Results from Spatial Skills Intervention among First-Year Engineering Students," in Proceedings of the 65th Midyear Meeting of the Engineering Design Graphics Division of ASEE, 2010.

[32] M. E. Smith, "The Correlation Between a Pre-Engineering Student's Spatial Ability and Achievement in an Elecetronics Fundamentals Course," All Graduate Theses and Dissertations, p. 254, 2009.

[33] R. Manseur, "Virtual reality in science and engineering education," in Frontiers in Education, 2005. FIE'05. Proceedings 35th Annual Conference, 2005, p. F2E–8.

[34] E. G. de Rave, F. Jimenez-Hornero, A. Ariza-Villaverde, and J. Taguas-Ruiz, "DiedricAR: a mobile augmented reality system designed for the ubiquitous descriptive geometry learning," Multimedia Tools and Applications, pp. 1–23, 2016.

[35] J. Martin Gutierrez, M. Contero, and M. Alcaniz, "Augmented reality to training spatial skills," Procedia Computer Science, vol. 77, pp. 33–39, 2015.

[36] S. A. Sorby, "Spatial abilities and their relationship to computer aided design instruction," age, vol. 4, p. 1, 1999.

[37] S. A. Sorby, "Developing 3D spatial skills for engineering students," Australasian Journal of Engineering Education, vol. 13, no. 1, pp. 1–11, 2007.

[38] M. A. Sadowski and S. A. Sorby, "A Delphi Study as a First Step in Developing a Concept Inventory for Engineering Graphics," in Engineering Design Graphics Division 66th MidYear Meeting Proceedings, 2012.

[39] T. Delahunty, S. Sorby, N. Seery, and L. Pérez, "Spatial Skills and Success in Engineering Education: A Case for Investigating Etiological Underpinnings," 2016.

[40] H. Wu, G.-J. Houben, and P. De Bra, "Supporting user adaptation in adaptive hypermedia applications," in Proceedings InfWet2000. Rotterdam, the Netherlands, 2000.

[41] Y. E. A. Mustafa and S. M. Sharif, "An approach to adaptive e-learning hypermedia system based on learning styles (AEHS-LS): Implementation and evaluation," International Journal of Library and Information Science, vol. 3, no. 1, pp. 15–28, 2011.

[42] S. Aammou, M. Khaldi, A. Ibrahimi, and K. El Kadiri, "Adaptive hypermedia systems for e-learning," in Education Engineering (EDUCON), 2010 IEEE, 2010, pp. 1799–1804.

[43] E. Popescu, "Dynamic adaptive hypermedia systems for e-learning," 2008.

[44] D. Barac, Z. Bogdanovic, A. Milic, B. Jovanic, and B. Radenkovic, "Developing adaptive e-learning portal in higher education," in Toulon-Verona Conference" Excellence in Services," 2015.

[45] T. Tsandilas, "Adaptive Hypermedia and Hypertext Navigation," 2003.

[46] A. Paramythis and S. Loidl-Reisinger, "Adaptive learning environments and e-learning standards," in Second european conference on e-learning, 2003, vol. 1, no. 2003, pp. 369–379.

[47] S. A. Sorby, T. Drummer, K. Hungwe, and P. Charlesworth, "Developing 3-D spatial visualization skills for non-engineering students," in Proceedings of the 2005 American Society for Engineering Education Annual Conference \& Exposition, 2005, vol. 10, pp. 1–10.

[48] S. A. Sorby, "Assessment of a' New and Improved' Course for the Development of 3-D Spatial Skills," Engineering Design Graphics Journal, vol. 69, no. 3, 2009.

[49] N. Veurink, A. Hamlin, J. Kampe, S. Sorby, D. Blasko, K. Holliday-Darr, J. Trich Kremer, L. Abe Harris, P. Connolly, M. Sadowski, and others, "Enhancing Visualization Skills-Improving Options aNd Success (EnViSIONS) of Engineering and Technology Students," Engineering Design Graphics Journal, vol. 73, no. 2, 2009.

[50] S. Y. Yoon, "Psychometric Properties of the Revised Purdue Spatial Visualization Tests: Visualization of Rotations (The Revised PSVT-R).," ProQuest LLC, 2011.

[51] A. C. Medina, H. B. Gerson, and S. A. Sorby, "Identifying Gender Differences in the 3-D Visualization Skills of Engineering Students in Brazil and in the United States.," 1998.

[52] Y. Maeda and S. Y. Yoon, "Measuring Spatial Ability of First-Year Engineering Students With the Revised PSVT: R," in American Society for Engineering Education, 2011.

[53] Y. Maeda and S. Y. Yoon, "Are Gender Differences in Spatial Ability Real or an Artifact? Evaluation of Measurement Invariance on the Revised PSVT: R," Journal of Psychoeducational Assessment, p. 0734282915609843, 2015.

[54] S. Y. Yoon and K.-H. Min, "College students' performance in an introductory atmospheric science course: associations with spatial ability," Meteorological Applications, vol. 23, no. 3, pp. 409–419, 2016.

[55] J. L. Mohler, "Examining the spatial ability phenomenon from the student's perspective," 2006.

[56] T. Tseng and M. Yang, "The role of spatial-visual skills in a project-based engineering design course," 2011.

Selection of Mathematical Problems in Accordance with Student's Learning Style

To the level of engineering by using an expert system

Elena Fabiola Ruiz Ledesma
Escuela Superior de Cómputo
Instituto Politécnico Nacional
México

Juan J. Gutiérrez García
Escuela Superior de Cómputo
Instituto Politécnico Nacional
México

***Abstract*—This article describes the implementation and development of an expert system as a support tool to tackle mathematical topics, by using Bayesian networks as engine of inference and a learning styles, as well as the difficulty level of problems and establish the base of the classifier probabilistic. The expert system makes decisions as which element to visualize at a specific moment, gives the student the best resource, and supervises the user progress. The article is divided into three sections, the first one deals with the construction of the expert system, the second one presents the operation of the system through the classification of students consistent with their profile, which is based on the prevailing learning style among them and in the difficulty level that problems have so that the student reaches in solving successfully. It also shows the operability of the system in respect of the allocation of digital resources in accordance with the identified profile and gradually provides more assignments with different difficulty levels, as the student progresses. An experimental study was performed by means of which the system was assessed under 30 students to the level of engineering and those who studied the Applied Calculus course in their second semester of the degree course. This group was named the study group (SG). The SG used the system for one semester. The results at the initial and final evaluation were from 3.58 to 7.37 for CG and SG respectively. Applying the F test, a statistically significant difference in increase was found (p <0.002). These results showed that SG identified the concept of derivative and applied that concept correctly in real problems solving correctly 74% of the final questionnaire, so it is concluded that the system expert opens a new way in educational research.**

Keywords—technology; research projects; education; learning

I. Introduction

Learning based on traditional computing systems and mobile devices converge in its use in what is called Mobile Learning also known as M-learning, by allowing users to stay connected to the learning environment, learning resources, members of the educational community, such as teachers and students, no matter where they are [1].

Thus, the learning process is no longer bounded to a specific location and will depend essentially of the student's willingness to accede to learning resources. The problem occurs when the student does not have enough incentive to find out the available content offered by some current systems, either because they have very complex exercises, or on the contrary, they have very simple ones, or due to they do not have a variety of activities that help them to understand that concepts as these are presented in only one form for all students, even though each of them may have different styles of learning, as noted by Felder and Silverman [2].

The applications of the learning systems are quite extensive, there are several e-learning platforms that allow to structure content based on a large variety of multimedia resources, however, they present each element in static way, they do not make decisions or determine which element visualizes in a certain moment; they do not check which is the most suitable. Furthermore, they do not oversee adequately the progress of the user. This fact supports the study of areas such as artificial intelligence applied to learning within the framework of e-learning, since in several contexts its efficient functioning has been demonstrated [3].

The field of expert systems (also associated with systems of knowledge representation or knowledge engineering) are an important part of the field of Artificial Intelligence. Fundamentally, they represent the knowledge of the expert in the domain (area of expertise), Being applied, including as a tool by the very same expert [4].

As part of the research reported in this article, a survey was carried out on a sample of 50% of the students from the Escuela Superior de Cómputo (ESCOM), an Academic Unit of the Instituto Politécnico Nacional (IPN), and there were obtained the following results:

1) Generally, students are used to the e-learning as well as the m-learning,

2) More than 80% of the sample considers that an online application, which supports learning and it could help in their learning process,

3) 90% of the students expressed their willingness to find out more about the use of mobile devices in the learning process.

4) More than 80% of the students have failed a basic training course (Mathematics or Physics) and they consider that it would be very useful to have a website that allows them t1o solve problems and practice what they have learned in class, as this would allow them to face better the exams they solve at their on-site courses.

In this study we also found that students know some sites on the Internet that offer resources either to reinforce

knowledge, such as videos that explain problem solving, or to practice what they learn, through exercises to solve, but point out that it does not adapt to their level since the problems are of equal degree of difficulty for all the users and they think that they would solve exercises and easy problems in the beginning and that the degree of complexity is increasing. They also feel that they would feel more relaxed if there is a supervision of what they do to be able to review their progress.

The present work intends to develop a support tool for the teaching of some topics that are worked in the Calculus, using Bayesian networks as the engine of inference and for the determination of the profiles of the students will be considered the registers of semiotic representation [5], which will be the basis of the probabilistic classifier

B. Expert Systems

An expert system can be defined as a computer system that simulates human experts in a given area of expertise. As such, an expert system should be able to process and store information, learn and reason in deterministic and uncertain situations and communicate with users and / or other expert systems, make appropriate decisions, and explain why they have taken such decisions.

This kind of systems encode a knowledge base and reasoning rules to determine or conclude the solution of a particular problem. Are formed by various interrelated parts: a rule base, a base of facts, an inference engine and an user interface [6].

The knowledge of the expert is represented by the rule base that are generally of the form $R_i : P_r(x) \Rightarrow C(x)$, where $P_r(x)$ it's a premise and $C(x)$ a conclusion. The conditions of application of a rule are the premises and new knowledge are conclusions.

a) Inference engine: Bayesian networks

The inference engines used rules in various ways, in particular the method of reasoning will be based on probabilistic classification with Bayesian networks. This method is characterized by a multivariable representation of data to be processed, which allows to describe complex relationships of certain elements and nonlinear, they represent causal relationships, thus allowing handling uncertainty in events unobserved [7].

A Bayesian network represents a joint probability distribution on a 4-tuple $(G, f_x Q, \Theta)$, where:

- $(G, f_x Q)$ it's a causal network.
- G it's an acyclic digraph.
- The set x of nodes G its defined by $\{x_i | \, i \leq n\}$ of random variables with r possible states and Θ it's a set $\{\Theta_i | \, i \leq n\}$.

So, the expansion of the joint distribution is:

$$P(Y|f_1, \dots, f_N) = \prod_{i=1}^{N} p(f_i|Y) \cdot P(Y) \qquad (1)$$

For the posterior probability of how it can be the status of variables in a Bayesian network, Bayes' theorem is used.

b) Classifiers probabilistic as an artificial learning

Learning from the perspective of Artificial Intelligence, it is considered as a process of induction of knowledge that allows us to generalize behaviors from an unstructured information provided in the form of examples incorporating design workable solutions to problems through the study of the computational complexity of these.

The computational analysis and performance of machine learning algorithms is a branch of statistics known as Computational learning theory. Different learning algorithms are grouped into a taxonomy based on the output thereof. Some types of algorithms are:

Supervised learning. The algorithm produces a function that establishes a correspondence between inputs and desired outputs of the system. An example of this type of algorithm is the classification problem, where the learning system tries to assign a label or classify a series of vectors using one of several categories (classes). The knowledge base system consists of labeled examples above. This type of learning can be very useful in biological research problems, Computational Biology and Bioinformatics.

Unsupervised learning. The entire process of modeling is performed on a set of examples formed only by system inputs. There is no information on the categories of those examples. Therefore, in this case, the system must be able to recognize patterns to assign labels to new entries.

Semi supervised learning. This type of algorithms combine the above two algorithms to classify properly. Marked and unmarked data is taken into account.

Reinforcement learning. The algorithm learns observing the world around him. Your input information is the feedback you get from the outside world in response to their actions. Therefore, the system learns based on trial and error.

Transduction. Similar to supervised learning, but not explicitly constructed a function. Tries to predict future categories of examples based on input examples, their respective categories and examples new system.

Multitasking learning. Learning methods using knowledge previously learned by the system face to face similar problems to those already seen

c) Registers semiotic representation

Semiotic representations are representations that use signs, they can expressed in natural language or algebraic, graphs or figures geometric. These semiotic representations are the means through which a person can externalize their mental representations in order to make them visible or accessible to others. The ability to change the records of semiotic representation is necessary in the mathematics learning and the importance of coordinating the different registers of semiotic representation. Many difficulties experienced by students can be described and explained as a lack of coordination between representation registers, in particular graphic, numeric and algebraic [5].

II. METHODOLOGY

The expert system has several components, modules and persistent entities. Optimization problems are working and interacting with students through a virtual tutor who tells the procedure to be followed at all times. As shown in Figure 1, the student begins to access the pretest module, where it presents a diagnostic test that, through various reagents, assesses student skills and deficiencies [6].

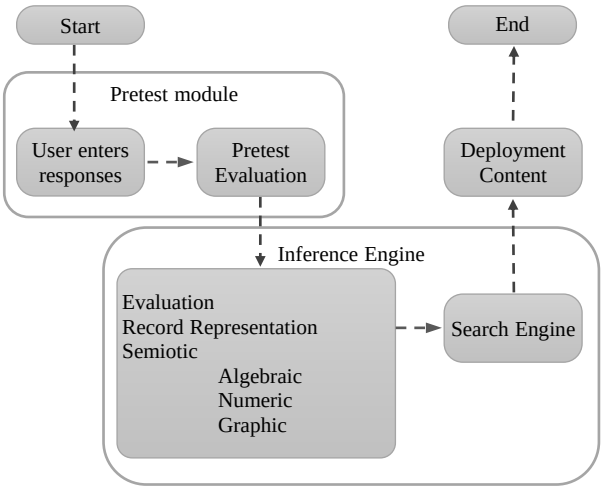


Fig. 1. Flow user action with the expert system

Figure 2 shows an example of the graphical user interface (GUI) provided to the student to work with the pretest. This interface provides the user with the following elements: the problem statement, a descriptive picture, different options to choose in response to problem and mechanisms of recurrent asynchronous storage

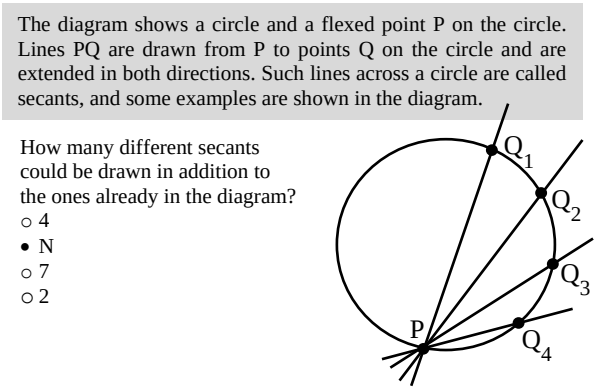


Fig. 2. GUI pretest for a record graphical representation

To determine which register of semiotic representation has the highest efficiency is assigned a score according to the complexity of the reagent. The evaluation is carried out in the inference engine using probabilistic classification based on Bayesian networks [7].

The search engine takes the result of the inference drawn and creates a data structure ramified from a collection of associative structures that link resources, activities and teaching materials. This collection is stored in a database. Finally, an exhaustive search is performed to determine the optimal route learning, that is to say, the set of suggested activities for the student.

A. System Overview

The expert system consists of a set of applications and Web services that engage in the following modules:

Users. In this module, the user has the ability to create your account, manage your profile, change personal data and set your preferences. This module is responsible for controlling the restricted access to the system.

Control of students. It allows to monitor student progress by monitoring system access and viewing activity logs and pretests presented. It provides the user with the role of teacher the ability to generate reports and statistics showing the student's behavior on the platform.

Administration. The user administrator role is the only one with access to this module. It allows management of secondary modules and specific components of search engine and inference.

Management pretests. Since this module creating, editing, and deleting query evaluation tests for the student is managed. The role of teacher has the privileges to access this module.

Content management. Enables management activities, resources and learning materials enabling the creation and editing content paths. These routes will be used by the search engine to determine the learning path that suits the profile of each student. The inference engine is responsible for composing the profile based on the records of semiotic representation.

Activity logs and backups. This module has tools that allow the export of records stored in the database through the scheduling of full or partial backups in plain text with SQL format.

III. RESULTS

We worked with a group of 30 college students, who started their Calculus course. This group was called the study group (SG). These students enrolled in the system and solved a questionnaire, which allowed them to determine their preference for the representation register, which they were used to work. Once they were placed in one of three profiles: Algebraic, numerical, and graphic, the system allocated resources to each student, according to their profile, in order to make the students feel comfortable resolving the problems. Once the activities were resolved, the students would upload the work developed as an attachment and in the system selected the response presented in subparagraphs. Subsequently, the system provides students with examples of solved problems, through simulations, using the three registers, graph, numerical and algebraic.

So, that the student can review how to obtain the same result using different semiotic register. About the part of digital educational resources there is also a league that re-addresses some program that allows the student to construct his or her own simulation and knowledge.

Finally, the system gives new problems to be solved by the student, offering him options of records to employ in its resolution. Figures 3 and 4 show the resolution of a rate of change problem using the algebraic register.

A village is situated 20 kilometers from a straight railroad track that passes through village A. Determine the position relative to B in which a train station C must be built on the railroad track so that the journey from village A to village B, making the trip by rail and CB by road, as long as possible, if the speed by rail is 80 Km/h and the road speed is 20 Km/h.

Algebraic procedure:

1. Find the time as a function of the distance x from the station to the point A.

$$v = \frac{e}{t} \quad \Rightarrow \quad t = \frac{e}{v}$$

$$t_1 = \frac{d}{20} = \frac{\sqrt{20^2 + (a-x)^2}}{20} \qquad t_2 = \frac{x}{80}$$

where: d is the distance elapsed at a time t_1 y x is the distances elapsed by the railway at a time t_2

2. Total time spent.

$$t = \frac{\sqrt{20^2 + (a-x)^2}}{20} + \frac{x}{80}$$

3. Function to optimize

$$t(x) = \frac{\sqrt{20^2 + (a-x)^2}}{20} + \frac{x}{80}, x \in [0, a]$$

4. Critical points

$$t'(x) = \frac{-(a-x)}{20\sqrt{20^2 + (a-x)^2}} + \frac{1}{80}$$

$$t'(x) = 0 \Rightarrow -80(a-x) + 20\sqrt{20^2 + (a-x)^2} = 0$$

$$4(a-x) = \sqrt{20^2 + (a-x)^2}$$

$$16(a-x)^2 = 20^2 + (a-x)^2$$

$$15(a-x)^2 = 20^2$$

$$a - x = \frac{20}{\sqrt{15}} \Rightarrow x_0 = a - \frac{20}{\sqrt{15}}$$

Fig. 3. Rate of change problem using the algebraic register

5. Discussion

- to $x = 0, t = \frac{\sqrt{20^2+a^2}}{20}$
- to $x = a, t_2 = 1 + \frac{a}{80}$
- to $x = a - \frac{20}{\sqrt{15}}, t_3 = \frac{\sqrt{20^2+\frac{20^2}{15}}}{20} + \frac{a-\frac{20^2}{15}}{8}$

Fig. 4. Rate of change problem using the algebraic register

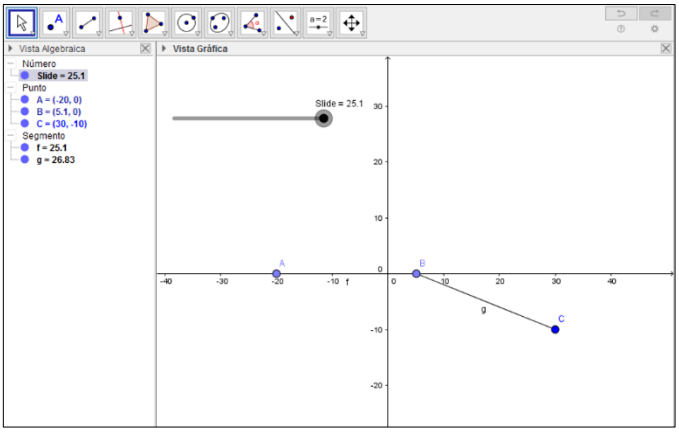


Fig. 5. Rate of change problem using the graphics register

The figures 5 and 6 show the construction of this problem using a dynamic geometry program.

Procedure in GeoGebra©:

1) With the straight-line tool take the option of segment and a length. Locate point A of coordinates (0, -20).

2) Determine the slides by giving values, minimum 0 and maximum 25.15.

3) Make another line segment from point B to point C (30, -20)

4) Built another line segment from point B to point D (30,0).

5) Use the slider to produce animation.

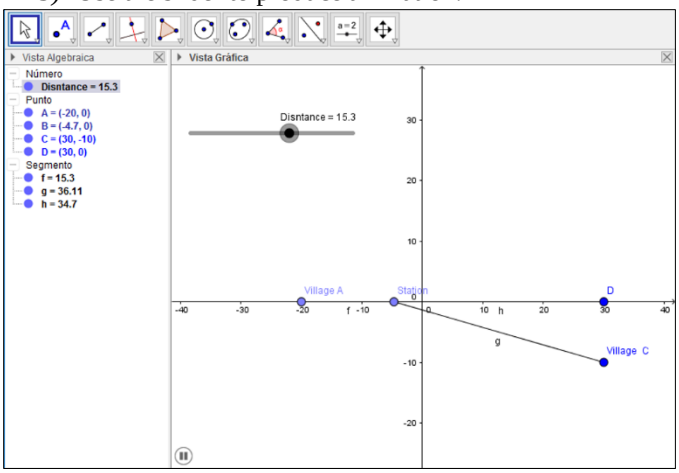
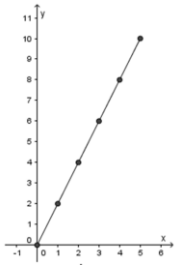


Fig. 6. Rate of change problem using the graphics register

The figure 7 and 8 show an example of numerical representation register that is related with the same concept rate of change.

A flow of water falls into a tank at a rate of constant change, such that for each unit that increases over time, the water depth increases by two units. The chart and graph illustrate this situation.

Hour (x)	0	1	2	3	4	5
Depth (y)	0	2	4	6	8	10
Difference		2	2	2	2	2



What is the rate of change of depth with respect to time if $x =$

Fig. 7. Rate of change problem using the numerical register

The study group (SG) used the system for one semester.

Once the students of the study group (SG) worked with the concepts of derivative and rate of change, by solving various activities and problems, using the expert system, they solved a final questionnaire, which asked them concepts applied to different situations of the same matter of Calculus. When

obtaining the average of the group in this questionnaire it was found that in relation to the initial had a significant increase since of having obtained 3.58, increased to 7.37. The F test was used and the increase was found to be ($p < 0.002$).

The table shows the amount of litter dumped into the sea at certain times of the year

Day (x)	1	2	3	4	5	6	7
Tons (y)	0.25	1.55	2.85	4.15	5.45	6.75	8.05
Difference		1.3	1.3	1.3	1.3	1.3	1.3

What is the rate of change of the amount of garbage with respect to the day if x = Thursday (day 5)?

Fig. 8. Rate of change problem using the numerical register

This result showed that SG identified the concept of derivative and correctly applied that concept to real problems solving correctly 74% of the final questionnaire, so it is concluded that the system expert open a new way in educational research.

IV. Conclusions

Using probabilistic classifiers for the implementation of inference engines in an expert system allows a multivariate representation of the data to be treated, which in turn shows a description of complex relationships of certain elements, also considering the handling of uncertainty. When using semiotic (graphic, numerical, iconic and algebraic) registers as primary classification elements, an appropriate form of presenting the aspects related to the learning paths of the area to be reinforced in the student is parameterized.

Acknowledgment

We are grateful for the support provided to IPN's SIP (research project 20164801. We also thank COFAA.

References

[1] Kurbel K., Hilker, J., (2002). Requirements for a mobile e-Learning Platform, IASTED Intl. Conf. on Commun., Internet and Information Technology, US Virgin Islands.

[2] Felder, R and Silverman, L. (1988) “Learning and Teaching Styles,” Journal of Engineering Education, Vol. 78, No.7, pp. 674-681.

[3] Micarelli, A., Stamper J., and. Panourgia, K. (2016). Intelligent Tutoring Systems. Sun, T., Liu H. (2013).

[4] Henderson, H. (2003). Encyclopedia of computer science and technology. New York, NY: Facts on File.

[5] Duval, R. (1998). Registros de representación semiótica y funcionamiento cognitivo del pensamiento. Investigaciones en Matemática Educativa II, p. 173–201.

[6] Sun, T., Liu H. (2013). Design of Fault Diagnosis Expert System of Transformer. AMM, vol. 291-294, pp. 2557-2561, 2013.

[7] Woolf, B. (2009). Building intelligent interactive tutors. Amsterdam: Morgan Kaufmann .

[8] Russell, S., Norving, P. (2009). Artificial Intelligence: A Modern Approach, 3rd. Prentice Hall, Englewood Cliffs.

RIN-Sum: A System for Query-Specific Multi-Document Extractive Summarization

Rajesh Wadhvani
CSE
National Institute of Technology
Bhopal, India

Manasi Gyanchandani
CSE
National Institute of Technology
Bhopal, India

Rajesh Kumar Pateriya
CSE
National Institute of Technology
Bhopal, India

Sanyam Shukla
CSE
National Institute of Technology
Bhopal, India

***Abstract*—In paper, we have proposed a novel summarization framework to generate a quality summary by extracting Relevant-Informative-Novel (RIN) sentences from topically related document collection called as RIN-Sum. In the proposed framework, with the aim to retrieve user's relevant informative sentences conveying novel information, ranking of structured sentences has been carried out. For sentence ranking, Relevant-Informative-Novelty (RIN) ranking function is formulated in which three factors, i.e., the relevance of sentence with input query, informativeness of the sentence and the novelty of the sentence have been considered. For relevance measure instead of incorporating existing metrics, i.e., Cosine and Overlap which have certain limitations, a new relevant metric called as C-Overlap has been formulated. RIN ranking is applied on document collection to retrieve relevant sentences conveying significant and novel information about the query. These retrieved sentences are used to generate query-specific summary of multiple documents. The performance of proposed framework have been investigated using standard dataset, i.e., DUC2007 documents collection and summary evaluation tool, i.e., ROUGE.**

Keywords*—Text summarization; maximum marginal relevance; sentence selection; DUC2007 data collection

I. Introduction

The notion of information retrieval is to locate documents that might contain the relevant information. Generally, when a user fires a query, his desire is to locate relevant information rather than locate a ranked list of documents. The retrieved documents contain the relevant information leaving the user with a massive amount of text. There is a requirement of a tool that shrinks this amount of text in order to comprehend the complete text [1]. The query focused summarization track at Document Understanding Conference (DUC) aims at doing exactly this. Conventional query focused text summarization systems rank and assimilate sentences based on maximizing relevance to the user's information need expressed via query [2]. These systems do not consider the important factor, i.e., informativeness and novelty of the sentence. In this paper, a novel summarization framework to generate a quality summary by extracting Relevant-Informative-Novel (RIN) sentences from topically related document collection called as RIN-Sum has been presented. This framework generates a query focused summary of multiple documents by using three factors, namely: sentence relevance with input query (discussed in section 2), sentence informativeness (discussed in section 3) and sentence novelty (discussed in section 4). In this work, ordering of these factors has been considered to rank the sentences. Firstly, relevance with input query is applied, and then sentence informativeness, and finally sentence novelty. For example, if a sentence is novel and highly informative in the document collection, but if it is not relevant to a user's query, it will not be considered for a final summary.

II. The Relevance Measure

Relevance measures can be divided into two types based on whether the ordering of vectors is taken into account, i.e., symmetric and asymmetric [3] [4]. For two sentence vectors S_i and S_j, a symmetric measure yields the same result regardless of the ordering of the sentence vectors, i.e., $Sim(S_i,S_j) = Sim(S_j,S_i)$. An asymmetric measure yields different results for different orderings of two sentence vectors, i.e., $Sim(S_i,S_j) \neq Sim(S_j,S_i)$. The Cosine measure is the most popular symmetric measure based on VSM for checking the extent of similarity between two texts. In VSM for text summarization, the sentence is usually presented as a vector of weighted terms. Cosine similarity between two weighted sentences $S_i = [w_{1i},\ldots..,w_{ni}]$ and $S_j = [w_{1j},\ldots\ldots,w_{nj}]$ can be define as:

$$Sim_{cos}(s_i, s_j) = \frac{vec(s_i).vec(s_j)}{|\,s_i\,| \;\; |\,s_j\,|} = \frac{\sum_{k=1}^{n} w_{ki} \times w_{kj}}{\sqrt{\sum_{k=1}^{n} w_{ki}^2} \;\times\; \sqrt{\sum_{k=1}^{n} w_{kj}^2}} \quad (1)$$

In Cosine measure, two sentence vectors S_i and S_j are compared on the basis of all terms which appear in S_i and/or S_j. In both sentences discriminative power of each term is well defined. Discriminative power of uncommon terms between S_i and S_j also affects the similarity measure. Hence this type of similarity measure performs well when two texts are

compared on the basis of a set of terms appearing in either first text and/or second text.

The Overlap measure is the asymmetric relevance measure between two texts. It is a relative measure to detect similarity or overlap among texts by making comparison between the current text and any other text with respect to all those terms which appears only in current text. The Overlap measure is computed by comparing the current sentence S_i with any sentence S_j, as define in [5] is given in (2):

$$Sim_{overlaps}(s_i, s_j) = \frac{\sum_{k=1}^{m} w_{ki} \times w_{kj}}{\sum_{k=1}^{m} w_{ki}^2} \quad (2)$$

This metric is a relative measure to detect similarity or overlap among sentences. This mechanism works on the comparison of the relative frequency of the words representing a sentence. One of the limitations with (2) is that it does not compare two sentences irrespective of their sizes. This causes problem in a situation when for a given common term weight in S_j dominates over weight in S_i, resulting in increase of the overlap score in the proportion to differences in their weights. In case of Cosine measure there is no such limitation. For getting the advantages of overlap measure, there is a need to improve (2) over this limitation. Proposed improvement over this metric has been formulated in (3).

$$Sim_{overlaps}(s_i, s_j) = \frac{\sum_{k=1}^{m} [min(w_{ki}, w_{kj})]^2}{\sum_{k=1}^{m} w_{ki}^2} \quad (3)$$

The above mentioned metrics for Overlap measure is used to determine whether sentences are copies of one another or not. One limitation with this metric is that it does not consider the discriminative power of the terms. In next section Overlap based Cosine measure is formulated for identifying all those sentences in which each term of current text appears with high discriminative power.

At the time of sentence extraction, applying Overlap measure technique as relevance measure returns a set of sentences without considering the discriminative power of query terms of those sentences. Here the use of Cosine measure may improve the match quality by considering the discriminative power of query terms in sentence ranking, but at the same time ranking of the sentence are declined in its non-query terms. Matching quality can be improved by adding the properties of Overlap in Cosine measures. In this respect, proposed methodology has been formulated called as Overlapped based Cosine measure which can be abbreviated as C-Overlap measure. In this formulation, at its first step, the terms appear in sentence S_j are decomposed into two groups having common and uncommon terms with respect to S_i. After decomposing sentence S_j into two groups, $S_j' = [w_{1j}, \ldots, w_{mj}]$ and $S_j'' = [w_{(m+1)j}, \ldots, w_{nj}]$ are obtained, where $S_j = S_j' \cup S_j''$. Overlap between S_i and S_j is nothing but cosine similarity between S_i and S_j'. Cosine similarity between two weighted sentences $S_i = [w_{1i}, \ldots, w_{mi}]$ and $S_j' = [w_{1j}, \ldots, w_{mj}]$, can be formulated as follows:

$$Sim_{Coverlap}(s_i, s_j) = Sim_{cos}(s_i, s_j') = \frac{\sum_{k=1}^{m} w_{ki} \times w_{kj}}{\sqrt{\sum_{k=1}^{m} w_{ki}^2} \sqrt{\sum_{k=1}^{m} w_{kj}^2}} \quad (4)$$

$$Sim_{Coverlap}(s_i, s_j) = \frac{vec(s_i).vec(s_j)}{\|s_i\| \sqrt{\sum_{k=1}^{m} w_{kj}^2}} \quad (5)$$

Here in normalization process of vector S_j, uncommon terms are neglected. As a result strength of common terms increases. Hence sentences will be ranked on the basis of discriminative query terms only.

III. The Informativeness Measure

Cosine, Overlap and C-Overlap all are pure relevance measurement techniques which do not consider the sentence informativeness. A ranking metric is required which improves the rank of relevant sentences on the basis of informativeness of the sentence. In this section, a ranking function which measures the informativeness score of the given sentence based on assumed hypothesis, i.e., "within a query relevant sentence, its non-query terms may convey information about the query terms" is formulated, which is defined as follows:

$$informative_{t \in S_i \setminus Q}(S_i) = \sqrt{\sum_{t \in S_i \setminus Q} w_{ki}^2} \quad (6)$$

Here informativeness of sentences S_i is measured by considering the weights of non-query terms only. A score of informativeness of the sentence is equal to L2 norm or Euclidean norm of the weights of discriminative non query terms. In this work, instead of preferring large number of low discriminative terms, small numbers of high discriminative terms are considered. Therefore, L2 norm is preferred over L1 norm as the L1 norm focuses on total weights while L2 norm considers the distribution of weights. Further, the value of the score may be greater than one and to use it with other scores it need to be normalized in the range of 0 to 1 for all sentences in the document collection. To normalize this score, initially score of informativeness for all sentences in document collection is calculated and then maximum score between them is found as:

$$Maxscore = \max[informative_{t \in S_i \setminus Q}(S_1), informative_{t \in S_i \setminus Q} \quad (7)$$

Now to obtain the normalized score, score of each sentence is divided with Maxscore and can be written as:

$$informative_{t \in S_i \setminus Q}(S_i) = \frac{\sqrt{\sum_{t \in S_i \setminus Q} w_{ki}^2}}{Maxscore} \quad (8)$$

Besides this, a ranking function for informativeness is used to formulate sentence informativeness based relevant metric.

This approach measures relevance and informativeness of the sentence separately and then uses a linear combination of the two to produce a single score for the ranking of a sentence. The informativeness based relevant metric can be formulated as:

$$Sim_{Relevance-informative}(Q, S_i) = \beta\, Sim_{relevance}(Q, S_i) + (1-\beta)\, informative_{t \in S_i \setminus Q}(S_i) \quad (9)$$

In this metric any one of Cosine, Overlap and C-Overlap can be used for relevance measurement. β is tuning factor and its theoretical value lies between 0 to 1. A sentence of our interest is primarily relevant to user query and then informative. To accomplish this in (9) relevant metric should get more weight as compare to informative metric. So practically, value of β should be close to one.

IV. The Novelty Measure

In automatic text summarization, precision of results will be increased by being very selective about the sentences and retaining only those in summary that are considered to be surely relevant. Therefore, necessary condition to retain a sentence in the summary is its relevance with input query. Along with precision a good coverage is required for improving recall, but at the same time another constrains with summary is that it is bounded in length [6]. Optimizing these three constrains, namely: relevance, coverage, and summary length is a challenging task. One of the solutions to maximize the coverage of summary by confirming its length is trying to include those relevant sentences which are novel to the sentences already retained in summary.

A. Maximum Marginal Relevance (MMR)

Carbonell et al. [7] encouraged Maximal Marginal Relevance (MMR) which considers novelty along with relevance to rank the text. Using this technique, partial or full duplicate information is prevented from being retrieved. In particular, MMR has been widely used in text summarization because of its simplicity and effectiveness, and it has shown a consistently good performance. MMR uses the Retrieval Status Value (RSV) as a parameter to measure the diversity among the sentences. The RSV value of the newly retrieved sentence is decided by sentences which have been already retrieved. It prevents the similar sentences by lowering their RSV value and as a result, it boosts up dissimilar sentences. The final score of given sentence S_i is calculated as follows:

$$MMR(R,S) = \underset{S_i \in R \setminus S}{argmax} \left[\lambda\{Sim_1(Q, S_i)\} - (1-\lambda)\left\{\max_{S_j \in S} Sim_2(S_i, S_j)\right\}\right] \quad (10)$$

Where R stands for the ranked list of sentences, S represents the sentences that have been extracted into the summary, Q denotes the query and S_i indicates a sentence. Sim_1 and Sim_2 are similarity measures, which can either be same or different. Different similarity measures have been explored in next session. λ is tuning factor which lies between 0 to 1.

In this approach, summaries are created using greedy sentence-by-sentence selection. At each selection step, the greedy algorithm is constrained to select the sentence that is maximally relevant to the user query and minimally redundant with sentences which have been already included in the summary. MMR measures relevance and novelty separately and then uses a linear combination of the two to produce a single score for the importance of a sentence in a given stage of the selection process. Xie et al. [8], Forst et al. [9] and Chowdary et al. [10] encouraged the concept of "relevant novelty", which claim that a sentence of input text will be retained in a summary if it is relevant to the user and should not convey the information which is already covered by the current summary sentences.

B. Relevant-Informative-Novelty (RIN) metric for sentence selection

Relevance, informativeness and novelty are the three basic measures which have been considered in the ranking during sentence extraction. Considering only relevance measure for generating the summary does not give the guarantee of novelty in the summary. In this section, a ranking metric is formulated which improves the rank of relevant and informative sentences based on their diversity with other sentences. In this formulation, MMR has been used. A ranking function which measures a novelty score of the given sentence with respect to current summary sentences is formulated. This formulation is based on the following assumptions:

- Those sentences in the current summary are put under considerations which are diverse on the basis of conveyed information.
- Sentences are retained in the current summary if they convey novel information about the query.
- Within a query relevant sentence, its non-query terms may convey information about the query terms.

Thus, novelty of given sentence with respect to current summary sentence can be measured in term of amount of overlap between non query term of given sentence S_i and current summary sentence S_j. This can be calculated as follows:

$$Sim_{novelty}(S_i, S_j) = Sim_{overlaps}(S_i, S_j) : \text{ if } k \in S_i \cap Q \text{ then } w'_{ki} = 0 \text{ else } w'_{ki} = w_{ki}, \text{ where } w'_{ki} \text{ is } k^{th} \text{ term in } S_i' \quad (11)$$

Now using linear combination of relevant and novelty metric final score is obtained. Relevant-novelty metric can be given as:

$$Score(R,S) = \underset{S_i \in R \setminus S}{argmax} \left[\lambda\{Sim_{Relevant}(Q, S_i)\} - (1-\lambda)\left\{\max_{S_j \in S} Sim_{novelty}(S_i, S_j)\right\}\right] \quad (12)$$

In case, when informativeness of the sentence is considered, RIN metric can be given as:

$$Score(R,S) = \underset{S_i \in R \setminus S}{argmax} \left[\lambda\{Sim_{Relavant-informative}(Q,S_i)\} - (1-\lambda)\left\{\max_{S_j \in S} Sim_{novelty}(S_i,S_j)\right\}\right] \quad (13)$$

In this metric, novelty of sentence S_i is measured in terms of amount of overlap between non query term of given sentence S_i and current summary sentence S_j. λ is tuning factor and its theoretical value lies between 0 to 1. More weight is given to informativeness based relevant metric because a sentence is significant if primarily relevant to the user query then it should be informative and finally it should be a novel.

V. RIN-SUM METHODOLOGY

To provide the methodology of sentence extraction from unstructured text to generate its query-specific summary, following are the steps that RIN-Sum takes to construct query-specific summary of multiple documents.

1) Select a query and set of associated documents for which summary is to be generated. These documents and the query constitute the input to RIN-Sum.

2) Each document in the collection is analyzed to obtained its structured representation using following steps:

- Firstly, each document is pre-processed to generate sentence set.
- Each sentence in the resultant set is represented by vector in dimensions of content terms of pre-processed document.
- Each sentence vector is weighted for content terms.

3) Finally a cluster of unstructured sentences is generated as a final summary by extracting salient and non-redundant sentences from given document collection. This process consists of following steps:

- Firstly, Sentence vectors are ranked by applying proposed C-Overlap measure based relevant metric to produces a cluster of relevant sentences.
- Resultant cluster sentence vectors are again ranked through proposed Relevant-Informative metric to produces a cluster of relevant and informative sentences.
- Finally, to retrieve sentences conveying novel information about query from group of identified relevant-informative sentences, a Relevant-Informative-Novelty (RIN) ranking function is used.

4) Further the performance of the proposed framework has been investigated using standard dataset, i.e., DUC2007 documents collection and summary evaluation tool, i.e., ROUGE, and simulation strategy of proposed methodology and analysis of results have been performed.

Thus RIN-Sum uses topically related documents to produce a summary. These summaries are deemed relevant to a user query. For example, to satisfy the user's information need about given topically related documents collection, a summary which contains user's intended information on that topic will be generated.

VI. EXPERIMENTS

DUC2007 dataset has been used for evaluation and it is available through [11] on request. A total of 45 documents were constructed by NIST assessors based on topics of interest and for each topic four reference summaries were produced by human experts to create gold collection for evaluation purposes. For performance evaluation ROUGE-1, ROUGE-2 and ROUGE-SU metrics of ROUGE-1.5.5 package [12] has been used. ROUGE-1 compares the unigram overlap between the candidate summary and the reference summaries. ROUGE-2 compares the bigram overlap between the candidate summary and the reference summaries. ROUGE-SU is an extended version of ROUGE-2 that match skip bigrams, with skip distance up to 4 words. Performance is measured in terms of Recall, Precision and F-score. Several experiments have been conducted in which for text representation the standard sequence of steps have been followed, which are:

1) Generate sentence set by separating sentences of DUC2007 document collection.

2) Remove functional and grammatical words of the sentences using stop word list, provided with DUC document collection.

3) For each sentence, apply stemming algorithm on each word of the sentence with the help of well-known Porter Stemmer [13] in order to find related words.

4) Calculate weight of each word within the sentence using standard tf.idf weighting scheme [14].

As an output of the above steps, sentences of each document are represented as sentence-terms weighted vector. Now using sentence ranking function as formulated in (13), sentences are ranked and then extracted to get a final summary. With ranking function, different experiments are performed for different relevance measure i.e. Cosine, overlap, C-Overlap. For informativeness and novelty measure, fixed measure as defined in (8) and (11) respectively are used. Experiments were performed in three different phases. In each phase four different ranking functions were used which are:

- Relevant Metric
- Relevant-Informative Metric
- Relevant-Novelty Metric
- Relevant-Informative-Novelty(RIN) Metric

Here results were obtained for different ROUGE metrics in term of Precision, Recall and F-Score.

Phase I: In this phase results were obtained for above four ranking functions. In these experiments Cosine measure was used as relevant metric and for informativeness and novelty

measure fixed metrics was used as defined in (8) and (11) respectively. The results are as shown in Table (1).

TABLE. I. EVALUATION RESULTS USING COSINE MEASURE BASED (A) RELEVANT RANKING FUNCTION; (B) RELEVANT-INFORMATIVE RANKING FUNCTION; (C) RELEVANT NOVELTY RANKING FUNCTION AND (D) RELEVANT-INFORMATIVE-NOVELTY RANKING FUNCTION

a

	Recall	Precision	F-score
ROUGE-1	0.42037	0.38742	0.40248
ROUGE-2	0.10729	0.10046	0.10369
ROUGE-SU	0.16919	0.14283	0.15391

b

	Recall	Precision	F-score
ROUGE-1	0.42203	0.39064	0.40494
ROUGE-2	0.10789	0.10155	0.10455
ROUGE-SU	0.16997	0.14686	0.15682

c

	Recall	Precision	F-score
ROUGE-1	0.42638	0.39512	0.40938
ROUGE-2	0.10775	0.10073	0.10397
ROUGE-SU	0.17403	0.14873	0.15935

d

	Recall	Precision	F-score
ROUGE-1	0.43557	0.40243	0.41786
ROUGE-2	0.11719	0.10980	0.11329
ROUGE-SU	0.18374	0.15546	0.16745

Phase II: In this phase results were obtained for above four ranking functions. In these experiments overlap measure was used as relevant metric and for informativeness and novelty measure fixed metrics was used as defined in (8) and (11) respectively. The results are as shown in Table (2).

TABLE. II. EVALUATION RESULTS USING OVERLAP MEASURE BASED (A) RELEVANT RANKING FUNCTION; (B) RELEVANT-INFORMATIVE RANKING FUNCTION; (C) RELEVANT NOVELTY RANKING FUNCTION AND (D) RELEVANT-INFORMATIVE-NOVELTY RANKING FUNCTION

a

	Recall	Precision	F-score
ROUGE-1	0.43321	0.40360	0.41715
ROUGE-2	0.11283	0.10536	0.10876
ROUGE-SU	0.17934	0.15429	0.16467

b

	Recall	Precision	F-score
ROUGE-1	0.44941	0.40532	0.42548
ROUGE-2	0.12231	0.11124	0.11635
ROUGE-SU	0.18958	0.15488	0.16949

c

	Recall	Precision	F-score
ROUGE-1	0.44848	0.41063	0.42826
ROUGE-2	0.12129	0.11273	0.11677
ROUGE-SU	0.19158	0.15975	0.17334

d

	Recall	Precision	F-score
ROUGE-1	0.45678	0.41583	0.43449
ROUGE-2	0.12971	0.12109	0.12511
ROUGE-SU	0.19761	0.16549	0.17886

***Phase III*:** In this phase results were obtained for above four ranking functions. In these experiments C-Overlap measure was used as relevant metric and for informativeness and novelty measure fixed metrics was used as defined in (8) and (11) respectively. The results are as shown in Table (3).

TABLE. III. EVALUATION RESULTS USING C-OVERLAP MEASURE BASED (A) RELEVANT RANKING FUNCTION; (B) RELEVANT-INFORMATIVE RANKING FUNCTION; (C) RELEVANT NOVELTY RANKING FUNCTION AND (D) RELEVANT-INFORMATIVE-NOVELTY RANKING FUNCTION

a

	Recall	Precision	F-score
ROUGE-1	0.44777	0.40909	0.42666
ROUGE-2	0.12581	0.11597	0.12048
ROUGE-SU	0.19081	0.16002	0.17262

b

	Recall	Precision	F-score
ROUGE-1	0.45646	0.41197	0.43217
ROUGE-2	0.13073	0.12003	0.12493
ROUGE-SU	0.19883	0.16304	0.17800

c

	Recall	Precision	F-score
ROUGE-1	0.45935	0.41621	0.43597
ROUGE-2	0.13176	0.12197	0.12643
ROUGE-SU	0.20092	0.16575	0.18005

d

	Recall	Precision	F-score
ROUGE-1	0.46487	0.41969	0.44042
ROUGE-2	0.13568	0.12304	0.12879
ROUGE-SU	0.20821	0.16813	0.18449

Graphically the F-scores (ROUGE-1, ROUGE-2 and ROUGE-SU) results are depicted in figures (1) - (3) respectively.

In these figures, while observing the curves of Relevant and Relevant-Informative Ranking, it can be concluded that in all cases, i.e., Cosine, Overlap and C-Overlap the performance of Relevant-Informative Ranking is better as compared to Relevant Ranking. While observing the curves of Relevant and Relevant-informative-Novelty Ranking it can be concluded that in all cases, i.e., Cosine, Overlap and C-Overlap the performance of Relevant-Informative-Novelty Ranking is better as compared to Relevant Ranking.

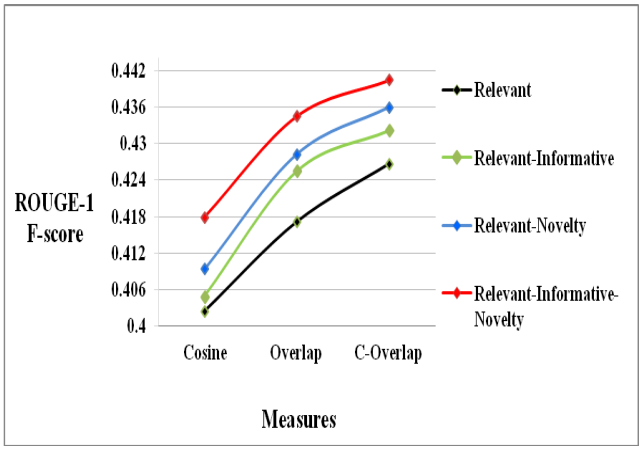


Fig. 1. ROUGE-1 F-score results comparison for Relevant, Relevant-Informative, Relevant-Novelty and Relevant-Informative-Novelty metrics over Cosine, Overlap and C-Overlap measures

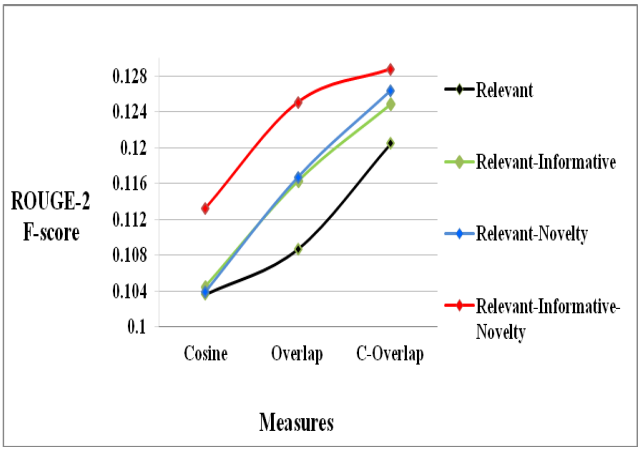


Fig. 2. ROUGE-2 F-score results comparison for Relevant, Relevant-Informative, Relevant-Novelty and Relevant-Informative-Novelty metrics over Cosine, Overlap and C-Overlap measures

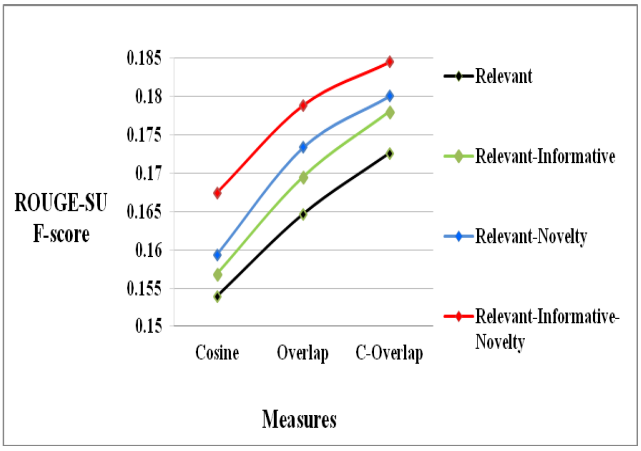


Fig. 3. ROUGE-SU F-score results comparison for Relevant, Relevant-Informative, Relevant-Novelty and Relevant-Informative-Novelty metrics over Cosine, Overlap and C-Overlap measures

Justification for this improvement is that ranking of the sentence based on proposed C-Overlap relevance measure does not consider the significance of non-query terms. When Informative Metric is applied in sentence ranking, it considers the significance of non-query terms also. As a result, this technique tries to retrieve all sentences having significant query terms as well as significant non-query terms. Also, in sentence ranking when Novelty Metric is applied, it prevents the retrieval of partial or full duplicate information and improves the coverage of bounded length summary. As a result, performance in terms of recall value increases.

VII. CONCLUSION

In this paper, a novel technique to query specific extractive text summarization for multiple documents has been presented. The utility of the approach is examined on DUC2007 dataset collection. In the proposed method, with the aim to retrieve user's relevant significant sentences conveying novel information, ranking of structured sentences has been carried out. A new method of sentence ranking has been developed which identifies the relevant, significant and novel sentences from a large volume of input text. To achieve this, RIN metric is formulated for sentence ranking depending on three factors, i.e., the relevance of sentences with input query, informativeness of the sentence as well as the novelty of the sentence. For relevance measurement, a new measure formally known as C-Overlap (Overlapped based Cosine measure) has been proposed with the aim to overcome the limitations of existing relevance measures, i.e., Cosine and Overlap measure. Experimentally it has been proved that C-Overlap measure outperformed the previous ones.

Finally, sentences in document collection were extracted using RIN ranking metric. Results were compared with the other standard sentence ranking functions, i.e., Relevant, Relevant-Informative, Relevant-Novelty and Relevant-Informative-Novelty, using ROUGE-1.5.5. It has been observed that in each case results of proposed function are found to be better as compared to other three ranking functions. Experimentally, it is also observed that Relevance alone is not a good choice as a ranking function.

REFERENCES

[1] Mani I and Maybury M T 1999 Advances in Automatic Text Summarization. MIT Press, Cambridge

[2] Kumar Y J and Salim N 2011 Automatic Multi Document Summarization Approaches. In: Journal of Computer Science, Volume 8, Issue 1, pp: 133–140

[3] Tsai S F S, Tang W and Chan K L 2010 Evaluation of novelty metrics for sentence-level novelty mining. In: Information Sciences, Volume 180, Number 12,pp: 2359–2374

[4] Zhang Y, Tsai F S and Kwee A T 2011 Multilingual sentence categorization and novelty mining. In: Information Processing and Management, Volume 47, pp: 667–675

[5] Alguliev R M, Aliguliyev R M and Isazade N R 2013 MR&MR-SUM: Maximum Relevance and Minimum Redundancy Document Summarization Model. In: World Scientific Publishing Company, International Journal of Information Technology and Decision Making, Volume 12, Number 3, pp: 361–394

[6] Alguliev R M, Aliguliyev R M, Hajirahimova M S and Mehdiyev C A 2011 MCMR: Maximum coverage and minimum redundant text summarization model. In: Elsevier, Expert Systems with Applications, Volume 38, pp: 14514-14522.

[7] Carbonell J and Goldstein J 1998 The use of MMR, diversity-based Reranking for reordering documents and producing summaries. In Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp: 335–336

[8] Xie S and Liu Y 2008 Using corpus and knowledge-based similarity measure in Maximum Marginal Relevance for meeting summarization. In IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp: 4985-4988

[9] Forst J F, Tombros A and Roelleke T 2009 Less is More: Maximal Marginal Relevance as a Summarization Feature. In Proceedings of the 2nd International Conference on the Theory of Information Retrieval, Lecture Notes in Computer Science, Volume 5766, pp: 350–353

[10] Guohua WU and Yutian G 2016 Using Density Peaks Sentence Clustering For Update Summary Generation. In Proceedings of 2016 IEEE Canadian Conference on Electrical and Computer Engineering.

[11] Document Understanding Conference: <http://duc.nist.gov>.

[12] Lin C-Y 2004 ROUGE: A package for automatic evaluation summaries. In Proceedings of the ACL Text Summarization Branches Out Workshop, Barcelona, Spain, pp: 74–81

[13] Porter M 2006 The Porter Stemming Algorithm, Official home page for distribution of the Porter Stemming Algorithm. <http://tartarus.org/~martin/PorterStemmer/ index.html>

[14] Polettini N 2004 The Vector Space Model in Information Retrieval Term Weighting Problem. http://sra.itc.it/people/polettini/PAPERS/ Polettini Information Retrieval.pdf

A Bus Arbitration Scheme with an Efficient Utilization and Distribution

Amin M. A. El-Kustaban
Department of Electronic Engineering
University of Science and Technology (UST)
Sana'a, Yemen

Abdullah A. K. Qahtan
Department of Electronic Engineering
University of Science and Technology (UST)
Sana'a, Yemen

***Abstract*—Computer designers utilize the recent huge advances in Very Large Scale Integration (VLSI) to place several processors on the same chip die to get Chip Multiprocessor (CMP). The shared bus is the most common media used to connect these processors with each other and with the shared resources. Distributing the shared bus among the contention processors represents a critical issue that affects overall performance of the CMP. Optimal utilization with fair distribution of the shared bus represents another challenge. This paper introduces a bus arbitration scheme, which is an Age-Based Lottery (ABL) Arbitration that combines the lottery and age-based algorithms to overcome the shared bus challenges. The results show that the developed bus arbitration scheme maximizes the bus utilization and improves the distribution by at least 13.5% with an acceptable latency time comparing to the traditional bus arbitration schemes.**

Keywords—Chip Multiprocessor; Round Robin; Lottery Algorithms; Latency; VHDL

I. INTRODUCTION

The technology revolution in Very Large Scale Integration (VLSI) has enabled today's designers to design and implement Chip Multiprocessor (CMP), where two or more processors with a shared memory are integrated on a single chip [1].

The contention between the processors in CMP systems adds significant overhead in order to manage the access to that shared bus [2]. Thus, scheduling mechanisms or "arbitration schemes", which are employed to synchronize and schedule the bus requesting from different bus masters in order to avoid contentions, have a major and important effect on the overall performance of the CMP design [3-5]. One of the challenges faced by the bus arbitration is to ensure that the sharing resources can be utilized and balanced distribution among the contention masters.

The improvements on the bus arbitration protocols are performed to enhance some of the protocols' aspects, such as: the fairness degree, latency time, bandwidth utilization, responding to priorities, cost, and power consumption [6].

In this paper, a bus arbitration scheme, which is called an Age-Based Lottery (ABL), is introduced. This scheme overcomes the static and dynamic lottery schemes shortcomings such as the unbalance distribution of the bus. Also, this paper improves the performance by maximizing the shared bus utilization and balancing the bus distribution with an acceptable latency. The results are shown and compared to the traditional bus arbitration schemes by implementing them using the Hardware Description Language (HDL) and illustrating the testing results using ModelSim tool.

This paper is organized as follows. Section II, reviews the related work. Section III, introduces the most knowing bus arbitration schemes. Section IV, discusses the developed bus arbitration scheme. Implementing, testing and comparing the developed bus arbitration scheme to the traditional schemes are presented in section V. Finally, conclusion and future work are summarized in section VI.

II. RELATED WORK

The related work of the bus arbitration can be divided into three categories. First, implementing the existing bus arbitration protocol. Second, enhancing existing protocols in order to improve the whole bus–base system performance. Third. introducing new bus arbitration schemes. This section clarifies some of these works as follow:

Two new bus arbitration algorithms, which are Request-Service and Age-Based algorithms are introduced in [2]. The new algorithms try to improve the existing algorithms in term of latency caused by the contention among the bus masters. The Request-Service algorithm attempts to remove all forms of starvation among the competing maters. It also sets an upper limit for the waiting time for each master. The Age-Based algorithm gives more priority to masters that have recently used the bus, which will lead to improve the performance. The starvation problem is solved in this algorithm by using CritNo flag. Each algorithm has been implemented in a software simulation. The results show that the Request-Service model works well under low load. The Age-Based model performs well as the Futurebus model and reduces the amount of starvation and it is suitable when there is a need to transfer large blocks of data.

A HDL implementation and analysis of the lottery bus arbitration techniques are presented in [3]. The problem of generating a pseudo-random number greater than the total tickets value, which cause that none of the masters will get access the bus, is solved by allowing the bus to be granted to the master that is given highest priority. Moreover, the priority is rotated among the masters in order to prevent a single master to grant the bus for long time when the random number falls outside the range of the total tickets value. The results of the implementation indicate that dynamic lottery is more efficient than static lottery since it improves the average waiting time of

the bus masters. In addition, dynamic lottery using rotating priority ensures the best average waiting time for the bus masters comparing with other lottery approaches. However, more resources and on-chip power consumption are the most disadvantages of dynamic lottery comparing to static lottery.

A novel Dynamically Adaptive Arbitration (DAA) algorithm and compares it with the traditional bus arbitration protocols through using MPEG-4 video encoder application on FPGA instead of the analytical simulation methods are presented in [7]. The new DAA algorithm has been inspired by Lottery bus, where a dynamic algorithm has been implemented for centralized arbiter. The algorithm adaptively allocates the bus bandwidth to the masters that need it based on the usage history. The bus is offered more to those masters that have been the most active lately. The comparing results show that DAA competes with RR in performance sense in every evaluated case. DAA delivers the best performance when a high clock frequency is used. However, DAA drawback is the highest area requirement. If the area is an important issue, RR is a safe choice that performs well in most cases.

A dynamic round robin arbiter based on lottery method using VHDL is implemented in [8]. The results of the implemented model, which are shown on ModelSim tool, show that the latency is improved with the dynamic tickets more than the static tickets and the starvation is avoided. Moreover, the latency of the highest priority master is lower than that of some conventional architecture. The proposed arbiter provides flexible design for efficient SoC. However, the limitation of the dynamic method is that the distribution of random number is not uniform [9].

There are many other researches use FPGA and VHDL to implement and test their proposed or existing bus arbitration algorithms such as [10-13].

III. Bus Arbitration Schemes

The bus arbitration schemes can be divided into two broad categories, which are centralized and decentralized or distributed arbitration. In the centralized arbitration, there is a single arbiter for the bus. Each master sends its request to that arbiter, and then the arbiter decides which the bus owner according to the applied protocol is. In the decentralized arbitration, there is no explicit device or unit to decide which master will own the bus. However, all of the devices on the bus work together to determine which device will get the bus access [14]. The most knowing centralized bus arbitration protocols are daisy chain, static fixed priority, round robin, time division multiplexed, and lottery bus arbitration. In the following sub-sections, the round robin and lottery bus arbitration, which are related to the work in this paper, will be discussed.

A. Round Robin

A round robin (RR) protocol is a simple and fair arbitration style where no master is allowed to get the bus ownership indefinitely [15]. Any master wants to access the bus will get it in an arranged manner as shown in Fig. 1. Whenever a master's turn ends, either unused, because of the end of the data transfer, or limited time length, the turn is passed to the next master.

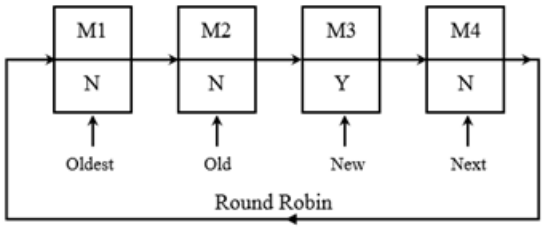


Fig. 1. Round Robin Arbitration

The RR has a disadvantage of checking all masters' interfaces even if they do not have pending requests. This action reduces the system performance as a result of bus distribution latency. Moreover, giving every master an equal share of the bus is not always a good idea. Because highly bus access masters will get scheduled as the idle masters [4, 7, 16].

The RR scheme can be improved by using a queue as shown in Fig. 2. This enhancement scheme has the same principle to serve all masters requests in an arranged manner. Instead of checking all masters' interfaces, it uses a queue to save the number of any master requests the shared bus. Then the masters' requests are served in First-In-First-Out (FIFO) manner. This scheme will be implemented in section VI under the name of queuing round robin (QRR) scheme.

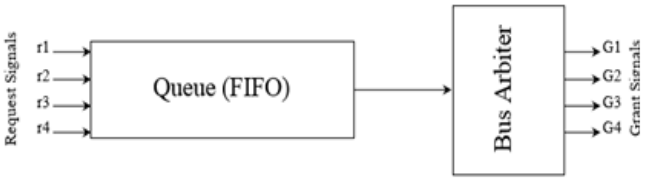


Fig. 2. Queuing Round Robin bus arbitration

B. Lottery Bus Arbitration

The role of the arbitration in the lottery bus arbitration algorithm is like a lottery manager that decides which lucky one can win the prize. The lottery manager accumulates the requests of the bus access from all of the masters. Each master is assigned a number of "lottery tickets". Then a pseudo random number is generated to choose one of the competing masters to be the winner of the lottery, favoring masters that have a larger number of tickets, and grant access is issued to the chosen master for a certain number of bus cycles. The random number guarantees that there is no master will monopolize the sharing resource [6, 9].

The inputs to the lottery manager are a set of requests and number of tickets held by each master. The output is a set of grant lines, one per master that defines which master had been allowed to access the bus.

According to the type of the tickets, lottery algorithms are divided into two types: static lottery and dynamic lottery [6]. In the static lottery, as shown in Fig. 3, each master has a fixed number of tickets. However, the number of tickets that is possessed by each master in the dynamic lottery are generated by a ticket generator, as shown in Fig. 4. For the both types, the same procedures are followed to decide the winner of the bus as the following:

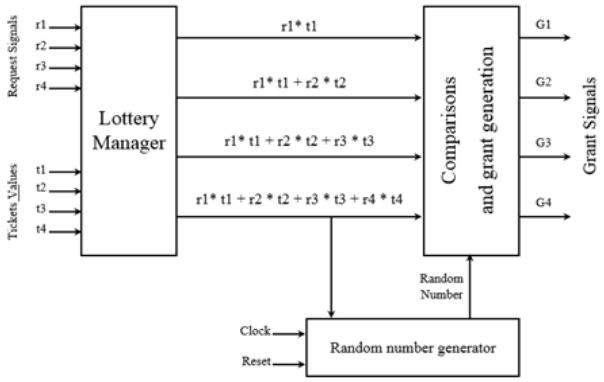


Fig. 3. Lottery arbiter with static tickets

- The lottery manager calculates the total tickets value for each master that has pending requests. This is given by $\sum_{j=1}^{n} r_j * t_j$, where n is the masters number, r is a Boolean variable represents the pending bus access request, and t is the number of tickets held by each master. For example, if the system has four processors and only three of them have pending requests, then n=4, r1=1, r2=0, r3=1, and r4=1. If the number of tickets that is possessed by each master are t1=1, t2=2, t3=3, and t4=4, then the total tickets values for processor1=1, processor2=1, processor3=4, and processor4=8.
- A pseudo-random number is generated in the range $[0, \sum_{j=1}^{n} r_j * t_j]$. It is supposed that the generated number is 5.
- If the generated number falls in the range $[0, r_1 * t_1]$, the bus is granted to master M1.
- In general, if the generated number lies in the range $[\sum_{k=1}^{i} r_k * t_k, \sum_{k=1}^{i+1} r_k * t_k]$ the bus is granted to master Mi+1. For our example, the generated number (5) falls in the range $[\sum_{k=1}^{3} r_k * t_k, \sum_{k=1}^{4} r_k * t_k] = [4, 8]$ so the bus is granted to processor4.

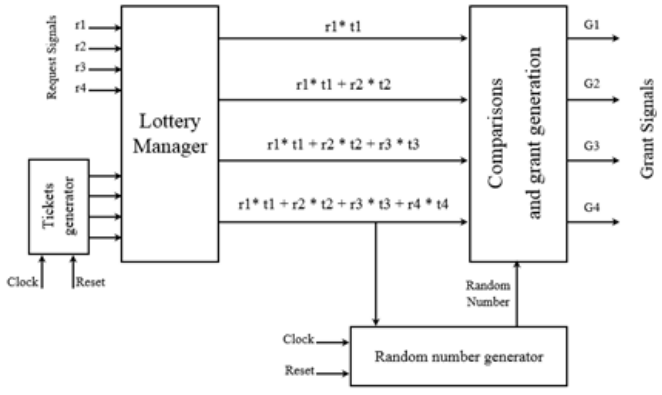


Fig. 4. Lottery arbiter with dynamic varying tickets

The advantages of the lottery algorithms are that all the masters that are requesting the bus get access to it (avoid starvation), and they improve the masters waiting time [3]. However, if the pseudo-random number is greater than the total tickets value, none of the masters will get access the bus. Moreover, the fixed ticket values in the static lottery algorithm give high chance to masters with high ticket values [6]. The limitation of the dynamic lottery algorithm is that the distribution of the ticket values is non-uniform [9]. In addition, it is more complex and required extra logic to calculate the tickets of each master at run time [3].

IV. Age-based Lottery Arbitration

As described in the related work section, the RR and the lottery bus arbitration compete in the performance sense. The developed scheme, in this work, represents the lottery bus arbitrations with additional enhancements to overcome their shortcoming.

The Age-Based Lottery (ABL), shown in Fig. 5, combines the dynamic lottery algorithm with the age-based algorithm from [2] to generate the ticket values. The ABL gives higher ticket values to masters that have recently won the bus. A preference during contention is given to the masters that are granted the bus recently. Each master has a ticket value can vary from 1 to MaxAge, which is a fixed parameter. The higher the ticket of a master, the more recently it has been granted the bus.

The algorithm shown in Fig. 6, illustrates the principle of ABL, which can be describe as follows: A CritNo flag is used for each master to balance the ticket value. When the master ticket value reaches MaxAge, the CritNo flag associated with that master is set. Then, if the ticket value is between the minimum age, which is 1, and MaxAge, then its ticket is decreased by one. If its ticket reaches 1, its flag is reset. If a master's flag is not set, its ticket is incremented by one after every bus grant. The integration between MaxAge and CritNo ensures a uniform distribution of the ticket values among the competing masters.

If there is only one request for the bus, the ABL will grant the bus to that request without any change on the corresponding ticket value since there is no any contention. On the other hand, the ticket values and CritNo flag must be changed when there are two masters or more compete on the bus. When there is more than one master request the bus and all of them reached the MaxAge, the associated ticket values reset to minimum age and their CritNo flags are reset.

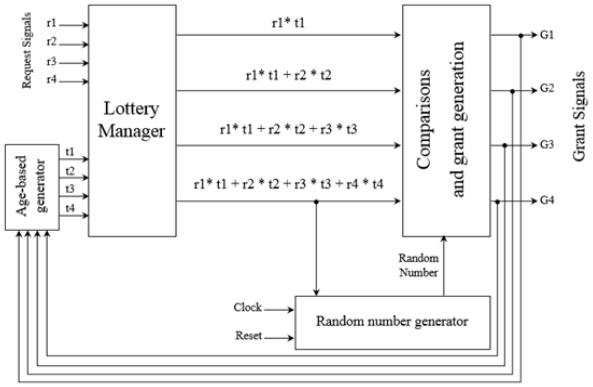


Fig. 5. Age-Based Lottery bus arbitration

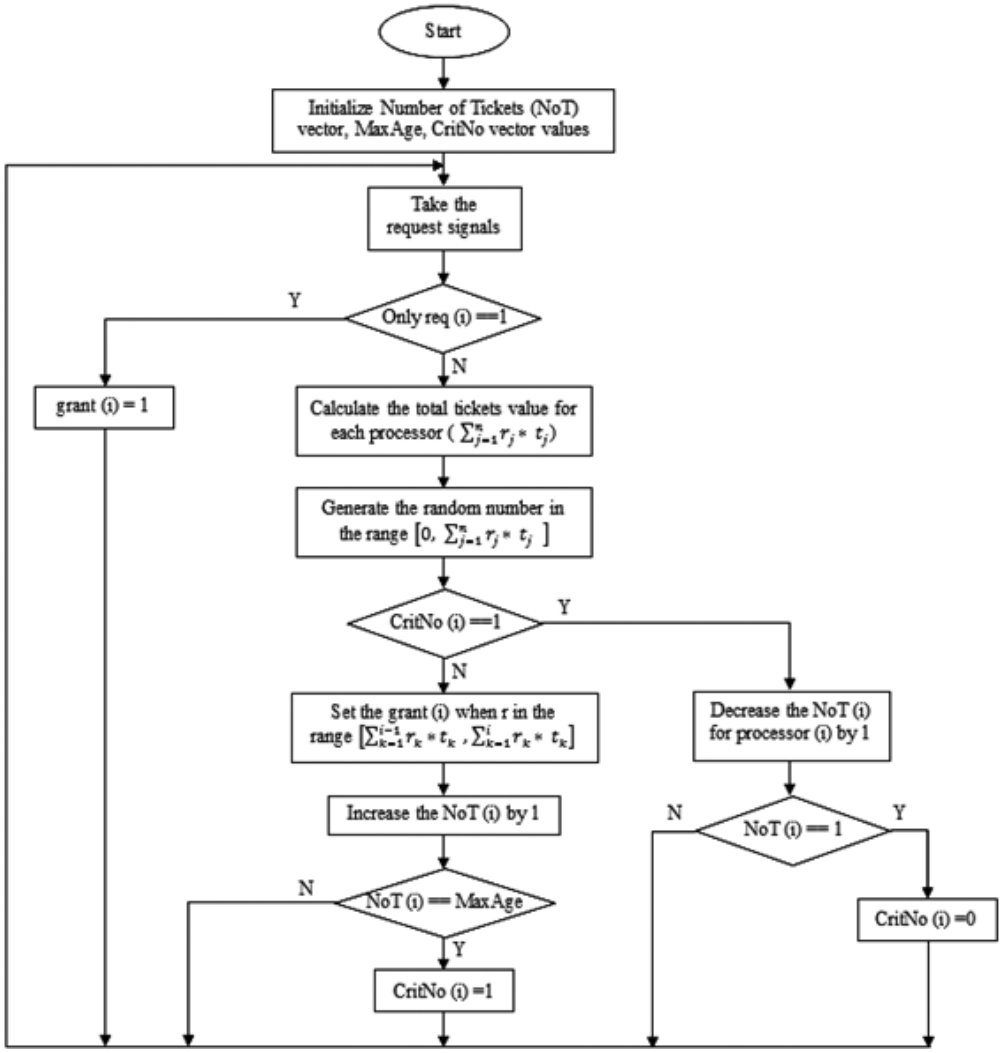


Fig. 6. The ABL algorithm for processor (i)

V. The Developed Bus Arbitration Scheme's Implementation and Testing

To test the developed scheme and compare it to the traditional schemes, the following schemes are implemented for four processors (masters) using VHDL language:

- Traditional Round Robin (RR)
- Queuing Round Robin (QRR)
- Age-based lottery (ABL)
- Dynamic lottery (DL)
- Static lottery (SL)

The results of testing the developed scheme and the traditional schemes are obtained by a VHDL simulation tool from Mentor Graphics Company, which is called ModelSim. The ModelSim has an ability to illustrate the simulation results as a waveform, which is an easy way to recognize the required results. The main parameters of the comparison are the bus utilization, the bus distribution, and the latency.

To compare the tested bus arbitrations, the grant output signals are observed by providing input signals such as bus requests, clock, reset, and additional signals related to the arbitration type.

For more illustration, two testing scenarios of requesting the bus are applied on the tested bus arbitrations. First, when all the four masters request the shared bus. Second, when only two masters request that bus. The simulation runs 100,000 clock cycles. In every cycle, one processor takes the permission to access the bus.

The simulation results appear as shown in

TABLE I and TABLE II. For the bus utilization parameter, results show that all schemes utilize the shared bus effectively in the first testing scenario since all processors request the bus as shown in TABLE I. However, in the second testing scenario, the RR suffers from idle bus cycles that are given to processor number 2 and 3 as shown in TABLE II. These cycles affect the overall performance of the CMP. They must be granted to the requested processors to improve the overall performance. The rest schemes utilize the bus effectively in the second testing

scenario, too. They serve the requested processors only so there is no idle bus cycle.

TABLE I. THE FIRST TESTING RESULTS (ALL PROCESSOR REQUEST THE SHARED BUS)

Processor / Arbiter	1	2	3	4	Divergence
RR	25000	25000	25000	25000	**0**
QRR	25000	25000	25000	25000	**0**
ABL	21642	25509	29964	22885	**3187.85**
DL	30329	24616	25138	19917	**3687.87**
SL	15132	20069	29918	34881	**7802.45**

TABLE II. THE SECOND TESTING RESULTS (ONLY PROCESSOR 1 AND 4 REQUEST THE SHARED BUS)

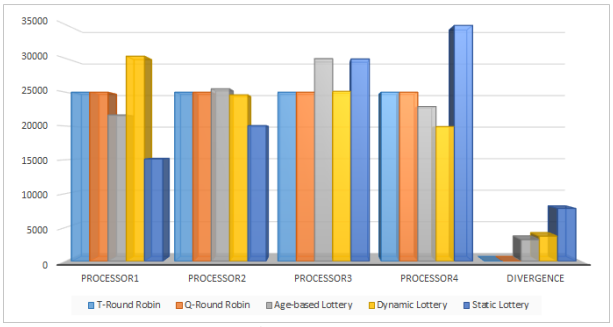
Processor / Arbiter	1	2	3	4	Divergence
RR	25000	0	0	25000	**0**
QRR	50000	0	0	50000	**0**
ABL	49999	0	0	50001	**1**
DL	39181	0	0	60819	**10819**
SL	30172	0	0	69828	**19828**

For the bus distribution parameter, results of the both testing scenarios show that the RR and QRR schemes surpass other schemes in the fair distribution. They give all processors the same priority degree to access the bus. In the first testing scenario, the ABL introduces fair distribution better than the dynamic and static lotteries. The ABL improves the distribution by 13.5% more than the DL and by 59% more than the SL. In the second testing scenario, the ABL has the same results of the QRR, which is better than the DL and SL by approximately 100%. Fig. 7 depicts the simulation results with the divergence in the bus distribution for each scheme.

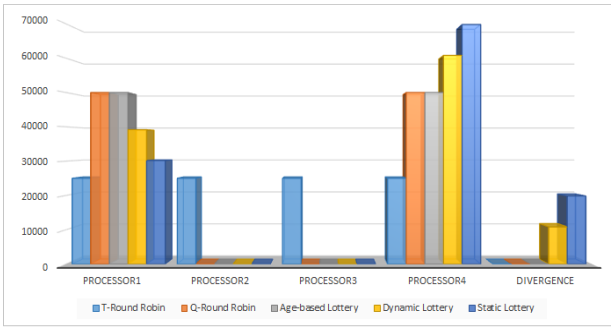
For the latency parameter, the latency time is static in the RR and QRR schemes since each processor gets access to the shared bus in its order as shown in Fig. 7. However, there is no chance for any processor to get access to the shared bus for two or more cycles successively. This problem has been solved by using the lottery schemes. The latency time is improved using the probabilistic lottery schemes. Moreover, in ABL the latency time to get access to the shared bus is improved by the term of age as shown in Appendix A.

VI. CONCLUSION AND FUTURE WORK

In this paper, a new bus arbitration scheme, which is called an Age-Based Lottery (ABL), is developed to improve the shared bus utilization and distribution. The ABL is a new combination scheme that combines Lottery algorithm with Age-Based algorithm. The ABL is designed to overcome the traditional static lottery (SL) and dynamic lottery (DL) arbitrations shortcomings. The simulation results illustrate that the developed scheme improves the bus utilization and distribution comparing to the traditional schemes by at least 13.5%.



a)The first testing results



b)The second testing results

Fig. 7. Simulation results with the divergence in the bus distribution for bus arbitration schemes

The shared bus in this paper limits the number of masters that can share it. This paper can be extended by designing an alternative bus implementation such as hierarchy of physical buses (tree bus) which may increase the number of masters in CMP systems.

REFERENCES

[1] K. Olukotun, L. Hammond, and J. Laudon, "Chip multiprocessor architecture: techniques to improve throughput and latency," Synthesis Lectures on Computer Architecture, vol. 2, no. 1, pp. 1-145, 2007.

[2] N. Ramasubramanian, P. Krishnan, and V. Kamakoti, "Studies on the Performance of Two New Bus Arbitration Schemes for MultiCore Processors," in International Advance Computing Conference (IACC), Patiala, pp. 1192-1196, 2009.

[3] B. Tiwari, R. Chandraker, and N. Goel, "Comparative analysis of different lottery bus arbitration techniques for SoC communication," in 2016 International Conference on Computational Techniques in Information and Communication Technologies (ICCTICT), pp. 495-499, 2016.

[4] J. Gupta and N. Goel, "Efficient Bus Arbitration Protocol for SoC Design," in International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM) India, pp. 396-400, 2015.

[5] F. Wang, "Bus arbitration techniques to reduce access latency," ed: Google Patents, 2013.

[6] P. Bajaj and D. Padole, "Arbitration schemes for multiprocessor Shared Bus," New Trends and Developments in Automotive Engineering, INTECH Publisher, 2011.

[7] A. Kulmala, E. Salminen, and T. Hamalainen, "Distributed bus arbitration algorithm comparison on FPGA-based MPEG-4 multiprocessor system on chip," Computers & Digital Techniques, IET, vol. 2, no. 4, pp. 314-325, 2008.

[8] D. Shanthi and R. Amutha, "Design of efficient on-chip communication architecture in MpSoC," in International Conference on Recent Trends

in Information Technology (ICRTIT), Chennai, Tamil Nadu, pp. 364-369, 2011.

[9] D. Shanthi and R. Amutha, "Design Approach to Implementation Of Arbitration Algorithm In Shared Bus Architectures (MPSoC)," Computer Engineering and Intelligent Systems, vol. 2, no. 4, pp. 185-196, 2011.

[10] A. K. Singh, A. Shrivastava, and G. Tomar, "Design and Implementation of High Performance AHB Reconfigurable Arbiter for Onchip Bus Architecture," in International Conference on Communication Systems and Network Technologies (CSNT), Katra, Jammu, pp. 455-459, 2011.

[11] D. Valle, G. Pablo, D. Atienza, I. Magan, J. G. Flores, E. A. Perez, et al., "Architectural exploration of MPSoC designs based on an FPGA emulation framework," in Proceedings of XXI Conference on Design of Circuits and Integrated Systems (DCIS), pp. 12-18, 2006.

[12] W. Zhang, G.-M. Du, Y. Xu, M.-L. Gao, L.-F. Geng, B. Zhang, et al., "Design of a hierarchy-bus based MPSoC on FPGA," in 8th International Conference on Solid-State and Integrated Circuit Technology (ICSICT'06) Shanghai, pp. 1966-1968, 2006.

[13] W.-T. Zhang, L.-F. Geng, D.-L. Zhang, G.-M. Du, M.-L. Gao, W. Zhang, et al., "Design of heterogeneous MPSoC on FPGA," in 7th International Conference on ASIC (ASICON'07) Guilin, pp. 102-105, 2007.

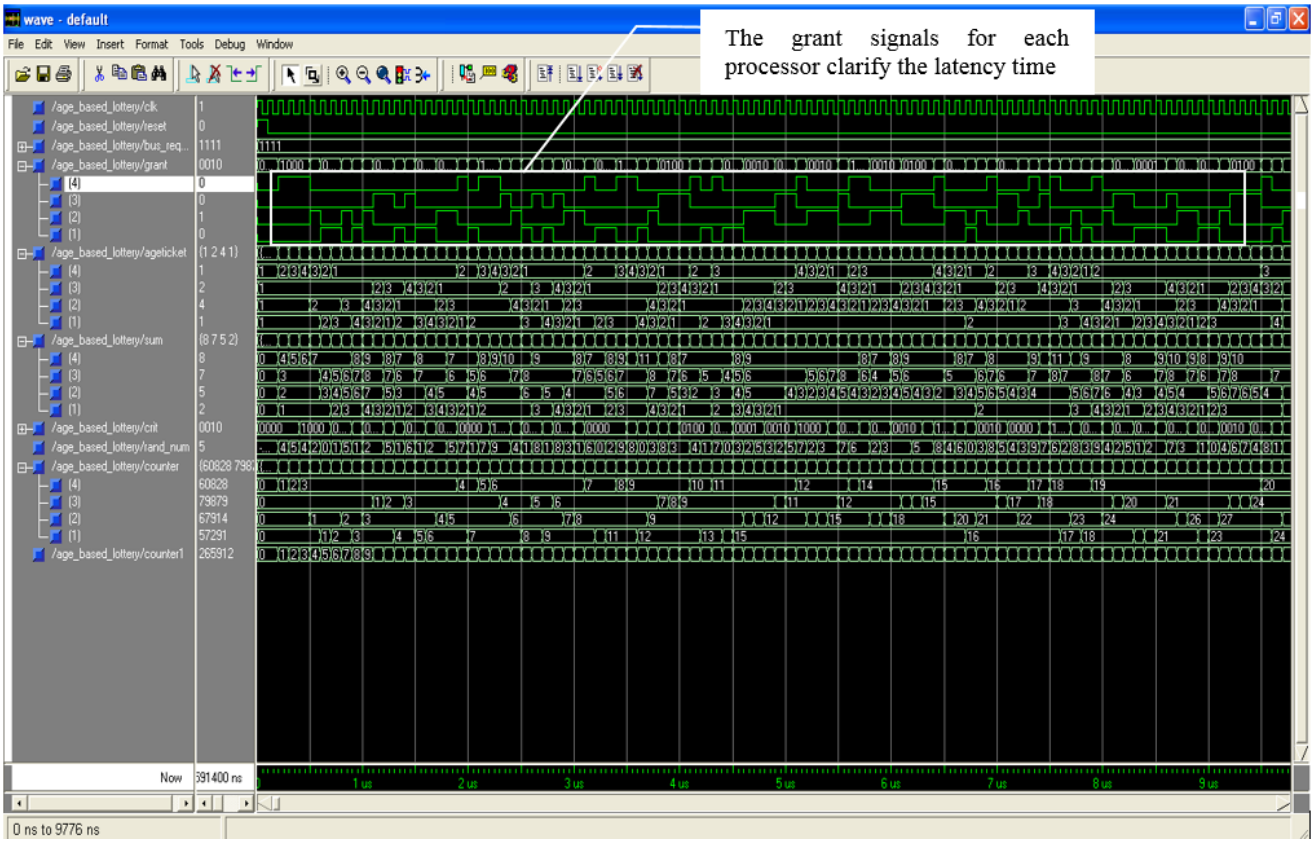
[14] D. A. Patterson and J. L. Hennessy, Computer organization and design: the hardware/software interface, Newnes, 2013.

[15] H. El-Rewini and M. Abd-El-Barr, Advanced computer architecture and parallel processing vol. 42, John Wiley & Sons, 2005.

[16] I. S. Rajput and D. Gupta, "A Priority based Round Robin CPU Scheduling Algorithm for Real Time Systems," International Journal of Innovations in Engineering and Technology, vol. 1, no. 3, 2012.

APPENDIX A

SIMULATION WAVEFORM FOR THE AGE-BASED LOTTERY BUS ARBITRATION



A Semantic Interpretation of Unusual Behaviors Extracted from Outliers of Moving Objects Trajectories

Sana CHAKRI, Said RAGHAY, Salah EL HADAJ
Laboratory of applied mathematics and informatics,
Cadi Ayyad University
Marrakesh, Morocco

***Abstract*—The increasing use of location-aware devices has led to generate a huge volume of data from satellite images and mobile sensors; these data can be classified into geographical data. And traces generated by objects moving on geographical territory, these traces are usually modeled as streams of spatiotemporal points called trajectories. Integrating trajectory sample points with geographical and contextual data before applying mining techniques can be more gainful for the application users. It contributes to produce significant knowledge about movements and provide applications with richer and more meaningful patterns. Trajectory Outliers are a sort of patterns that can be extracted from trajectories. However, the majority of algorithms proposed for discovering outliers are based on the geometric side of trajectories; our approach extends these works to produce outliers based on semantic trajectories in order to give meaning to the outliers extracted, and to understand the unusual behaviors that can be detected. To prove the efficiency of the approach proposed we show some experimental results.**

***Keywords*—Moving objects analysis; spatial databases; data mining; Semantic clustering; semantic trajectories**

I. INTRODUCTION

Researchers from spatial databases, GIS, data mining, and knowledge extraction communities have developed several techniques for mobility analysis. As consequence three research areas have been expended; The first one focuses on data modeling to provide definitions and extensions of trajectory related data types such as moving objects, points, lines, or regions. The second deals with data management to optimize the storage of mobility data with suitable indexing and querying techniques. And the last one that is the main topic of this research deals with the analysis of patterns that can be extracted from stored data like trajectories by using spatiotemporal data mining algorithms. Several data mining methods have been proposed for extracting patterns from trajectories. However, the majority of them use trajectories without looking for any additional information, and yet by considering only the raw trajectory data, discovering why an object followed a different route become very complex since no additional information (called semantic) is given about the moving object. This additional information can hide behind a lot of meanings; in fact it can lead to a better understanding of the patterns extracted. This is can be achieved by combining the raw mobility tracks (e.g., the GPS records) with related contextual data in order to use semantic trajectories instead of focusing only on the geometric side of trajectories. Therefore, applying mining techniques on semantic trajectories continues to prove success stories in discovering useful and non-trivial behavioral patterns of moving objects. Several data mining methods have been proposed for extracting behaviors from trajectories such as chasing behaviors [1], flocks [2], avoidance [3], etc. In this paper, we focus on trajectory outlier detection. Trajectory outliers are sort of patterns that can be extracted from semantic trajectories of moving objects. The objective in trajectory outlier detection is to find trajectories that do not comply with the general behavior of the trajectory dataset. While most of pattern analysis focuses on patterns that are common in the trajectory dataset, outlier detection focuses on rare patterns such as trajectories that follow a path different from the common path followed by most of the other moving objects, or objects following the same path but behave differently than the other objects (very slow or fast objects compared to the majority of the moving objects). Trajectory outlier detection can be very useful in traffic analysis, it helps understand the flow of people that move between regions, how this flow is distributed and what are the characteristics of the movements. In high traffic routes, outliers can give some alternative paths that can reduce the volume of traffic, or give the best or worst path that links two areas, by extracting outliers, users can easily discover suspicious behaviors like company cars that escape from their normal route. In fact, detecting semantic outliers proves his efficiency, especially to discover suspicious behaviors in a group of people, to find alternative routes in traffic analysis in many applications such as transportation, ecology, animal tracking, health sector, crime sector, and climatology, etc. Indeed, by adding semantics to outliers, the analysis became more performed; we can discover the reasons for each behavior extracted. The interpretation of outliers can provide more information to the decision maker. Thus, many new applications are interested in understanding and using semantic interpretation of the moving object behavior. Semantics refers essentially to additional contextual and geographical information available about the moving object, apart its position. Semantics contain both the geometric properties of the moving object as well as the geographic properties and any other additional information like the moving object's activity, mode of transportation, speed or any data that can help give more meaning to the behavior extracted. The purpose of this research is to find spatial, spatiotemporal and temporal outliers among semantic trajectories, analyzing them

taking into account their semantic data to understand the meaning of the outliers detected, especially to give an answer to the famous question "why an object could deviate from a group?"

The rest of the paper will be organized as follows: in section 2 we will present the related work, in section 3 we will present the semantic outlier detection in which we will discuss the flow to construct semantic trajectories then apply mining algorithms for extracting outliers, section 4 will provide with the methodology used to give meaning to outliers extracted. Section 5 illustrates the algorithms used, section 6 gives case of study and in section 7 we will discuss the work proposed and gives some comparisons.

II. Related Work

To our best knowledge, there are a medium number of researches to detect outliers in trajectories. However, only a few of them focus on semantic as they focus on geometric data, so we can split this research area to two essential fields: the first filed focuses only on the geometrical side of outliers like [4] which is an efficient technique to discover spatiotemporal outliers and causal relationships between them. Another one is proposed in [5] used for detecting outlier sequences in precipitation data. A roughest approach is described in [6] for spatiotemporal outlier detection. A survey was presented in [7], in which more approaches for outlier detection in temporal and spatiotemporal data were discussed. The second filed handles semantic data besides than geometrical one, their approaches are closer to our research like [8 9]. For the first work, authors try to find outliers between regions of interest, in the second authors try to find the specific standard path that the outlier deviates and propose to give a meaning to it. In [10], the main objective is to discover outliers among trajectories that have the same goal and move between the same regions and to give a meaning to these outliers extracted. Authors in [11] tries to extract anomalous behaviors in single-trajectory data, in [12], authors propose a method of detecting avoidance behaviors between moving objects, and the paper [13] tries to detect abnormal pedestrian behavior based on a new trajectory model, [14] and [15] are recent works that tries to detect outliers based on vehicle trajectories and multi-factors. Our work extends these works by giving a global approach which starts by merging GPS feeds with semantic data to produce semantic trajectories, then applying the mining algorithm proposed in order to give a very deeper analysis to the outliers extracted, we also try to analyze the outliers extracted according to semantic data to give more precision to the reasons for which some moving objects deviate from the main route.

III. Semantic Outlier Detection

A. Enriching trajectories with semantics

Trajectories of moving objects present a huge data warehouse where users can extract several information according to the application domain studied, this is can be achieved by applying data mining techniques based on both temporal and spatial data mining algorithms. However, spatiotemporal data mining is only one step between all the knowledge discovery processes. In fact, to extract meaningful knowledge, the trajectories must follow several steps to be ready to use for data mining, our approach gives the whole process that trajectories pursuit to be structured and enriched before being used. First, it consists of enriching trajectories with semantic data throughout a process where the raw trajectories will be built from GPS feeds, cleaned, well structured, and enriched before applying data mining algorithms. Figure 1 illustrates the process pursued to build semantic trajectories; it is structured in three steps to prepare trajectories for data mining. The first step is raw trajectories building, where we try to prepare trajectories by cleaning and structuring the GPS points which can be defined as: Definition1: A point p is a tuple (x,y,t), where x and y are spatial coordinates and t is the time instant in which the coordinates were collected. The formatted points produce a healthy raw trajectory that is defined as: Definition2: A trajectory T is a list of points (p1, p2, p3,..., pn), where pi = (xi,yi,ti) and t1< t2 < t3 < ... < tn.

The second step (Semantic Trajectory Enrichment) takes as input these structured trajectories, and tries to segment it into episodes (sub-trajectories) of stops and moves, then annotated them with related contextual data to product semantic trajectories. A sub-trajectory can be defined as:

Definition3: Let T = (p1, p2, p3,..., pn), be a trajectory. A sub-trajectory S of T is a list of consecutive points (pk, pk+1, pk+2,..., pm), where p Є T, k ≥ A, and m ≤ n.

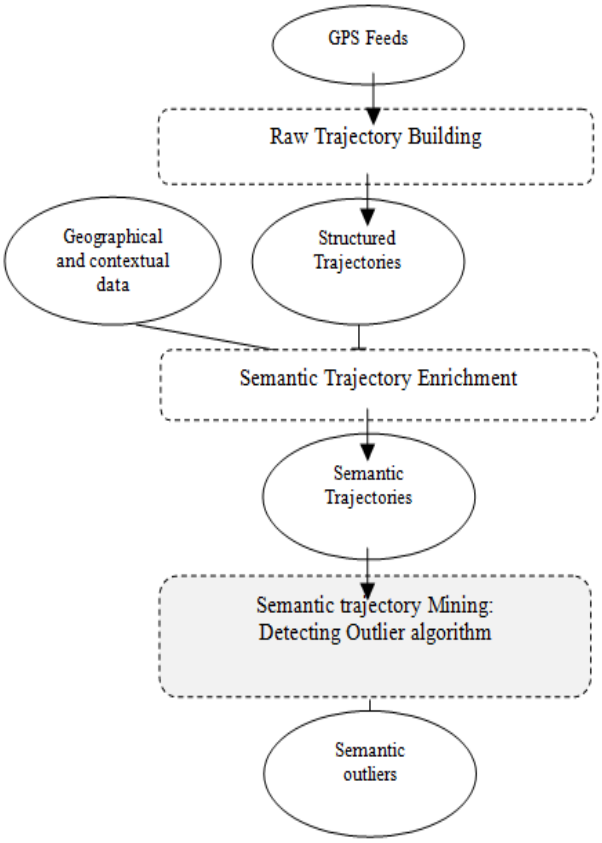


Fig. 1. Trajectory enrichment and extraction process

These semantic trajectories will be the input of the third phase that is semantic trajectory mining, where we will be able to apply mining algorithms to extract suspicious behaviors of moving objects (outliers), more details about the process of enrichment are explained in [16].

B. Extracting Semantic Outliers

Globally, outlier analysis in classical databases reveals odd objects which appear to be inconsistent with the other objects in the database. This definition implies that the object is significantly different from the overall database as a whole. However, in case of spatiotemporal databases, it is possible for an object to appear consistent with the entire database objects, but appear unusual with a local neighborhood [17, 18]. Therefore, we can say that an outlier is a spatiotemporally geo referenced object whose non-spatiotemporal attribute values differ from objects in its spatiotemporal neighborhood. Otherwise, a spatiotemporal outlier is a local shakiness or inconstancy. An outlier can refer either the whole trajectory, or more often it refers parts of trajectories called sub-trajectories, where the moving object chooses to behave differently compared to the rest of the other moving objects trajectories and then becomes suspicious [19].

1) Methodology

The purpose is to find spatiotemporal and temporal outliers between regions of interest [20], Analyzing them with semantic data to understand the meaning of the outliers detected. Spatiotemporal outliers refers to sub-trajectories that have spatial and temporal difference compared to common trajectories, while temporal outliers refers to moving objects that behave spatially like the majority of the other moving objects, but temporally they are different; for instance moving objects that took the same route but they accelerate or they mark an important number of stops which make them seen as suspicious moving objects. The analysis presented in this paper are made on sub-trajectories that rely regions of interest which are shapes that have different size and format, depending on the application, they can be regions ROI, lines LOI, or even points POI, they can be districts, dense areas, hotspots, important places, etc. generally a region of interest can be a pre-defined important place or computed by an algorithm that finds dense areas. In our case we consider a region as a point, line or polygon, which is a well-known concept in GIS community. The use of regions allows filtering from the whole dataset only the sub-trajectories that move between the same regions, outliers will be searched among these sets what significantly reduces the search space for outliers. Among the trajectories that cross all regions, we are only interested in the part of trajectories (sub-trajectories) that move between specific regions, we call these sub-trajectories Nominees. After defining the set of nominees, we start looking for temporal outliers, and spatial outliers in which we extract from them spatiotemporal outliers. A nominee will be a spatial outlier when it follows a different path in relation to the majority of the sub-trajectories from its group, and it can be a temporal outlier if it follows the same path, but shows different behaviors compared to the other moving objects. In general, we have two types of path: Populated path that have many trajectories in its proximity. And depopulated Path, it has less trajectories around. The spatial and the spatiotemporal outliers will be extracted from depopulated paths, while the temporal outliers will be extracted from the populated paths.

To detect if the nominee is in the populated or the depopulated path, we introduce the concept of proximity; A nominee is in proximity to a point if it is close to the point, if a point has a few nominees in its proximity, then at that time the moving object was following a path different from the majority of Nominees, it is in a depopulated path. The maximal distance for a nominee to be in proximity to a point is called PD (proximity distance). In general, there is at least one main route used between two regions, which is more frequent than alternative ways. The minimal amount of a nominee (MA) is the minimal number of points that each point of a nominee should have in its proximity to be part of this main route. The nominee that has all its points in a populated path is considered common trajectory. The nominee that has at least one point where the cardinality of its proximity is less than MA is called expected outlier. So the nominee will always be either Common trajectory or expected outlier. The spatial outliers will be extracted from expected outliers, and the temporal outliers will be extracted from common trajectories. When two nominees leave the start region at a time interval inferior to Maximal Tolerance MT. we can say that they are synchronized.

2) The process

The general process, as shown by images in figure 2, starts by looking at sub-trajectories that have the same arrival, in order to define the nominees (figure 2.a). Like said before; a nominee will be an outlier when he follows different path compared to the majority of the sub-trajectories from its group, or when it behaves differently even if he follows the same

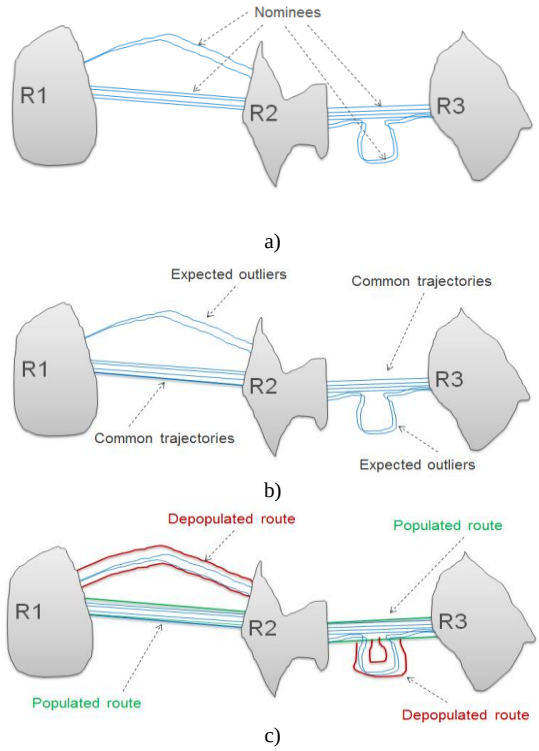


Fig. 2. Steps for detecting outliers

route. So we try to define the expected outliers which are sub-trajectories that have a few neighbors in their proximities, and the common trajectories which have a lot of neighbors in their proximities (figure 2.b). The route followed by the expected outliers is considered as depopulated route, while the route followed by common trajectories is a populated route (figure 2.c).

After grouping the expected outliers and to select from them the spatial outliers, we verify two conditions; the first one is that the expected outlier connects two regions? If yes, we move to the second one that is for these two regions, is there any populated route detected? Because if we want to discuss the existence of spatial outliers, it should be at first a populated route that the majority of moving objects follows, then the deviation can be seen as spatial outlier, if there is no populated route, we can't discuss spatial outliers. When all nominees between two regions are expected outliers, which means there is no common trajectories; there is no populated path that an object could avoid or deviate. Contrariwise, if there is at least one populated path, then the expected outlier did really perform a detour, and becomes spatial outlier. No spatial outlier will exist if there is no common trajectories, as assumption to define a spatial outlier, It should move between two regions of interest, and there must be a populated path that connects the regions such that the spatial outlier should deviate from it, therefore, any sub-trajectory that uses a path different from the populated path is a spatial outlier. For the temporal outliers, they will be extracted from common trajectories. As said before; the temporal outliers are sub-trajectories that follow the same path used by the most of moving objects, but behave differently than the other objects; for example some moving object can make several stops in his way, so it can be seen as a very slow object compared to the majority of the other moving objects, or contrariwise, it can be seen more fast. After extracting the outliers detected, we classify them first according to their speed, and then we try to analyze each group of outliers classified by proposing a meaning to their deviations by looking for the reasons of deviation.

IV. Giving Meaning to Unusual Behaviors

After extracting outliers from semantic trajectories, the main goal of the next step is to add meaning to the outliers extracted. The next step is about splitting the outliers extracted to several types according to their semantic interpretation;

A. Spatiotemporal outliers

Figure 3 illustrates the classification of spatiotemporal outliers extracted from spatial outliers.

1) Stop outliers

It occurs when the moving object made a stop for some time during the deviation, for instance the moving object had an appointment, a meeting, go shopping after work, pick up the children at school, go with friends, pass by a market, or something to do somewhere else that was not in the standard path. This is an intentional detour with a reason. To discover if an outlier has a stop we need to look for stops not in the complete outlier trajectory, but only in the sub-trajectory that corresponds to the outlier (deviation), i.e., the outlier segment. We consider as a stop a sub-trajectory that its speed is close to zero for a minimal amount of time (MT).

2) Emergency outliers

It occurred when the moving object took an alternative route and shows an important acceleration of its speed, the reasons can be almost about an emergency case like an ambulance transporting patient, or someone trying to escape from police, etc. to detect if there is an emergency we need to compare the speed of the fast outlier with the speed of the synchronized outliers that took the same deviation. We consider that there is an emergency outlier if the speed of the fast outlier is higher than the double of the average speed of the synchronized outliers detected in the same derived route.

3) Regular outliers

It occurs when the moving object deviates from the populated route without an important change of speed, or with a degradation of speed. This may reveal that the populated route is temporarily busy or is under reconstructions, or there is an accident, or even there is an event that block the path, so the moving object is forced to deviate from the populated route, Which can cause a big traffic on the alternative ways, and as consequence, the speed of the moving object may decrease. Our algorithm assembles all these reasons in three types of outliers: the blocked route outlier, the avoided route outlier, and the traffic jam outlier.

a) Blocked route outliers

Expresses any deviation because something happens close to the populated route which causes some blockage, for instance, an accident, route reconstructions, or some artistic events like a carnival or a concert. The challenge is how to discover the case that blocked the populated route; we start by analyzing only the part of the closest populated route deviated by the outlier (we call it the main segments), then we look if there is an activity around the main segments, if yes, we verify the time of this activity to be sure that the outlier was generated in the moment of the action. And finally we verify that at the time of the activity, there are no synchronized segments in the populated route, to prove that the path was blocked by the event, so the moving objects were forced to take an alternative route. Thus, a blocked route outlier is an outlier that deviates from the populated route because a blocking activity is happening close to the populated route at the same time of the deviation.

b) Avoided route outliers

This type of outliers is similar of the first type, the only difference is that there is an activity in the populated route, but this activity doesn't cause any blockage, an example could be a police checkpoint; In this case, the majority of moving objects will take the populated route normally, but some of them choose to avoid this event. For discovering such type of outliers we verify if there is an activity in the populated route. If yes we verify if there are some trajectories which Travers the populated route in the time of the activity to prove that the activity doesn't block the route. At this time we can say that this outlier is of type avoided route outlier.

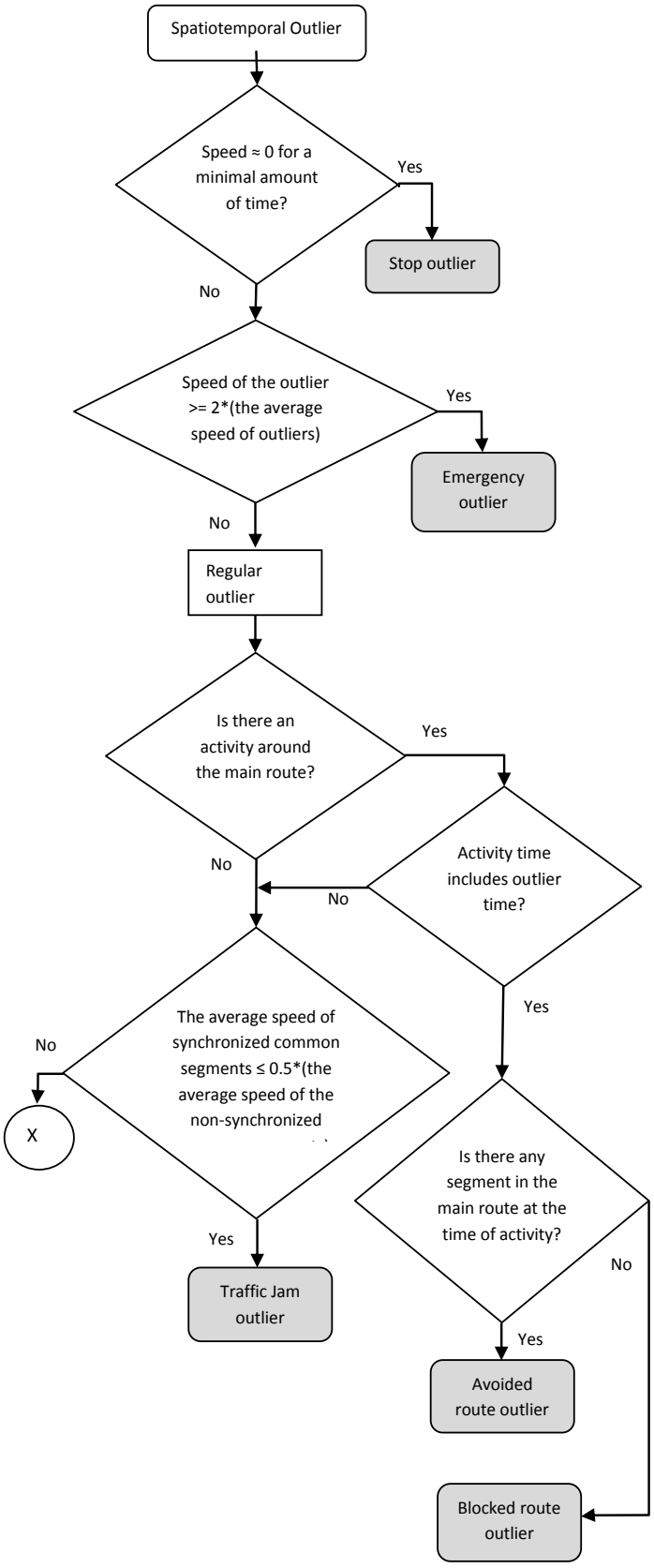


Fig. 3. Operating logic schema for giving meaning to spatiotemporal outliers extracted

c) Traffic jam outliers

Expresses deviations due to a heavy charge at the rush hour, it occurs if we found an outlier, but no activity is blocking the populated route, so we start looking if there is a traffic jam. For that we look for the slow traffic in the populated route at the time of the outlier. To measure the speed on the populated route at the same moment that the outlier deviated from it, we need to look only at the segments of the synchronized common trajectories. The average speed of all synchronized common segments in the same populated route is compared to the speed of the non-synchronized common segments in the same route. We consider that there is a traffic jam when the average speed of those who are synchronized is less than half of the average speed of the non-synchronized.

B. Temporal outliers

Temporal outliers are common trajectories that follow the populated route, but with an important difference of the speed compared to the other common trajectories. For extracting such type of outliers, we make use of the average speed used by the moving objects in the populated route, we make a comparison between each sub-trajectory from the common trajectories and the average speed for all common trajectories that traverse the same route with some tolerance, and we extract two essential types; temporal emergency outliers, and temporal stop outliers.

1) Temporal emergency outliers

This type of outliers is extracted from fast common trajectories that traverse the populated route. It occurred when the moving object stay in the populated route but shows an important acceleration of its speed, the reasons can be almost about an emergency case. To detect if there is a temporal emergency, we need to compare the speed of the fast common trajectory detected with the speed of the synchronized common trajectories that took the same populated route. We consider that there is a temporal emergency outlier if the speed of the fast common trajectory is higher than the double of the average speed of the synchronized common trajectories detected in the same populated route.

2) Temporal stop outliers

The temporal stop outliers are common trajectories that Travers the populated route with a very slow speed compared to the synchronized common trajectories in the same route, it occurs when the moving object made a stop for some time in the populated route. To discover if the common trajectory has a stop we need to look for stops in the sub-trajectory that corresponds to the common trajectory. We consider as a temporal stop outlier a sub-trajectory that its speed is close to zero for a minimal amount of time (MT).

V. ALGORITHM

In this section we present the algorithms used to detect and interpret the outliers extracted. Figure 4 shows the pseudo-code of the main algorithm.

The algorithm starts by computing the nominees that move between two regions, with the function detectNominee. This function checks for every trajectory if it intersects the pair of regions. Once the nominees are computed, the algorithm searches for the common trajectories (trajectories that follow

the populated route) with the function findCommon, considering the parameters PD and MA, this function checks for all points of a nominee in the set if the number of points in proximity is greater than MA.

```
INPUT:
T; // set of trajectories
R; // set of regions
PD; // Proximity distance
MA; // Minimal amount
MT; // Maximal tolerance
MS; // minimal speed
MD; // minimal duration
ML; // minimal lenth
MSA; // minimal Stops allowed
OUTPUT:
ClassifiedOutliers; // set of outliers
for each (reg1,reg2)in R
{   N = detectNominees(T,reg1,reg2); // find nominees
    CommonSet = FindCommon(N,PD,MA); // find common trajectories
    if(not_empty(CommonSet)) then
    {   ClassifiedOutliers.TemporalEmergencyO(CommonSet); // find temporal emergency outliers
        ClassifiedOutliers.TemporalStopO(CommonSet,MS,MD,MSA); // find temporal outliers
        OutlierSet = N  MINUS  CommonSet; // Set of spatial outliers
        For each outlier in OutlierSet
        {   outlierSegments = getOutSeg(outlier,ML); // all segments of the outlier
            ClassifiedOutliers.StopOutlier  = findStops(outlierSegments, MS,MD);
            O = outlierSegments - ClassifiedOutliers.StopOutliers;
            ClassifiedOutliers.EOutlier = findEmergencyOutlier(O);
            O = O - ClassifiedOutliers.EOutlier;
            ClassifiedOutliers.ROutlier = findRegularOuteOutlier(O);
        }
    }
    return ClassifiedOutliers;
}
```

Fig. 4. Pseudo Code of the main algorithm

If this is the case, then the nominee is considered as common trajectory. If the set of common trajectories is not empty, the algorithm tries to extract temporal emergency outliers and temporal stop outliers, and then it goes for finding the spatial outliers, since there is a common path that connects both regions.

In the next step, the algorithm goal is to add meaning to the outliers extracted. So we go further in semantics by extracting the types of outliers; temporal stop outliers, Temporal Emergency outliers, stop Outliers, Emergency Outliers, Blocked Route Outliers, Avoided Route Outliers and Traffic Jam Outliers. Figure 5 illustrate the algorithms used. The temporal stop outliers and Stop Outliers are classical types that the majority of data mining algorithms use to detect stops of moving objects, The Emergency outliers are extracted from fast outliers, and the temporal emergency outliers are extracted from fast common trajectories. The Regular Outlier captures all outliers that keep almost the same or less speed, and then the algorithm tries to detect from this type the blocked Route Outliers, the avoided Route Outliers and the traffic Jam Outliers.

```
INPUT:
OutlierSegments outSegs; // set of outlier Segments

OUTPUT:
EmergencyOutlier  EO; // set of emergency outliers

for each Out in outSegs
{
    if(out.speed) >= (2*(avgSpeed(outSegs)))
    {
        EO.add(out);
    }
}
Return EO;
```

A

```
INPUT:
CommonSet commonSet; // set of common trajectories

OUTPUT:
TemporalEmergencyOutlier  TEO; // set of temporal emergency outliers

for each ComnT in commonSet
{
   if(ComnT.speed) >= (2*(avgSpeed(commonSet)))
   {
       TEO.add(ComnT);
   }
}
Return TEO;
```

B

```
INPUT:
OutlierSegments outSegs; // set of outlier Segments

OUTPUT:
RegularOutlier RO; // set of Blocked route outliers

for each out in outSegs
{
    ComnSeg CS = getComnSeg(Out);
    Regions R = getIntersection(Cs); // get all regions that intersect the common segment
    Activities A = R.getActivities(out.time) // get the activities that
                                             //appnehed at the same time of the outlier
   if (not_empty(Activities))
   {
       SynCmnSeg SCS = getsynchronizedCommon(out);
       if(empty(SCS))
           RO.addBlockedRouteOutlier(out);
       else
           RO.addAvoidedRouteOutlier(out);
   }
   else
   {
       SynCmnSeg SCS = getsynchronizedCommon(out);
       NSynCmnSeg NCS = getNNsynchronizedCommon(out);
   if(avg(SCS.speed) <= (2*(avg(NCS.time)))
       RO.addTrafficJam(out);
   }
}
Return RO;
```

C

Fig. 5. Pseudo code of semantics outliers; A : Pseudo Code of emergency outliers, B : Pseudo code of temporal emergency outliers, C : Pseudo code of regular outliers

VI. Experimental Results

In this section we present the results of experiments with real data, before that we provide with a presentation of the general architecture of our approach in the figure 6. Our approach contains tree main phases in the general architecture, the first one concerns the data preprocessing where the GPS

feeds will be treated to become sample trajectories, then they will be able to be structured in the enrichment process [21]. In the second phase we make use of the Weka-STPM toolkit [22] which is a java toolkit for semantic trajectory data mining and visualization, we have used the CB-SMot algorithm to create Stops and Moves [23].

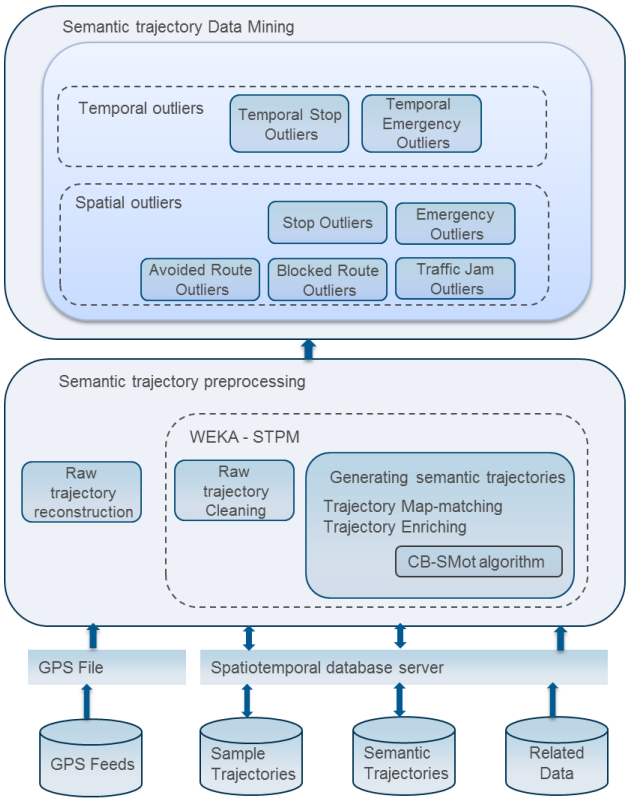


Fig. 6. General architecture

After the Semantic process we move to the last phase when we apply the Semantic Outlier Analysis algorithm in which we extract the outliers then add meanings.

For the experimental results we try to analyze two data sets to prove the efficiency of our method, these datasets rare taken from [24 25 26 27 28]. The first one contains trajectories of School Buses dataset which consists of 145 trajectories of two school buses collecting and delivering students around Athens metropolitan area in Greece for 108 distinct days. Notice that we analyzed only trajectories from Monday to Friday. The second are Trucks dataset which consists of 276 trajectories of 50 trucks delivering concrete to several construction places around Athens metropolitan area in Greece for 33 distinct days. The structure of each record is as follows: {obj-id, traj-id, date(dd/mm/yyyy), time(hh:mm:ss), lat, lon, x, y} where (lat, lon) is in WGS84 reference system and (x, y) is in GGRS87 reference system. These datasets are interesting for analyzing outliers because this type of drivers, in general, knows different routes to reach the same place. Therefore, we can find the alternative routes (outliers) in relation to the standard path. In this experiment we consider as interesting regions the districts around Athens metropolitan area. The application domain data are all about information about drivers, the number of students for the school buses, the type and the number of products for the trucks, the noun of the districts and the activities of the drivers and regions in this period.

The results for school buses are displayed below;

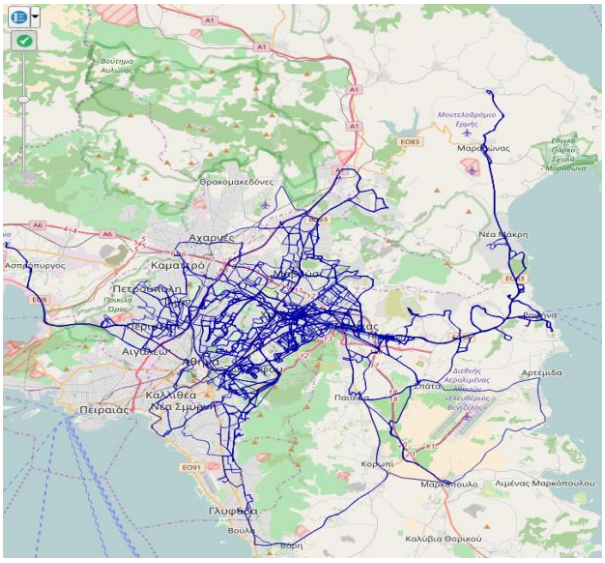


Fig. 7. School bus trajectories

Fig. 8. Common trajectories

Fig. 9. Outliers extracted

TABLE. I. SCHOOL BUS OUTLIERS EXTRACTED

Nominees	Expected outliers	Spatiotemporal outliers	Common trajectories	Temporal outliers
54598	2402	1778	52196	54

TABLE. II. SEMANTIC SPATIOTEMPORAL OUTLIERS FROM SCHOOL BUS TRAJECTORIES

Spatiotemporal outliers					
Stop	***Emergency***	***Regular***			
448	-	1330			
		Blocked route	***Avoided route***	***Traffic Jam***	***Others***
		21	706	454	149

TABLE. III. SEMANTIC TEMPORAL OUTLIERS FROM SCHOOL BUS TRAJECTORIES

Temporal outliers	
Stop	***Emergency***
54	-

The experimental results for school buses outliers show that the trajectories contain 1778 spatiotemporal outliers from 2402 expected outliers, and contain 54 temporal outliers from 52196 common trajectories, the spatiotemporal outliers contain 448 stop outliers and 1330 regular outliers, in which there are 21 blocked route outliers, 706 avoided route outliers, 454 traffic jam outliers, and 149 outliers none defined.

The results for trucks are displayed below

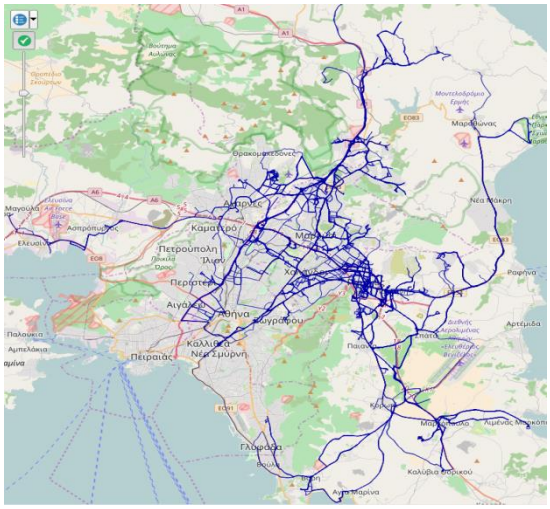


Fig. 10. Trucks trajectories

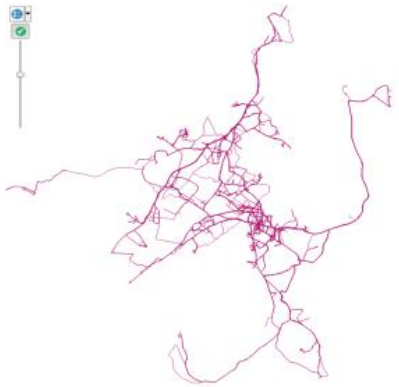


Fig. 11. Common trajectories

Fig. 12. Outliers extracted

TABLE. IV. TRUCKS OUTLIERS EXTRACTED

Nominees	Expected outliers	Spatiotemporal outliers	Common trajectories	Temporal outliers
35750	9402	1157	26348	421

TABLE. V. SEMANTIC SPATIOTEMPORAL OUTLIERS TRUCKS TRAJECTORIES

Spatiotemporal outliers					
Stop	***Emergency***	***Regular***			
512	14	631			
		Blocked route	***Avoided route***	***Traffic Jam***	***Others***
		14	345	225	47

TABLE. VI. SEMANTIC TEMPORAL OUTLIERS FROM TRUCKS TRAJECTORIES

Temporal outliers	
Stop	***Emergency***
387	43

The experimental results for trucks outliers show that the trajectories contain 1157 spatiotemporal outliers from 9402 expected outliers, and contain 421 temporal outliers from 26348 common trajectories, the spatiotemporal outliers contain 512 stops, 14 emergency outliers, and 631 as regular outliers, in which we have 14 blocked route outliers, 345 avoided route outliers, 223 traffic jam outliers, and 47 other outliers none defined.

VII. CONCLUSION

Several algorithms have been proposed for trajectory data mining, but only a few consider semantics, and very few of them deal with semantics on trajectory outlier detection. In this paper, we gave importance to outliers extracted from semantic trajectories, for that we have proposed a conceptual approach that consist to build trajectories from GPS points, enrich them with semantic data, then apply mining algorithm to detect semantic outliers from moving objects, the algorithm shown in this experiment discovers the populated route that the majority of trajectories followed, then detect all other deviations that trajectories can follow to reach the same place, after that the algorithm divided the results to spatiotemporal outliers and just temporal outliers. The spatiotemporal outliers are extracted

from spatial outliers, and they contain stop outliers, emergency outliers, and regular outliers in which three types are discussed; blocked route outliers, avoided route outliers and traffic jam outliers. The temporal outliers contain stops and emergency outliers that can exist in the populated route. The next step will be the introduction of the direction of outliers extracted, and the introduction of mode of transportation to distinguish the types of moving objects that can use the routes [30, 31], giving more details about results, and studying the parameters of the algorithm.

ACKNOWLEDGMENT

I would like to thank my two supervisors; professor Said Raghay and professor Sala El hadaj for their commitments and guidelines as well as their mental support during the preparation of this work, also I would like to present my gratitude to authors of [9,8] for their indirectly contribution of the realization and the implementation of this work.

REFERENCES

[1] Fernando de Lucca Siqueira and Vania Bogorny, "Discovering chasing behavior in moving object trajectories", T. GIS, vol. 15, no. 5, pp. 667–688, 2011.

[2] Monica Wachowicz, Rebecca Ong, Chiara Renso, and Mirco Nanni, "Finding moving flock patterns among pedestrians through collective coherence", International Journal of Geographical Information Science, vol. 25, no. 11, pp. 1849–1864, 2011.

[3] Luis Ot´avio Alvares, Alisson Moscato Loy, Chiara Renso, and Vania Bogorny, "An algorithm to identify avoidance behavior in moving object trajectories", J. Braz. Comp. Soc., vol. 17, no. 3, pp. 193–203, 2011.

[4] Wei Liu, Yu Zheng, Sanjay Chawla, Jing Yuan & Xing Xie, (2011) "Discovering SpatioTemporal Causal Interactions in Traffic Data Streams", KDD'11, August 21 -24, San Diego, California, USA.

[5] Elizabeth Wu, Wei Liu & Sanjay Chawla, (2010) "Spatio-temporal outlier Detection in Precipitation Data", Knowledge discovery from sensor data, Volume 5840, pp 115-133.

[6] Alepsia Albanese, Sankar K Pal & Alfredo Petrosino, (2011) "A rough set approach to spatiotemporal outlier detection", Proceedings of 9th international conference on Fuzzy logic and applications, Springer-verilog, LNCS Volume 6857, pp 67-74.

[7] Gupta, M., Gao, J., Aggarwal, C. C., and Han, J. (2013). Outlier detection for temporal data: A survey. TKDE, 25.

[8] Fontes, V. C., de Alencar, L. A., Renso, C., and Bogorny, V. (2013). Discovering trajectory outliers between regions of interest. In GeoInfo.

[9] AR Aquino. Alvares L. O., Renso C. and Bogorny V .Towards Semantic Trajectory Outlier Detection. GeoInfo. (2013).

[10] Fontes, V. C, Bogorny V. (2013). Discovering Semantic Spatial and Spatio-Temporal Outliers from Moving Object Trajectories.CoRR, abs/1303.5132.

[11] Huang, H. Anomalous behavior detection in single-trajectory data. International Journal of Geographical Information Science, 29 (12), 2015 .

[12] Lettich, F., et al . Detecting avoidance behaviors between moving object trajectories. Data & Knowledge Engineering . 2016 .

[13] Li, X., et al . . A method of abnormal pedestrian behavior detection based on the trajectory model. In : ICTE 2013 : proceedings of the fourth international conference on transportation engineering, , 2013 .

[14] Shen, M., Liu, D.-R., and Shann, S.-H. Outlier detection from vehicle trajectories to discover roaming events.Information Sciences, 2015 .

[15] Zhang,L.,Hu,Z.,andYang,G. Trajectory outlier detection based on multi-factors. IEICE TRANSACTIONS on Information and Systems, 2014.

[16] Sana Chakri, Said Raghay and Salah El Hadaj. Modeling, Mining, and Analyzing Semantic Trajectories: The Process to Extract Meaningful Behaviors of Moving Objects. International Journal of Computer Applications 124(8):15-21, August 2015. Published by Foundation of Computer Science (FCS), NY, USA.

[17] Lee, J.-G., Han, J., and Li, X. (2008). Trajectory outlier detection: A partition-and-detect framework. In ICDE, pages 140–149. IEEE.

[18] Mateus Barragana, Luis Otavio Alvares & Vania Bogorny, Unusual behavior detection and object ranking from movement trajectories in target regions, International Journal of Geographical Information Science, 2016.

[19] Zheng, Y., Zhou, X.: Computing with spatial trajectories. Springer (2011).

[20] de Graaff, V., de By, R.A., van Keulen, M., Flokstra, J.: Point of interest to region of interest conversion. In: Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (SIGSPATIAL GIS 2013), Orlando, FL, USA, (New York), pp. 378–381. ACM, November 2013.

[21] Parent, C., Spaccapietra, S., Renso, C., Andrienko, G., Andrienko, N., Bogorny, V., Damiani, M.L., Gkoulalas-Divanis, A., Macedo, J., Pelekis, N., et al.: Semantic trajectories modeling and analysis. ACM Computing Surveys (CSUR) 45(4), 42 (2013).

[22] Alvares L, Palma A, Oliveira G, Bogorny V Weka-STPM: from trajectory samples to semantic trajectories. In: Proceedings of the Workshop Open Source Code.2010.

[23] Nanni, M., Pedreschi, D. 2006. Time-focused clustering of trajectories of moving objects. Journal of Intelligent Information Systems 27(3) (2006) 267–289.

[24] C. Panagiotakis, N. Pelekis, I. Kopanakis, E. Ramasso, Y. Theodoridis, "Segmentation and Sampling of Moving Object Trajectories based on Representativeness", IEEE Transactions on Knowledge and Data Engineering, 07 Feb. 2011. IEEE computer Society Digital Library. IEEE Computer Society.

[25] N. Pelekis, I. Kopanakis, E. Kotsifakos, E. Frentzos and Y. Theodoridis. "Clustering Uncertain Trajectories", Knowledge and Information Systems (KAIS), DOI 10.1007/s10115-010-0316-x, 2010.

[26] E. Frentzos, K. Gratsias, N. Pelekis and Y. Theodoridis. "Algorithms for Nearest Neighbor Search on Moving Object Trajectories". Geoinformatica, 11:159–193, 2007.

[27] N. Pelekis, I. Kopanakis, E. Kotsifakos, E. Frentzos and Y. Theodoridis. "Clustering Trajectories of Moving Objects in an Uncertain World", In the Proceedings of the IEEE International Conference on Data Mining (ICDM'09), Miami, U.S.A., 2009.

[28] O. Abul, F. Bonchi, M. Nanni, Never Walk Alone: uncertainty for anonymity in moving objects databases, in: Proceedings of the 24nd IEEE International Conference on Data Engineering (ICDE'08), 2008.

[29] Lee, J.-G., Han, J., and Li, X. (2008). Trajectory outlier detection: A partition-and-detect framework. In ICDE, pages 140–149. IEEE.

[30] Prelipcean, A.C., Gidofalvi, G., and Susilo, Y.O. Measures of transport mode segmentation of trajectories. International Journal of Geographical Information Science, 2016.

[31] Ahmad, M., Karagiorgou, S., Pfoser, D., Wenk, C.: A comparison and evaluation of map construction algorithms using vehicle tracking data. GeoInformatica Journal (2015).

Block Wise Data Hiding with Auxilliary Matrix

Jyoti Bharti
Deptt. of Computer Science & Engg.
MANIT
Bhopal, India

R.K. Pateriya
Deptt. of Computer Science & Engg.
MANIT
Bhopal, India

Sanyam Shukla
Deptt. of Computer Science & Engg.
MANIT
Bhopal, India

***Abstract*—This paper introduces a novel method based on auxiliary matrix to hide a text data in an RGB plane. To hide the data in RGB planes of image via scanning, encryption and decryption. To enhance the security, the scanning technique combines two different traversals – spiral and snake traversal. The encryption algorithm involves auxiliary matrix as a payload and consider the least significant bits of three planes. To embed the text message would in the form of ASCII values which are similar to the red plane values and least significant value of pixels in blue plane marks the position of pixels. The least significant bit of boundary values of green-plane signifies the message. These three planes are recombined to form the stego-image, to decrypt the message with the help of scanning in the red-plane and blue plane and green plane simultaneously. Performance evaluation is done using PSNR, MSE and entropy calculation and generated results are compared with some earlier proposed work to present its efficiency with respect to others.**

Keywords—Steganography; RGB planes; Scanning; Stego-image; ASCII value

I. INTRODUCTION

Steganography is the process to conceal a message or data in an image which is not detectable by human visual system. Message would be in the form of text, image, audio etc. Unlike cryptography transform the message into another form and hide in an image and then passed over the attack prone network to the receiver; it is more secure, as the existence of the message embedded in the image is concealed [1]. In this paper, a new technique is proposed to hide text message in planes of RGB image, so as to enhance the security of the information being hidden in the image. The accuracy has been evaluated on comparison of MSE and PSNR values. Some of the most popular techniques that have already been discussed in this field in the past years are adaptive data hiding in edge areas of images with spatial lsb domain systems [2], reversible data hiding using integer wavelet transform and companding technique[3], robust image-adaptive data hiding using erasure and error correction [4], reverse data hiding [5] and many more. Steganography is very useful and commercially important application in the digital world for example digital watermarking. In this application, to ensure the integrity or authenticity of intellectual property or product, owner can embed the message hidden in the file. This kind of mechanism is used by intelligence agencies for secret works [6].

II. PROPOSED WORK

In this paper the proposed methodology consists of three Steps- Scanning, Encryption and Decryption.

A. Scanning

Scanning means, In a two dimension array, the way or pattern in which each element or pixel is accessed. As a purpose of security, a hybrid scanning techniques has been used which is based on spiral and snake traversal. The carrier image is divided into smaller size of blocks. Each block contain 50 x 50 pixels. The blocks are accessed in a snake pattern as shown in Fig. 1(a).

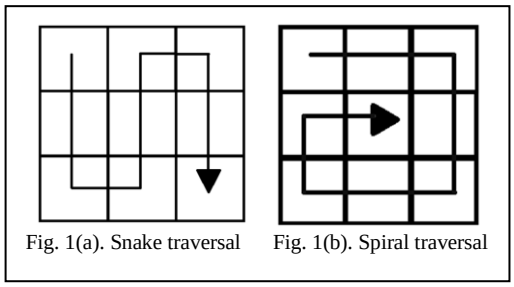


Fig. 1(a). Snake traversal Fig. 1(b). Spiral traversal

Fig. 1. Scanning Traversal

In the snake pattern, starting from the first block, blocks are accessed vertically downwards then accessing the adjacent block then moving vertically upwards. This pattern continues until all the blocks are traversed. Then within each block, pixels are accessed using the spiral technique as shown in Fig. 1(b). In the spiral technique, pixels are accessed starting from the first pixel, moving along the boundary towards the center. Once all the pixels within a block are accessed the technique again initializes the accessing pointer to the first pixel of next block to be accessed in the snake pattern which is shown in Fig. 2

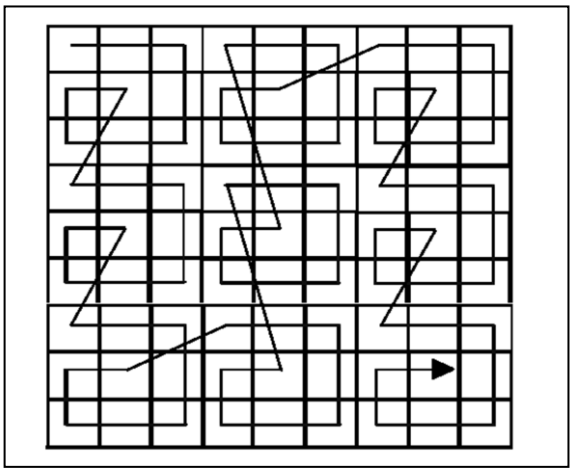


Fig. 2. Block wise path traversal of pixels in an image

B. Encryption

Encryption is the technique of hiding the text message in the carrier image. For this method RGB image is taken as the carrier image. The text message can be in any language. Here, English language is taken for the text message. The ASCII values can be mapped with the pixel intensities of RGB plane as ASCII values of English alphabets i.e. 0-127 lie within the range of pixel intensities i.e. 0-255. RGB image consists of three planes – red plane, green plane and blue plane, each playing a specific role in the proposed method that is described later in the paper. Pixels in the plane comprises of 8 bits which shows the intensity values within the range of 0 to 255.

In this methodology, RGB image is divided into three planes, namely: Red plane, Blue plane and Green plane. The LSB planes of blue and green plane are extracted using bit plane slicing. Along with this, an auxiliary matrix with size equal to the size of the image is maintained with all values set to 0. Auxiliary matrix stores the positions/ indices of the letters in the message. For example if the message is "HELP" then index of H is 1, E is 2, L is 3 and P is 4. To hide the message in RGB plane, it requires four steps.

1) Convert each letter of the message into its ASCII values as shown in Fig 3.

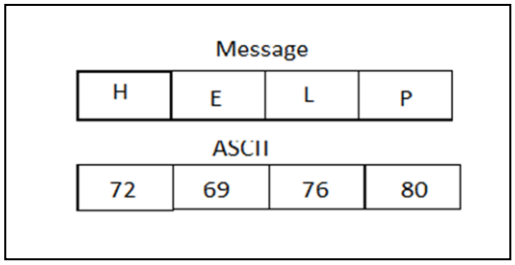


Fig. 3. Payload: The message to hide

2) The red plane is scanned using the proposed scanning technique (Originally the image size will be large enough to implement the proposed scanning technique, but for demonstration, here an image of size 5X5 pixel is taken, which smaller than 50X50, then spiral technique is applied. In the red plane, ASCII value of each letter in the message is compared with the pixel intensities of the red plane. If any pixel intensity in red plane is matched with ASCII value, then, its position is marked in LSB of blue plane using the method in step 3. If no such pixel intensity is found then, closest pixel intensity to the ASCII value is searched in the red plane and it is replaced with the ASCII value of the letter being searched. The position of this modified pixel is also marked in blue plane using the method in step 3.

3) Least significant bits of the blue plane act as an indicator plane and that signifies that the red plane contains the message. LSB of blue plane is set to 0 indicating no modification in pixel intensity or pixel intensity is not equal to ASCII value of message. After scanning the red plane, if the pixel intensity matches the ASCII value of the letters in the message or any closest pixel intensity is replaced by the ASCII value of the letter then the corresponding pixel in blue-plane is marked by setting the LSB of that pixel, i.e. LSB 0 is turned to 1.

Step 2 and 3 are repeated for each letter in the text message. Hence, all the ASCII values of the letters in the message will be available in the red plane. It is shown in Fig 4.

Red plane of carrier image

71	68	50	69	52
59	61	63	89	42
72	41	73	88	59
41	102	116	99	80
77	76	84	58	79

LSB of Blue Plane after Scanning

0	0	0	1	0
0	0	0	0	0
1	0	0	0	0
0	0	0	0	1
0	1	0	0	0

Fig. 4. ASCII value Comparison in red plane and converting the corresponding LSB of blue plane as 1

Auxiliary matrix is traversed simultaneously with red plane. When the ASCII value is found or nearest ASCII value is found then the index of that letter is set in auxiliary matrix (in the same position as in the LSB of blue plane). So auxiliary matrix holds the indices of the letters in the message in the exact position where the LSB of blue plane is set to 1, this is shown in the Fig. 5.

The auxiliary matrix is traversed with the scanning technique proposed earlier so as to get a jagged sequence of indices. These indices are converted into their binary forms as shown in Fig. 6.

Auxillary array showing indices

0	0	0	2	0
0	0	0	0	0
1	0	0	0	0
0	0	0	0	4
0	3	0	0	0

Fig. 5. Auxiliary matrix to hide the indices of payload

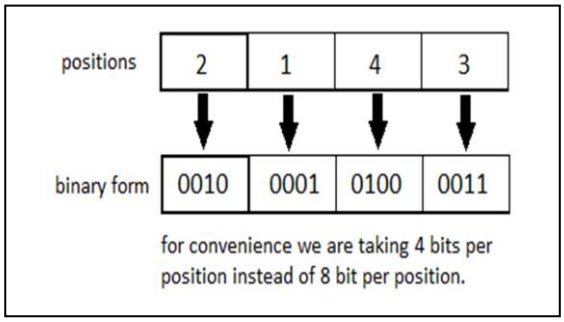


Fig. 6. Positions of indices of payload retrieved after scanning the auxiliary matrix and convert it into an 4 bit binary number

4) The LSB of green plane is used to hide the indices of letters in the message in their binary forms. The indices obtained after scanning the auxiliary matrix, are converted into their 8 bit binary format as shown in Fig 8 and substituted in the LSB of green plane at its boundary as shown in Fig 7. The whole LSB substitution is only done on the boundary values of the plane ensuring least modification in the LSB of green plane. These indices are hidden contiguously in the boundary of green plane at LSB positions.

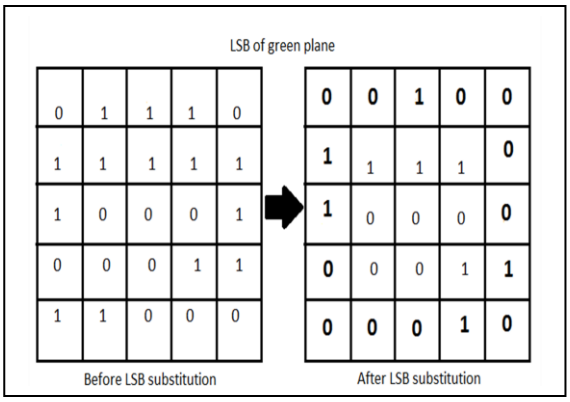


Fig. 7. LSB Substitution in Green plane

LSB of blue plane and LSB of green plane are merged with the higher bit blue plane and green plane respectively. To produce the Stego-image all the three planes are merged together. At the receiver end, the message retrieved from Stego-image using the Decryption technique.

C. Decryption

Decryption is a technique to decipher the information hidden in the Stego- image. Since the message is confidential, it is assumed that, the technique is known to the concerned sender and receiver only. In order to retrieve the message, the Stego-image is again divided into three planes-red plane, green plane and blue plane. The LSB of blue plane and green plane is extracted to retrieve the ASCII values of the letters in the message using red plane and to extract indices of those ASCII values respectively. Using the scanning technique proposed above, the red plane and the LSB of the blue plane is searched simultaneously, for the hidden letters. In case, a pixel intensity is found in the red plane such that its corresponding LSB of the blue plane is 1, then that intensity value (belonging to red plane) is stored in an array. Similarly the whole red plane and LSB of blue plane is searched and such values are extracted. These values are the ASCII values of letters in the message. These values are stored in an array in the same order as extracted from red plane. Hence the array contains a random sequence of ASCII values. To retrieve the message from Stego-image, these values are arranged in the same order as they were present in the original message. For this, the indices of the letters, which are hidden in green plane are extracted. The LSB of boundary pixels in green plane are processed, eight pixels at a time and an 8 bit binary number is formed. These binary numbers are converted into their decimal forms. These decimal numbers are the indices of the random ASCII values retrieved from the red plane earlier. These are stored sequentially into another array. Following the same procedure, indices for all the intensity values are extracted. After this, the intensity values are rearranged according to their indices. These intensity values are converted into their corresponding ASCII characters. The decrypted string of characters forms the original message hidden in the Stego-image.

III. ALGORITHM

A. Encryption technique

- Load an RGB Image.
- Extract red, green and blue plane. Store it in matrix red, matrix blue, matrix green.
- Extract LSB of green and blue planes and store it in matrix green1, matrix blue1 respectively.
- Extract size of image in variables rows, cols.
- Convert LSB of blue plane to zeros so that values in matrix blue1 are 0.
- Create an auxiliary matrix aux with all values 0.
- Input message from console and store in message array.
- Traverse the message from beginning to end one letter at a time
- Store letter in variable k;
- Flag stores the search result after scanning. Flag is set to 1 if ASCII value is found else set to 0.
- Call function of Scanning_technique and pass variables:- flag, pos_x and pos_y.
- pos_x and pos_y stores position where ASCII values are found or intensity values are the closest.
- If ASCII value matches pixel intensity value of red plane then set the corresponding position of the blue1 plane to 1.
- Store the position of the character in the auxiliary matrix.
- If the ASCII value does not match pixel intensity value then find the pixel intensity value closest to the ASCII value.

- Replace pixel intensity value with the new intensity value and set the corresponding location in the blue1 plane to 1 and then store the position of the character in the auxiliary matrix.
- Create an array- bit_arr.
- Scan auxiliary matrix using the proposed scanning technique.
- If a non-zero element is found then store the element in array in binary form.
- Store the bit array in the boundary values of green1 plane.
- Merge green1 and green plane.
- Merge blue1 and blue plane.
- Merge red, blue and green to get the StegoImage.

B. Scanning technique

- Divide the image into 50 X 50 blocks.
- Snake technique: In the snake technique, the image matrix is traversed block wise. From the first block, move vertically downwards until all the blocks are traversed in a column and then the adjacent block are traversed and move vertically upwards.
- Spiral technique: In the spiral technique, a 50 X 50 block is traversed starting from the first pixel and moving towards the boundary and moving inside towards the center pixel.

C. Decryption technique

- Load Stego-Image
- Extract red, green and blue plane.
- Extract size of the image.
- Initialize message array, position array and bit_arr array to store the ASCII value of the message, position in decimal and position in binary respectively.
- Call Scanning technique for red and blue planes.
- Consider the LSB of green plane and traverse the boundary values and 8 pixels at a time.
- Store 8 LSB values in bit_arr array.
- Convert bit_arr array into decimal and store into position array.
- Arrange message array according to the position array.
- Message array is our original message.

IV. Result And Analysis

The security analysis compares the Original image with the Stego-image based on the histogram of the images. If the change in histogram is minimal, then the encryption algorithm is considered secure. Fig.8(a)& (b) shows the size of original image size 200 x 450 pixels and Stego-image created after embedding the message using the above proposed technique.

The modified image (Stego-image) after applying the proposed algorithm does not release any identifiable visual difference. The histograms of the original and stego images are shown in Fig.9. Both the histograms show no such significant changes.

The experimental results obtained are subjected to various statistical techniques, to evaluate the performance parameters of the steganographic images viz., (i) PSNR values of the Stego-image (ii) Mean Square Error (iii) Entropy.

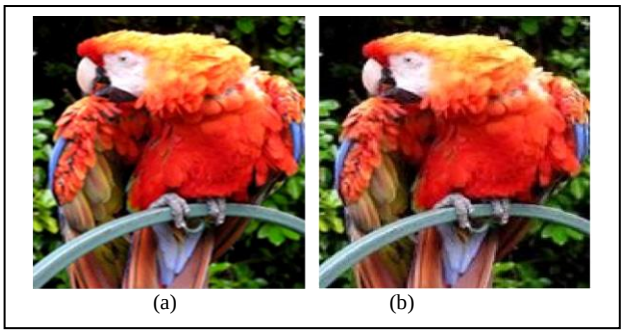


Fig. 8. (a). Original Image (b). Stego-image

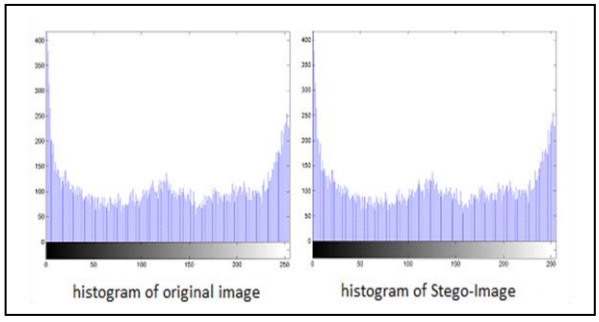


Fig. 9. Histogram original and stego-image

PSNR as a metric computes the peak signal-to-noise ratio, in decibels, between two images [7]. The higher the PSNRvalue, the better degraded image has been reconstructed. To match the original image and the stego-image calculate the PSNR value using (1) [8].

$$\text{PSNR} = 10 \times \log_{10}\left(\frac{R^2}{MSE}\right) \quad (1)$$

Where, R is the maximum pixel intensity value for an image.

The MSE represents the average of the squares of the "errors" between our actual image and our noisy image. The error is the amount by which the values of the original image differ from the degraded image [8] . It is given in (2) [9].

$$MSE = \left(\frac{1}{(m \times n)}\right) \times \sum_i \left(\sum_j (f - g)^2\right) \quad (2)$$

where, f: matrix data of our original image. g: matrix data of our degraded image. m: number of rows of pixels of the images . i: index of that row. n: number of columns of pixels of the image. j: represents the index of that column.

The more data hidden in a file, the higher that file's entropy. That is, if bits are too disorderly and data are too random, steganography may be suspected [10]. Entropy is defined as:

$$E = -\sum P \times \log(P) \quad (3)$$

For the above case of parrot image PSNR value comes out to be 52.65, MSE comes out to be 0.35 and entropy value comes out to be 0.19. The simulation done with other images viz. Lena, Baboon, Pepper and Butterfly. The results of the simulation for these images, the histogram analysis and the results based on the quality metrics (i.e. values of the PSNR, MSE and Entropy) are tabulated in table-I and shown in the following Fig.10-13.

TABLE I. PSNR, MSE AND ENTROPY OF IMAGES OF THE PROPOSED ALGORITHM.

Figure no. and name	PSNR	MSE	Entropy
Lena	52.19	0.39	0.00
Baboon	52.44	0.37	0.00
Pepper	53.50	0.29	0.11
Butterfly	52.99	0.33	0.66

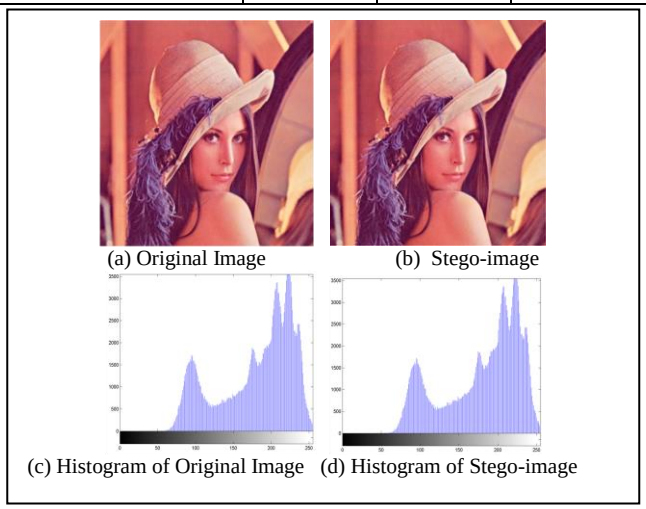

(a) Original Image (b) Stego-image
(c) Histogram of Original Image (d) Histogram of Stego-image

Fig. 10. Result for Lena

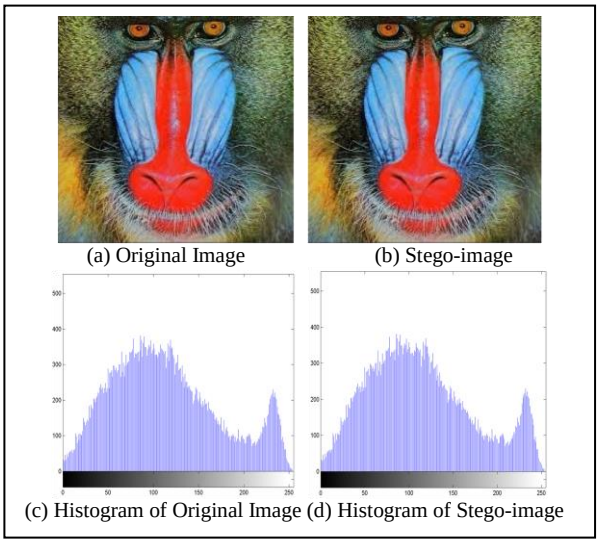

(a) Original Image (b) Stego-image
(c) Histogram of Original Image (d) Histogram of Stego-image

Fig. 11. Results for Baboon

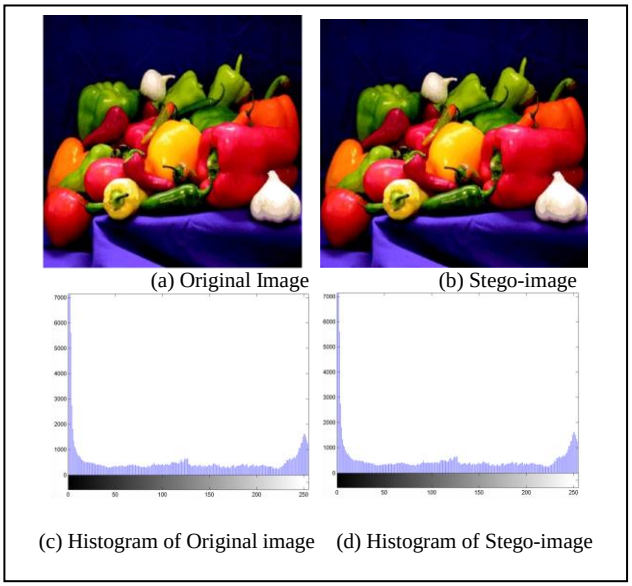

(a) Original Image (b) Stego-image
(c) Histogram of Original image (d) Histogram of Stego-image

Fig. 12. Result for Pepper

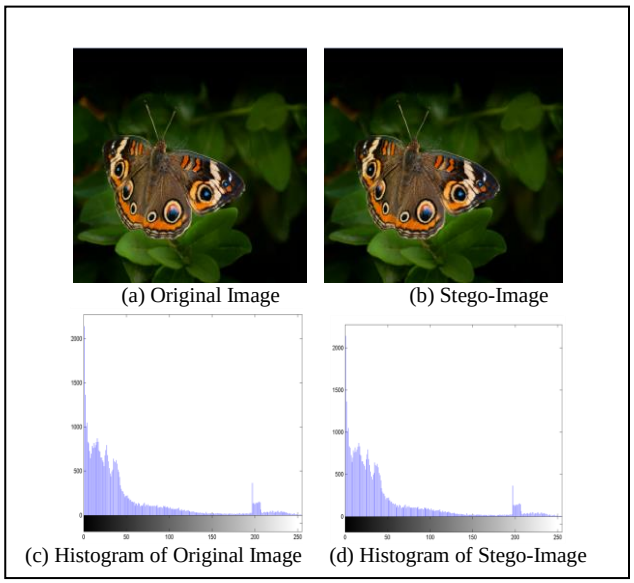


Fig. 13. Result for Butterfly

A. *Effect of message length*

Increase in the length of the message causes, manipulation of more number of pixel intensities. This results in decrease in PSNR values with increase in message length. The probability of error within a region increases with more number of modifications in pixel intensities, hence the value of MSE with increase of message length increases. The messages are embed in to a parrot image The change in PSNR and MSE values are shown in fig. 14 & 15.

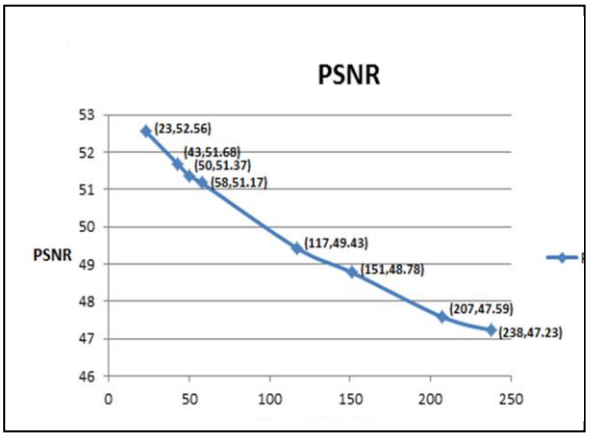


Fig. 14. Change in PSNR values with increase in message length for image 'Parrot'

Fig.14 shows the PSNR values for image Parrot. PSNR values lie on the Y-axis and message length on X-axis. For message length ranging from 25 to 250, PSNR values for the proposed method, lie within 47 to 53. The PSNR value decreases on enhancing size of the message lenght.

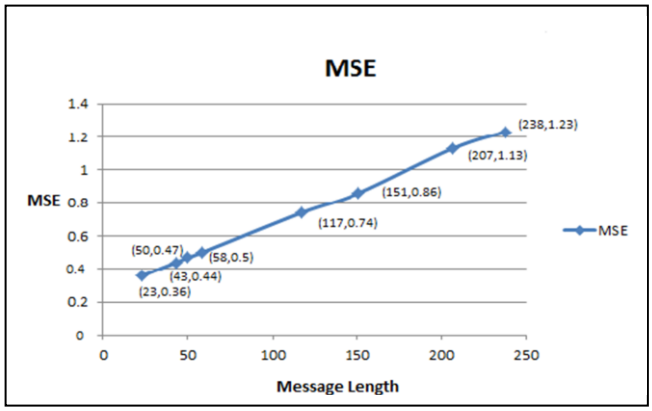


Fig. 15. Graph of change in MSE values with increase in message length for image 'parrot'

Fig. 15 MSE values lie on the Y-axis and message length on X-axis. For message length ranging from 25 to 250, PSNR values for the proposed method, lie within 0.2 to 1.3 for the image "parrot".

B. *Comparitive Study*

Results obtained for the proposed method are compared with the results of some Steganographic techniques proposed earlier in well-known research papers and journals, on the basis of quality metrics MSE and PSNR values for different images (based on the availability of the results). It is observed, the proposed algorithm provides better results than all the previous schemes. The results of comparison for the MSE and PSNR values of images- "Lena" and "Baboon" are shown in Table II-IV.

TABLE II. COMPARISON OF MSE VALUES FOR LENA

S.No	Steganography Technique	MSE
1.	Adaptive Data Hiding in Edge Areas of Images With Spatial LSB Domain Systems [2]	7.337
2.	OLS Technique [10]	2.34
3.	OLSGA Technique [10]	2.34
4.	Proposed method	0.39

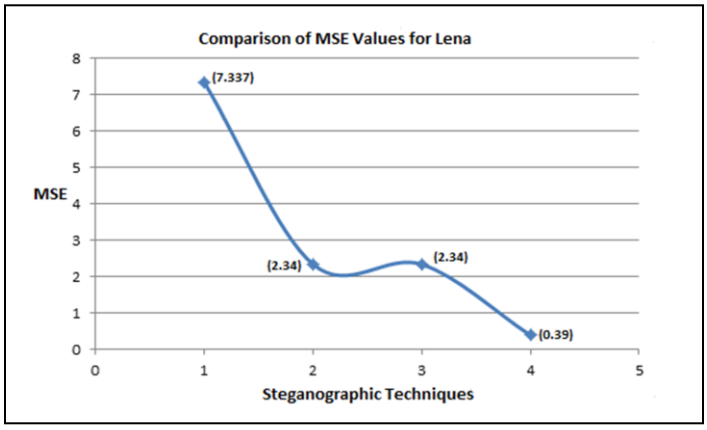


Fig. 16. Graph for MSE values for Lena

TABLE III. COMPARISON OF PSNR VALUES FOR LENA

S.No	Steganography Technique	PSNR
1	Adaptive Data Hiding in Edge Areas of Images With Spatial LSB Domain Systems [2]	37.61
2	Robust Image-Adaptive Data Hiding Using Erasure and Error Correction [4]	41.43
3	Reversible Data Hiding using integer wavelet transform and campanding technique. [3]	46.23
4	A Variable Depth LSB Data Hiding Technique (3k message length) [11]	46.35
5	Reversible data hiding [5]	48.20
6	A DWT based approach for image steganography[12]	50.8021
7	Proposed method	52.19

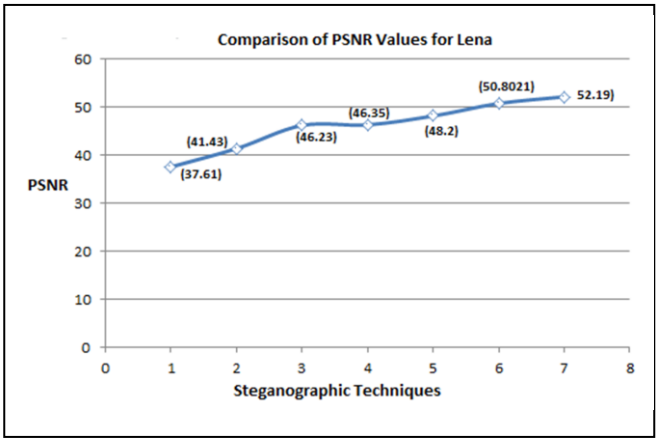


Fig. 17. Graph for PSNR values for Lena

The Fig.16 is a graphical representation of the Table.II. which presents comparative study between different Steganography techniques proposed earlier and the proposed method here, on the basis of MSE generated on the image "Lena". For the graph X- axis represents serial numbers of methods mentioned in Table.II and Y- axis represents their corresponding MSE values. It can be observed, the proposed method produces the least MSE value for "Lena" compared to rest.

The Fig.17 is the graphical representation of the Table III, on the basis of PSNR generated on the image "Lena". For the graph X- axis represents serial numbers of methods mentioned in Table III and Y- axis represents their corresponding PSNR values. It can be observed, the proposed method produces the comparatively higher PSNR value i.e 52.19 for "Lena

TABLE IV. COMPARISON OF PSNR VALUES FOR BABOON

Steganography Technique	PSNR
1. Adaptive Data Hiding in Edge Areas of Images With Spatial LSB Domain Systems [2]	34.26
2. Robust Image-Adaptive Data Hiding Using Erasure and Error Correction [4]	35.98
3. Reversible Data Hiding using integer wavelet transform and campanding technique. [3]	39.66
4. Reverse data hiding [5]	48.2
5. Proposed method	**52.44**

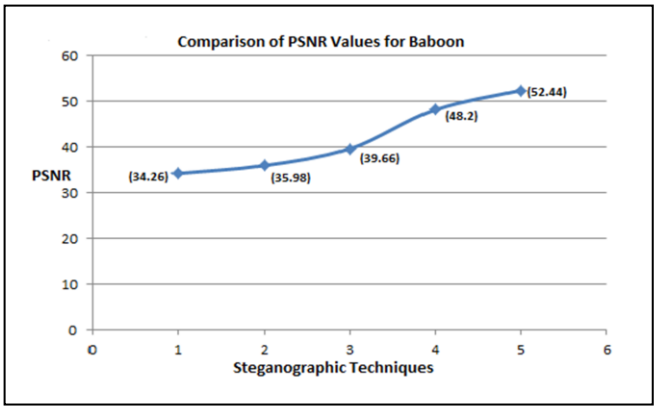


Fig. 18. Graph for PSNR values for Baboon

The Fig.18 is the graphical representation of comparison PSNR with other. For the graph X- axis represents serial numbers of methods mentioned in Table IV and Y- axis represents their corresponding PSNR values. It can be observed, the proposed method produces the comparatively higher PSNR value i.e 52.44 for "Baboon".

V. CONCLUSION

In this paper, the proposed is a new steganography technique to hide a text message in an RGB image with minimum manipulation with the intensity values and LSB of the pixel. Instead of hiding the ASCII value of the letter, which would have taken 8 bits, it tries to search the pixel in the red plane whose intensity value matches the ASCII value of each letter of the message and changes the LSB value of the corresponding pixel in the blue plane resulting in a change of only one bit of the pixel. This results in less modification of bits resulting in less randomness in the image. As per the experimental results shown on different images, it is found that the PSNR value ranges between 50 and 55 which is near to ideal, the entropy values are closer to 0.0 and the MSE values are less. These experimental analysis shows that after embedding the data, less distortion in the stego image is not noticeable as the histogram of the cover image and stego image are very similar which accounts for better stego image quality. Comparing with other techniques as well, it is found that this proposed technique gives better results for PSNR and MSE values.

REFERENCES

[1] S.Ashwin, S.Aravind Kumar, J.Ramesh, K.Gunavathi, "Novel and secure encoding hiding techniques using image steganography : A survey", IEEE International Conference on Emerging Trends in Electrical Engineering and Energy Management (ICETEEEM-2012), pp. 171-177, December 2012.

[2] [Cheng-Hsing Yang, Chi-Yao Weng, Shiuh-Jeng Wang, "Adaptive data hiding in edge areas of images with spatial lsb domain systems", IEEE Transactions On Information forensics And Security, vol. 3, pp. 488-497, September 2008.

[3] GuorongXuan, Chengyun Yang, Yizhan Zhen,Yun Q. Shi, and Zhicheng Ni, "Reversible data hiding using integer wavelet transform andcampanding technique", Third International Workshop(IWDW-2004), Springer Verlag Berlin Heidelberg, pp. 115 – 124, 2005.

[4] Kaushal Solanki, Noah Jacobse,Upamanyu Madhow, Shivkumar Chandrasekaran,"Robust image-adaptive data hiding using erasure and error correction", IEEE Transactions On Image Processing, vol. 13, pp. 1627-1639, December 2004.

[5] Zhicheng Ni, Yun-Qing Shi, Nirwan Ansari,and Wei Su, "Reversible data hiding", IEEE transactions on circuits and systems for video technology, vol. 16, No. 3, March 2006.

[6] [6]. Gary C. Kessler, "An overview of steganography for the computer forensics examiner", Issue of Forensic Science and Communication, Vol 6 - Number 3, February 2004 [updated February 2015]

[7] Gabriel Macharia Kamua, Stephen Kimani,Waweru Mwangi, "An enhanced least significant bit steganographic method for information hiding", Journal of Information Engineering and Applications, ISSN 2224-5782 (print) ISSN 2225-0506 (online), vol 2, No.9, 2012.

[8] Z.Wai, S. Than, "Data hiding techniques depended on pseudorandom sequences",International Journal of Scientific Engineering and Research, vol. 1, Issue 2, October 2013.

[9] Sudipta Kr Ghosal, "A new pair wise bit based data hiding approach on 24 bit color image using steganographic technique", Proceedings of IEMCON-2011,, Kolkata, West Bengal, India, pp.123–129, January 2011

[10] Xiaoyan Qiao, "A new method of steganalysis based on image entropy", Springer Verlag Berlin Heidelberg, CCIS 2, pp. 810-815, 2007.

[11] Smo-hui liv, Tun-hang chen, Hongxunyao, Wen gao, "A variable depthlsb data hiding technique in images", Proceedings of the third Intemational conference on Machine Learning and Cybernetics, Shanghai, pp. 26-29, August 2004.

[12] Po-Yueh Chen and Hung-Ju Lin, "ADW based approach for image steganography", International Journal of Applied Science and Engineering, vol. 4, pp. 275-290, 2006.

Detection of Edges using Two-Way Nested Design

Asim ur Rehman Khan, Syed Muhammad Atif Saleem, Haider Mehdi
Multimedia Lab,
National University of Computer and Emerging Sciences (NUCES),
Pakistan

***Abstract*—This paper implements a novel approach of identifying edges in images using a two-way nested design. The test comprises of two steps. First step is based on an F-test. The sums of square (SS) of various effects are used to extract the mean square (MS) effect of respective effects and the unknown effect considered as noise. The mean square value has a chi-square distribution. The ratio of two chi-square distributions has an F-distribution. The final decision is based on testing a hypothesis for the presence or absence of an effect. The second step is based on contrast function (CF). This test identifies the presence or absence of an edge in four directions that are horizontal, vertical, and the two diagonal directions. The test is based on Tukey's T-test. The performance of nested design is compared with the edge detection using Sobel filter. A rigorous testing reveals that the nested design yields comparable results for images that are either free of noise or corrupted with light noise. The nested design however outperforms the Sobel filter in situations where the images are corrupted with heavy noise.**

***Keywords*—Analysis of variance (ANOVA); Edge detection; F-test; nested design; T-test**

I. Introduction

The detection of edges, in a digital image, has several industrial, biological, medical, scientific, and other real life applications. In a recent paper, the tracking of wild life has been performed by detecting the edges of animals and then keeping their record in a database [1]. The FPGA has enabled us implement advanced algorithms that were previously considered impossible due to their longer processing time. Several fast real time edge detection schemes have been demonstrated in [2]-[3]. An algorithm of image segmentation using genetic algorithm (GA) has been proposed in [4]. A wavelet transform based technique for SAR (synthetic aperture radar) images is given in [5]. Several other wavelet transform based solutions are given in [6]-[8]. The nonlinear techniques generally outperform linear filters for edge detection. A comparison of several nonlinear techniques, like order statistics filters, hybrid filters, neural filters, and bilateral filters is made in [9]. Various statistical approaches for edge detection are demonstrated in [10]-[12]. A Kalman-based edge detection scheme is demonstrated in [13]. A few advance gradient based edge detection techniques are Marr-Hildreth, and Canny edge detectors [14]. The edge detection using cellular neural network (CNN) is given in [15]. A combination of ant colony optimization (ACO) and wavelet transform based edge detection technique is given in [16]. The linear vector quantization for edge detection has been demonstrated in [17]-[18].

There are generally two distinct approaches followed in digital image processing. The first approach is by using gradient analysis, and the second approach is by using some kind of transform. The gradient analysis identifies an edge with a significant change in pixel value. Some of the earlier gradient operators are Roberts, Prewitt, and Sobel filters [19]-[21]. The transform based approach uses discrete cosine transform (DCT), or wavelet transform [21]. A significant advantage of the gradient type approach is that their results are based on the local pixel analysis. The wavelet transform considers local effects to some extent, but still the fine details are lost. The second transform technique like DCT completely ignores the local details. All the above approaches fail in case the given image is corrupted with heavy noise.

The mathematical detail of analysis of variance (ANOVA) is available in standard textbooks of statistics [22]-[23]. The detection of edges by using Graeco-Latin square (GLS) design involves a template of 5x5 pixels, such that the Greek & Latin letters are assigned to each pixel. The presence or absence of an effect in four directions is tested statistically by testing a hypothesis for each of these letters [24]. The contrast function (CF) is also a well-tested statistical approach, where the mean of a set of pixels within the template is statistically compared with the remaining pixels. The approach used Tukey's T-test for testing the hypothesis of an edge that is present at a particular location [25]. The classification of multispectral imaging data is given in [26]. The statistical analysis of moving object detection that is previously corrupted with noise is given in [27]. In this paper, we have used two distinct techniques comprising of two-way nested design (TND), and a contrast function (CF). Both the approaches help in identifying edges in an image that are previously corrupted with significantly higher degree of Gaussian noise. After a brief introduction, the next section discuses two-way nested design. The mathematical details of analysis of variance (ANOVA) are given in section III. This is followed by the mathematical background of contrast function (CF) in section IV. The significant results and their critical analysis are given in section V. Section VI concludes this paper.

II. Two-Way Nested Design

A two-way nested design comprises of two levels A, and B such that the level-B is nested through level-A. In literature, this is mentioned as B(A). Graphically this is represented as in Fig. 1. The two-way nested design is quite appropriate for spatial image analysis, which identifies small homogenous regions with sufficient regional details. The level-A comprises

of i levels where $i = 1, \ldots I$. Level-B comprises of j_i levels, such that $j = 1, \ldots, J_i$. In principle, for each j the numbers of elements can vary for each i. Though, in this particular situation the value j is same for each i. Further, there is nothing in common for various levels of i. Theoretically, the subscript j should be writing as j_i, and the nested factor-B should be written as B_{ij_i}. Instead of this complicated notation, a more friendly notation of B_{ij} is used. The complete analysis is performed on a square mask of 8×8 pixels. The subsequent mask is taken by scanning the raster from left to right and from top to bottom. The algorithm is extremely fast when the mask positions are non-overlapping. However, this results in missing out several edges. The mask locations can be overlapped that identifies more edges, but also results in larger processing time. The analysis of variance (ANOVA) is applied statistically for identifying and marking regions having considerable gray level changes within a mask. The final decision is made by testing a hypothesis. In case there is significant confidence developed by rejecting the Null hypothesis of either the effect-A or effect-B (alternately, accepting the presence of an edge), then a second test comprising of contrast functions (CF) further identifies edges in four directions: vertical, horizontal, 45^o diagonal, and 135^o diagonal. Only one edge in any one direction is allowed. However, edges in multiple directions within a particular mask are possible.

A
↓
B

Fig. 1. Two-way Nested Design (TND)

A. Mask Partition

The partition of mask is given in Fig. 2. The mask of 8×8 pixels is partitioned into four segments each comprising of 4×4 pixels. The subscripts in equation do not represent rows and columns as used in standard images. Instead they represent various regions of a mask. Each of these regions has four rows and four columns. Different regions are represented by $i = 1,..,4$. The top-left region is considered as first in the effect-A. The segments are marked in clockwise direction starting from the top-left 4×4 pixel as the first region.

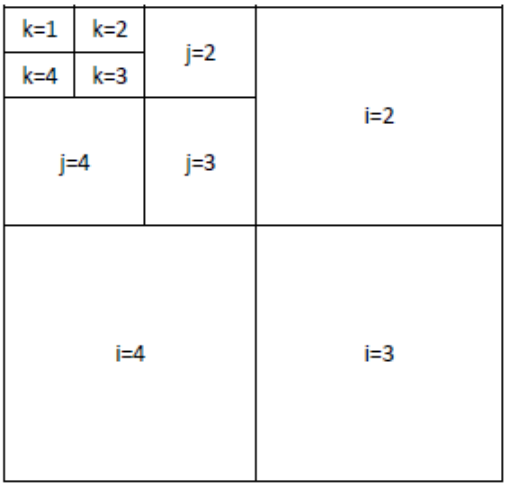


Fig. 2. Mask comprising of 8×8 pixels

Effect-A (subscript i) compares the effect of four regions of a mask each comprising of 4×4 pixels. Each of the four regions in level-A are further divided into four equal size sub-regions each comprising of 2×2 pixels. The sub-regions is represented by j such that $j = 1, \ldots, 4$. The value of j is repeated for each i. It is clear that for different values of i, there is nothing in common for the same j. Each individual pixel in the (2×2) sub-region is represented by $k = 1, \ldots, 4$ again in clockwise direction starting from top-left pixel.

III. The Analysis of Variance (ANOVA)

The gray level change in a large image results in building chipsets that together form interesting features for human and computer analysis. The micro information in the form of pixels is combined to form the macro information in mask comprising of a small set of pixels. The most critical information is, effectively, contained within each pixel. A pixel is designated by y_{ijk} where the subscripts i, j, k correspond to effect-A, effect-B, and an unknown effect considered as noise. All parameters are assumed to have unknown but fixed value with no random value. All randomness is present in the third parameter considered to be a random noise that has Gaussian distribution with zero mean and constant variance. The model is represented by,

$$\Omega: \begin{cases} y_{ijk} = \mu + \alpha_i + \beta_{ij} + \varepsilon_{ijk} \\ \varepsilon_{ijk} \sim N(0, \sigma^2 I) \end{cases} \quad (1)$$

Where μ is the general mean, α_i and β_{ij} are two specific fixed effects with no randomness, and ε_{ijk} represents Gaussian noise of zero mean and independent variance. The assumption of error having Gaussian distribution with zero mean and independent variance results in a simple mathematical model. Fortunately, this assumption holds true in most of the real images. If zero mean condition is violated, then the pixel values can be recalculated by subtracting the mean value from each pixel generating a new image that has zero mean. The mathematical analysis can then be performed on the new image. In some applications, like texture analysis, a nonzero mean and dependence across various observations may in fact help in the image analysis. The zero mean assumption is considered to hold in all the subsequent analysis.

B. The Least Square Estimate

A nested design identifies the least square estimate (LSE) of various parameters. Matrices are used for simplification. A set of observations $\boldsymbol{y} \in \mathbb{R}$ are equal to

$$\boldsymbol{y} = \boldsymbol{X^T \beta} + \boldsymbol{\varepsilon} \quad (2)$$

Where the observation matrix $\mathbf{y}$ is a column matrix $(n \times 1)$. A 2-D image is easily converted into 1-D column matrix by scanning rows horizontally from top-left to the top-right, and then from top-to-bottom of a mask. $\boldsymbol{X^T}$ is the transpose of $\boldsymbol{X}$ which has p rows and n columns $(p \times n)$. $\boldsymbol{X}$ is the transformation matrix comprising of p equations each having n number of parameters. $\boldsymbol{\beta}$ is an unknown parameter matrix of size $(p \times 1)$, and $\boldsymbol{\varepsilon}$ is the error of size $(n \times 1)$. The objective is to find the LS estimate $\widehat{\boldsymbol{\beta}}$ of parameter space $\boldsymbol{\beta}$. The analysis of variance (ANOVA) is very similar to the regression analysis. The only difference is that in regression

analysis, there is no restriction on the elements of $\boldsymbol{X^T}$ as these can be integers or real numbers; whereas, ANOVA requires the elements of $\boldsymbol{X^T}$ to be strictly zero or one. The model essentially assumes that a particular effect is either present or absent. This assumption simplifies the mathematical derivation, and result in an efficient and fast processing. The sum of square of error, ε is

$$\epsilon^2 = (\boldsymbol{y} - \boldsymbol{X^T\beta})^2 \tag{3}$$

By setting $\partial\epsilon^2/\partial\beta$ to zero and then solving for β, the LS estimate of parameter matrix $\hat{\beta}$ is found. The objective is to find their estimated values $\widehat{\boldsymbol{\beta}}$. In case $\boldsymbol{X^T}$ is a square matrix with full rank, then the estimated value $\widehat{\boldsymbol{\beta}}$ is given by,

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X^T})^{-1}\boldsymbol{y} \tag{4}$$

If however $\boldsymbol{X^T}$ is not a square matrix, then this is converted into a square matrix by multiplying both sides by *X* and then solving for estimator matrix $\widehat{\boldsymbol{\beta}}$.

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{XX^T})^{-1}\boldsymbol{Xy} \tag{5}$$

If $\boldsymbol{X^T}$ is not a square matrix or it is not having full rank, then a set of side conditions are added to make it a full rank matrix. The estimates are then found by

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{XX^T} + \boldsymbol{HH^T})^{-1}\boldsymbol{XX^T}\,\boldsymbol{y} \tag{6}$$

The set of side conditions must satisfy,

$$\boldsymbol{H^T}\widehat{\boldsymbol{\beta}} = \boldsymbol{0} \tag{7}$$

The above general mathematical analysis is applied to the nested design comprising of effect-A and effect-B.

C. Sum of Square (SS) of Various Effects

Under the assumption **Ω**, an observation y_{ijk} is approximated by $y_{ijk} = \eta_{ij} + e_{ijk}$, where $\eta_{ij} = \mu + \alpha_i + \beta_{ij}$ is the sum of the general mean μ, and the various effects α_i (level-A), and β_{ij} (level-B). The error is equal to $e_{ijk} = (e_{ijk} - \eta_{ij})$. The subscript 'k' considers the unaccounted for effects and includes all pixels in a mask. The sum of square of error (SSE) is equal to

$$SSE = \sum_{i=1}^{I}\sum_{j=1}^{J_i}\sum_{k=1}^{K_{ij}}\left(y_{ijk} - \eta_{ij}\right)^2 \tag{8}$$

The least square estimate of mean μ is found by differentiating SSE with respect to μ and then setting it equal to zero

$$\frac{\partial(SSE)}{\partial\mu} = -2\sum_i\sum_j\sum_k\left(y_{ijk} - \eta_{ij}\right) = 0 \tag{9}$$

The summation is taken over all possible values of subscripts i, j, and k. By replacing $\eta_{ij} = \mu + \alpha_i + \beta_{ij}$ and then solving for the estimate of μ. The estimated value of mean $\hat{\mu}$ is

$$\hat{\mu} = \frac{1}{n}\sum_i\sum_j\sum_k y_{ijk} = \frac{1}{n}\,y_{\dots} = \bar{y}_{\dots} \tag{10}$$

Where n is the total number of observations. The dot notation helps in simplifying an otherwise complex equation. Throughout the paper, the summation is taken across all parts, that is $i = 1, \dots, I$, and $j = 1, \dots, J_i$, and $k = 1, \dots, K_{ij}$. The sum of square (SS) of level-A (α_i), the level-B (β_{ij}), the sum of square of error (SSE), and the total SS (SST) are taken from [22]. The various parameters are,

$$\begin{aligned}\bar{y}_{ij.} &= \frac{1}{n_{ij}}\sum_k y_{ijk}\\ \bar{y}_{i..} &= \frac{1}{n_i}\sum_j\sum_k y_{ijk} = \frac{1}{n_i}\sum_j n_{ij}\bar{y}_{ij.}\\ \bar{y}_{...} &= \frac{1}{n}\sum_i\sum_j\sum_k y_{ijk} = \frac{1}{n}\sum_i n_i\,\bar{y}_{i..}\end{aligned} \tag{11}$$

TABLE. I. SUM OF SQUARE OF EFFECTS

Effect	Sum of Squares (SS)
A	$SSA = \sum_i n_i\ \bar{y}_{i..}^2 - n\,\bar{y}_{...}^2$
B(A)	$SSB = \sum_i\sum_j n_{ij}\,\bar{y}_{ij.}^2 - \sum_i n_i\ \bar{y}_{i..}^2$
Error	$SSE = \sum_i\sum_j\sum_k y_{ijk}^2 - \sum_i\sum_j n_{ij}\bar{y}_{ij.}^2$
Total	$SST = \sum_i\sum_j\sum_k y_{ijk}^2 - n\,\bar{y}_{...}^2$

The $n, n_i,$ and n_{ij} correspond to the number of pixels at various partitions of mask.

$$\begin{aligned}&n_{ij} = \sum_k M_{ijk} = 4 \qquad n_i = \sum_j n_{ij} = 4(4) = 16\\ &n = \sum_i n_i = 4(16) = 64 \quad M_{ijk} = 1\ (each\ pixcel)\end{aligned} \tag{12}$$

The degrees of freedoms (df) are given in Table II.

TABLE. II. DEGREE OF FREEDOM

Effect	Degree of Freedom (df)
A	(I-1) = 3
B(A)	$\sum_i(J_i - 1)$ = 4(3) = 12
Error	$\sum_i\sum_j(\sum_k 1 - 1)$ = 4(4)(3) = 48
Total	(n-1) = 63

The mean square (MS) of each effect is found by dividing the sum of square (SS) of an effect with the respective degree of freedom. The MS value with a degree of freedom $'v'$ has a Chi-square distribution with $'v'$ degree of freedom. This is represented by χ_v. The ratio of two Chi-square distributions with the respective degrees of freedom v_1 and v_2 gives F-distribution, that is F-test = χ_{v_1}/χ_{v_2}. The tables of F-test for various degrees of freedom are given in standard textbooks of

statistics [22]-[23]. The MS value of various effects is given in Table III.

TABLE. III. MEAN SQUARE OF EFFECTS

Effect	Mean Square (MS)
A	$MSA = \frac{SSA}{(I-1)}$
B(A)	$MSB = \frac{SSB}{\sum_i (J_i - 1)}$
Error	$MSE = \frac{SSE}{\sum_i \sum_j (\sum_k M_{ijk} - 1)}$

The respective F-tests are given Table IV.

TABLE. IV. F-TESTS OF VARIOUS EFFECTS

Effect	Mean Square (MS)
A	$F_A = \frac{MSA}{MSE}$
B(A)	$F_B = \frac{MSB}{MSE}$

Under the **Ω**-assumption, the presence of significant effect of a factor is confirmed by testing the hypothesis H_A against the Null hypothesis H_o as

$$\Omega: \quad \begin{cases} \mathrm{H_A}: \text{effect A is present} \\ \mathrm{H_o}: \text{effect A is not present} \end{cases} \tag{13}$$

Similar hypothesis is tested for effect-B by testing hypothesis H_B. In case either the effect-A, or effect-B are found to be present, then the next step is to find the exact location of an effect as derived in the contrast function discussed in next section.

IV. THE CONTRAST FUNCTION (CF)

A contrast function is applied in case the Null hypothesis of either effect-A, or effect-B is rejected against the alternate hypothesis. The primary objective is to identify if there is a significant variation in the horizontal, in the vertical, or in the two diagonal (45^o and 135^o) directions.

Definition: A contrast among a set of parameters, $\alpha_1, \alpha_2, \ldots\ldots, \alpha_I$ is a linear function of the α_i, $\sum_{i=1}^{I} c_i \alpha_i$, with known constant coefficient such that the condition $\sum_{i=1}^{I} c_i = 0$ holds.

As per above definition, the difference of two rows, $\alpha_i - \alpha_j;\ i,j$ are $1,2,\ldots,I$ form a valid contrast function. Similarly, a combination of rows with an appropriately selected coefficients form a valid contrast function. Other useful contrast functions can be formed in the vertical, and in the diagonal directions. The Gauss-Markov Theorem helps in finding the least square (LS) estimates.

Gauss-Markov Theorem: Under the assumption **Ω**: if $E(y) = \boldsymbol{X'\beta}, and\ \sum_{\boldsymbol{y}} \boldsymbol{I}\sigma^2$, then every estimable function $\psi = \boldsymbol{c'\beta}$ has a unique unbiased linear estimate $\hat{\psi}$ which has minimum variance in the class of all unbiased linear estimates. The estimate may be found by $\psi = \sum_{j=1}^{p} c_j \beta_j$ by replacing the $\{\beta_i\}$ with any set of LS estimates $\{\hat{\beta}_1, \hat{\beta}_2, \ldots\ldots, \hat{\beta}_p\}$.

$\boldsymbol{X'}$ is the transpose of coefficient matrix **X** consisting of zeros and ones. The matrix **X** is considered to have a full rank. The column matrix $\boldsymbol{\beta}$ represents the parameter matrix. The matrix $\sum_y I\sigma^2$ represents the covariance of observation matrix **y** which is assumed to be independent. All elements of this covariance matrix are zero, except the diagonal elements which are constant with the value equal to the variance σ^2. The matrix $\mathbf{c}'$ is the transpose of a column matrix c, which is a coefficient matrix fulfilling the requirement $\sum_{i=1}^{p} c_i = 0$.

A least square (LS) estimate of observations is found by taking the sample mean of observations in four directions. These are horizontal, vertical, diagonal 45^o, and diagonal 135^o. The sample mean in horizontal direction is found by summing pixels of a row across all columns. This is represented by $y_{i.} = \sum_j y_{ij}$. Similarly the sample mean in the vertical direction is found by $y_{.j} = \sum_i y_{ij}$. The LS estimates can be found by summing appropriate pixels in the diagonal 45^o, and diagonal 135^o directions. Using the Gauss-Markov theorem, an unbiased estimate of contrast function $\hat{\psi}$ in the horizontal direction is,

$$\hat{\psi} = \sum_i c_i \hat{\beta}_i = \sum_i c_i y_{i.} \tag{14}$$

The variance of $\hat{\psi}$ is found by

$$\sigma^2_{\hat{\psi}} = \sum_i c_i^2\, Var\,(y_{i.}) = \sigma^2 \sum_i \left(\frac{c_i^2}{J_i}\right) \tag{15}$$

J_i is the number of observations of each column to find the LS estimate. σ^2 is the variance with constant value of all observations. The σ^2 has an unbiased estimate that is equal to the mean square error (MSE) such that $E(\hat{\sigma}^2) = MSE$. The MSE is found by dividing the sum of square of error (SSE) with the respective degree of freedom. The estimate of contrast function is found by,

$$\sigma^2_{\hat{\psi}} = MSE \sum_i \left(\frac{c_i^2}{J_i}\right) \tag{16}$$

The objective is to test the hypothesis, H which tests significant variation across $\{\beta_i\}$,

$$\Omega: \quad \begin{cases} \mathrm{y_{ij}} = \beta_i + \mathrm{e_{ij}} \\ (\mathrm{i} = 1,..,\mathrm{I},\ \ \mathrm{j} = 1,..,\mathrm{J}) \\ \{\mathrm{e_{ij}}\} \sim \mathrm{N}(0, \sigma^2) \end{cases} \tag{17}$$

There are generally two methods for multiple comparisons of estimated values. These are Scheffe's S-method, and the Tukey's T-method. The T-method is preferred for pair-wise comparison, and the confidence interval is narrower than S-method. The S-method is applicable to all other types of comparisons. Here, the T-method is used as only the pair-wise comparison is required. Given the gray levels of two set of pixels as β_i and β_j, the confidence interval of the parameter $\psi = (\beta_i - \beta_j)$ is found by using the Tukey's T-test,

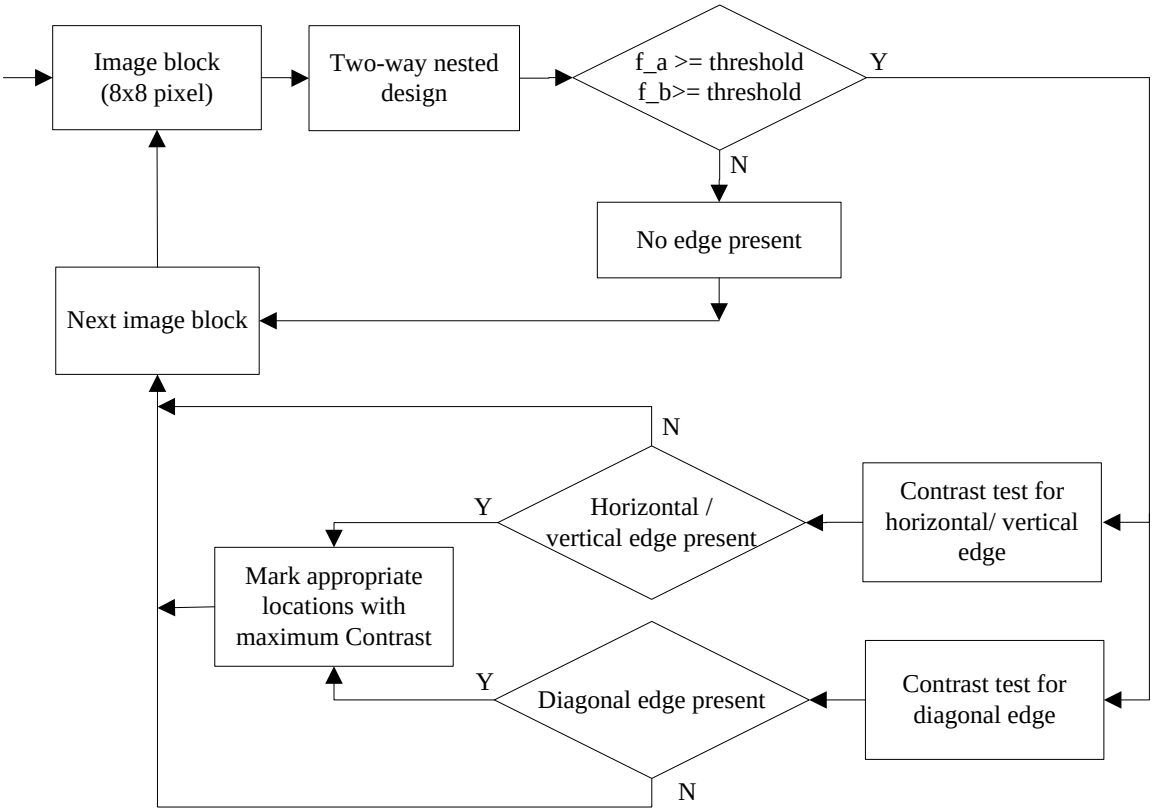


Fig. 3. Flow chart of edge detection algorithm

$$\hat{\psi} - Ts \leqq \psi \leqq \hat{\psi} + Ts \quad (18)$$

The $\hat{\psi} = (\hat{\beta}_i - \hat{\beta}_j)$ represents estimate of ψ. The unbiased estimate s^2 of variance σ^2 has 'v' degree of freedom and this is independent of samples. The ratio ψ/s is the Studentized range given by $q(r, v) = \psi/v$. The distribution of 'q' has been tabulated for various values of 'q' and 'r' in several standard textbooks of statistics. For reference see Table A-9 in [22]. An upper α-level of confidence interval corresponds to $(1 - \alpha/2)$ percentile level. As an example an upper $\alpha = 0.1$ level of confidence interval corresponds to a percentile of 95%. The test confirms presence of an edge, if the above confidence interval does not include zero value; that is either the entire range is positive or the entire range is negative.

V. Simulation Results

An overview of various steps is presented in Fig. 3. The algorithm considers a mask of 8x8 pixels. This mask size is selected to have four equal partitions, each 4x4 pixels. Each of these 4x4 pixels is further partitioned into four equal partitions, each 2x2 pixels. The processing is initiated from the top-left corner of an image, and scanned throughout from left to right and from top to bottom. A two-way nested design is applied on this mask. This generates two thresholds f_a and f_b. The threshold f_a signifies that there is enough variability among the four quarters of a mask each comprising of 4×4 pixels. The threshold f_b signifies that there is enough variability within each quarter of a mask. These thresholds are compared with the values from tables given in standard textbooks [22] using $f_{\_a} = MSA/MSE \geq F_{\alpha;v_1,v_3}$, and $f_{\_b} = MSB/MSE \geq F_{\alpha;v_2,v_3}$. If any of these inequalities do not hold then it is considered that the variability across four quarters of a mask, and the variability within each quarter is not significant. This is deduced in accepting the Null hypothesis of no significant variation at two granular levels. The mask is moved to the next adjacent location. In case the f_a or f_b is greater than the threshold, then the Null hypothesis is rejected, against the alternate. This demonstrates that there is enough variability within the mask and may contain an edge. The mask needs to be subjected to further analysis.

The next step involves in testing the mask for the presence of an edge in four directions using contrast functions. This step uses Tukey's T-test to mark edges in any of the four directions that are horizontal, vertical, 45 degree diagonal, and 135 degree diagonal. An edge in the horizontal direction can be present anywhere between 1st and 8th row of a mask. Several contrast functions are therefore generated, and the highest of them is compared with the threshold for testing the hypothesis for presence of an edge. The location of an edge is marked at the specific location with the largest disparity level. Similarly, the location of an edge in the vertical direction is marked at the highest contrast location. This results in exact identifying the most appropriate location of an edge. The edges in diagonal directions are simply marked on the respective diagonals of mask. In case the test fails to identify

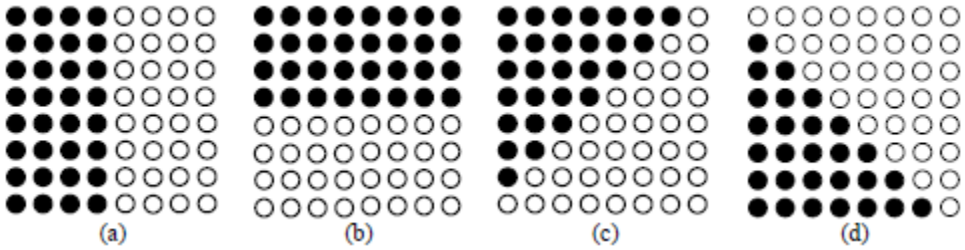


Fig. 4. Mask templates (a) horizontal (b) vertical (c) 45o diagonal (d) 135o diagonal

an edge in any of the four directions then the next mask is selected. Only one edge is marked at a particular mask location, in a particular direction. The mask is moved left-to-right and from top-to-bottom to scan the whole image.

A. Nested Design

The formulae for sum of square of effect-A (SSA), the sums of square of effect-B (SSB), and the sums of square of error (SSE) are given in Table 1. The aggregate of three sums of squares are always equal to the total sums of squares (SST). The respective mean square of effect-A (MSA), mean square of effect-B (MSB), and the mean square of error (MSE) are found by dividing the respective sums of squares with their corresponding degrees of freedom (d.f.) as given in Table 2 and Table 3. The mean square (MS) of an effect with a degree of freedom $'\nu'$ has chi-square distribution with a $'\nu'$ degree of freedom. This is represented by χ_ν. The ratio of two chi-square distributions is represented by an F-test. The F-test for effect-A is measured by $f_A = MSA/MSE$ which has chi-square $\chi^2_{\nu_1}/\chi^2_{\nu_3}$ distribution; where ν_1, and ν_3 are the degrees of freedoms of MSA and MSE, respectively. These degrees of freedoms are respectively equal to $(I-1)$, and $\sum_i\sum_j(n_{ij}-1)$ which are equal to 3, and 4(4)(3) = 48 respectively. The F-test for effect-B is measured by $f_B = MSB/MSE$, which is again a chi-square distribution, $\chi^2_{\nu_2}/\chi^2_{\nu_3}$ with a degrees of freedom ν_2 and ν_3. The corresponding values are equals $\sum_i(J_i-1)$, and $\sum_i\sum_j(n_{ij}-1)$, respectively. These are correspondingly equal to 4(3) = 12, and 4(4)(3) = 48. The details are given in Table II.

B. The Contrast Function

The contrast function is applied in four directions as in Fig. 4. The marked and unmarked pixels are represented by y_m, and y_{um}, respectively. The four contrast functions are formed as,

$$\psi = \beta_{m,w} - \beta_{um,w}$$
$$\hat{\psi} = \hat{\beta}_{m,w} - \hat{\beta}_{um,w} = \bar{y}_m - \bar{y}_{um} \quad (19)$$

The $\beta_{m,w}$ and $\beta_{um,w}$ corresponds to marked set of pixels and unmarked set of pixels. The $'w'$ corresponds to four directions as given in Fig. 4. The $\hat{\psi} = \hat{\beta}_{m,w} - \hat{\beta}_{um,w}$ is the corresponding estimated value, and the $(\bar{y}_m - \bar{y}_{um})$ represents pixel sample mean in marked and unmarked mask area. The total number of pixels in an 8×8 pixel mask is 64. The total mean square of contrast function MS_{cf} is partitioned into mean square of treatment MS_{tr}, and the mean square of error, MSE_{cf}. The corresponding degrees of freedom are respectively equal to 3, 12, and 63. Using the Table A-9 in [22], the threshold is taken as 4.31 for $r = 7$, and $v = 60$ for an upper 0.1 level of confidence interval corresponding to a percentile of 95%. The value of $v = 60$ is taken as the closest value to the required $v = 56$ value available in the Table A-9.

C. Discussion

The simulation results are given in Fig. 5 and Fig. 6. Fig. 5(a) gives a set of five test images consisting of Lena, house, chilies, cameraman, and baboon. Fig. 5(b) reproduces the images of Fig. 5(a) with an additive Gaussian noise of N(0, 400). The edges of the original image are detected using Sobel filter in Fig. 5(c). The mask size is a standard 3x3 pixels. Fig. 5(d) gives edges detected by nested design using an 8×8 pixel mask. The pixels of the mask are tested for the presence of level-A, and level-B effects. In case the hypothesis is affirmative then a follow-up contrast test are performed in above four directions. In order to be consistent with Sobel filter, the mask is shifted at every 3 pixels. This results in considerable overlap, but gives much improved results that are compared with those of Sobel filter. The comparison of Fig. 5(c) and Fig. 5(d) reveals that both the approaches exhibit

TABLE. V. PEAK SIGNAL-TO-NOISE RATIO, AND NUMBER OF EDGES IN PERCENTAGE OF PIXELS

S.No	Images	PSNR (dB)			Edges (% of pixels)					
		Original	N(0,25)	N(0,400)	No Noise		N(0,25)		N(0,400)	
					Sobel	Nested	Sobel	Nested	Sobel	Nested
1	Lena	14.5322	13.8472	13.8429	7.65	8.4629	8.6552	8.2973	5.5805	7.9597
2	House	12.8944	12.8612	12.4262	5.1731	5.9555	4.0482	5.5634	6.4716	6.0989
3	Chilli	13.5315	13.4983	13.035	7.2453	8.4164	8.0463	8.3641	6.8874	7.7858
4	Cameraman	12.2835	12.2604	12.0398	7.6683	8.4145	9.1682	7.8503	4.6974	7.6912
5	Baboon	16.1056	16.0374	15.1521	7.6286	12.9337	7.7011	13.0974	4.1538	14.2601

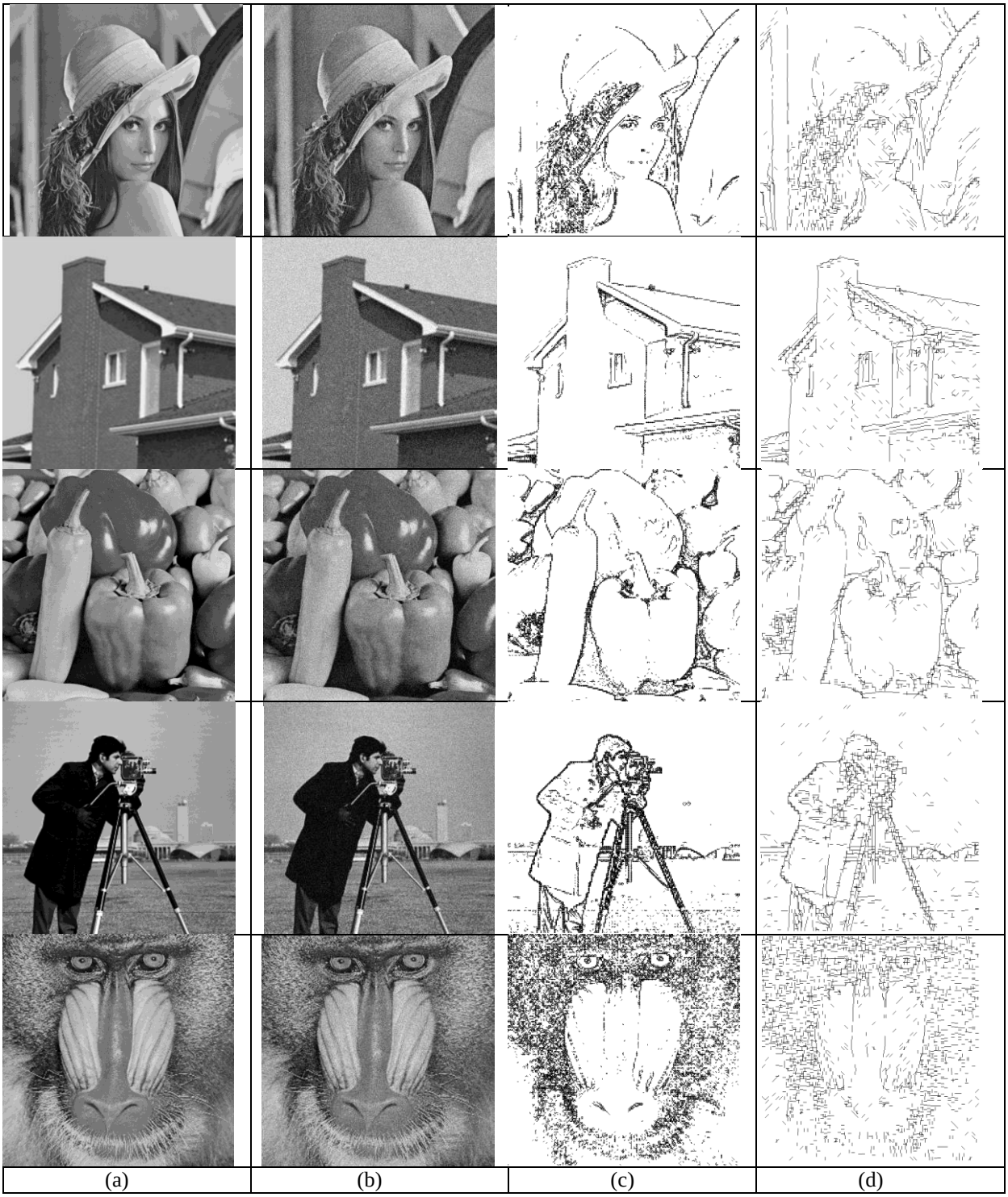


Fig. 5. (a) Original Images of Lena, House, Peppers, Cameraman, Baboon. (b) Images with additive noise of N(0, 400). Edge detection of noise free images using (c) Sobel filter (d) nested design

comparable results in identifying the edges. The nested design has performed slightly better in terms of few additional edges marked than the Sobel filter.

The edge detection with a moderate Gaussian noise of N(0, 25) for Sobel and nested design is given in Fig. 6(a), and Fig. 6(b). A comparison clearly explains that the Sobel filter is able to extract edges, but some of the background details are also marked. The nested design is able to identify clean edges. The performance of both algorithms under extremely heavy noise of N(0,400) is given in Fig. 6(c) and Fig. 6(d) for Sobel and nested design, respectively. A quick comparison of these images clearly reveals that the Sobel filter is unable to clearly

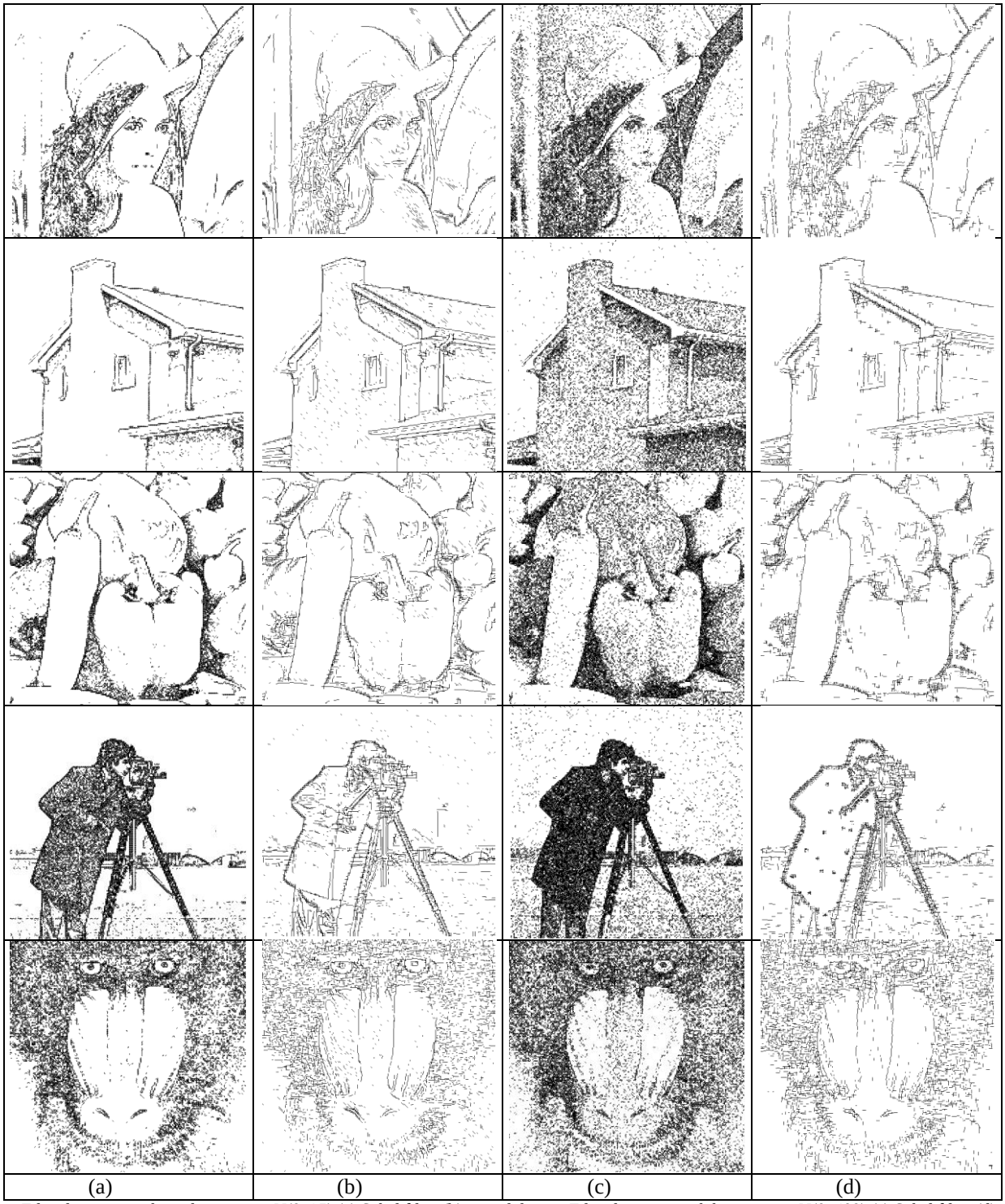


Fig. 6. Edge detection with moderate noise N(0, 25) (a) Sobel filter (b) nested design. Edge detection with heavy noise N(0, 400) (c) Sobel filter (d) nested design

mark the edges, while the nested design is able to identify edges with little or no background noise.

A comparison of numerical result is performed by comparing peak-signal-to-noise (PSNR) ratio,

$$PSNR = \frac{(255)^2}{\frac{1}{MxN}\sum_{i=1}^{M}\sum_{j=1}^{N}\left(I(i,j) - \hat{I}(i,j)\right)^2} \quad (20)$$

Where, M, N are the number of pixels in horizontal and the vertical directions. $I(i,j)$, and $\hat{I}(i,j)$ are respectively the original, and the estimated image.

A second criterion is the percentage of pixels marked as edges. The following simple formula is used for this measure,

$$\% \, pixels = \frac{\#\ of\ pixesl\ marked\ as\ edges}{Total\ pixels} \qquad (21)$$

VI. CONCLUSIONS

This paper presents a novel approach based on a nested design that marks the edges in an image. The complete design comprises of two steps. The initial test is based on two-way nested design which tests the variability of pixels in a mask. The decision is made by testing a hypothesis using an F-test. The variability of pixels is statistically tested to find if there is sufficient variability at two granular levels. The mask is subjected to a second test if there is sufficient confidence of enough variability.

The second test is based on contrast function (CF) using Tukey's T-test. The test identifies edges in four directions that are horizontal, vertical, and the two diagonals. The contrast function tests for the maximum contrast value at one of the several other possible locations. This selection is based on identifying the best location for marking an edge. In the diagonal directions, however, only the location at the middle of the mask is marked as an edge. The results are compared with edge detection using Sobel filter. A rigorous testing reveals that both the nested design and Sobel filter yields comparable results for noise-free images. The nested design, however, out performs the Sobel filter in situations where the image is corrupted with heavy Gaussian noise. It is clear that the nested design requires more processing than the Sobel filter. The processing time can be significantly reduced by parallel processing using advanced hardware like FPGA, and VLSI.

REFERENCES

[1] K. Arai, T. Higuchi, T. Murakami, "Estimation method of the total number of wild animals based on modified Jolly's method", International Journal of Advanced Computer Science and Applications (IJACSA)', no.1, vol. 8, pp. 341-346, 2017.

[2] T. Khairnar, Harikiran, A. Chandgude, S. Sivanantham, K. Sivasankaran, "Image Edge Detection in FPGA", International Conference on Green Engineering and Technologies (IC-GET)", 2015.

[3] Ziyan LIU. *et el*: 'Statistical pattern recognition for real-time image edge detection on FPGA', Proc., IEEE International Conference of Signal Processing (ICSP), 2014.

[4] A. Sheta, M. S. Braik, S. Aljahdali, "Genetic algorithms: A tool for image segmentation", IEEE Conference, 2012.

[5] M. T. Alonso, *et el*: 'Edge enhancement algorithm based on the wavelet transform of automatic edge detection in SAR images', IEEE Transactions on Geo Science and Remote Sensing, 49, (1), 2011.

[6] H. Zhang, J. Li, M. Wang, H. Li, "Image edge detection based on fusion of wavelet transform and mathematical morphology", The 11th International Conference on Computer Science & Education (ICCSE 2016), Nagoya University, Japan, August 23-25, 2016.

[7] S. Wu, Z. Zhang, H. Chen, C. Zhan, "Application of an improved method of wavelet transform in image edge detection", 11th International Conference on Computational Intelligence and Security, pp. 450-453, 2015.

[8] Z. Zhang, N. Saito, "Harmonic wavelet transform and image approximation", Journal of Math Imaging Vision, 38, pp.14-34, 2010.

[9] H. Hu, G. D. Haan, "Adding explicit content classification to nonlinear filters", pp. 291-305, 2011.

[10] J. Su, *et el*: "A fully statistical framework for shape detection in image primitives", 2010,

[11] F. Khellah, et el: "Statistical processing of large image sequence", Transaction of Image Processing, 14, (1), 2005.

[12] R. R. Rakesh, P. Choudhri, C. A. Murthy, "Threshold in edge detection: a statistical approach", Transaction of Image Processing, 13, (7), 2004.

[13] S. Borse, P. K. Bora, "A novel approach to image edge detection using Kalman filtering", IEEE 7th Annual Information Technology Electronics and Mobile Communication Conference (IEMCON), 13-15 October 2016.

[14] S. Guiming, S. Jidong, "Remote sensing image edge-detection based on improved Canny operator", 8th IEEE International Conference Software and Networks", 2016.

[15] W. Xue, X. Wenxia, L. Guodong, " Image edge detection algorithm research based on the CNN's neighborhood radius equals 2", International Conference on Smart Grid and Electrical Automation, 2016.

[16] A. Patel, A. Patel, "Performance enhancement in image edge detection technique", International Conference on Signal and Information Processing (IConSIP), pp.1-5, 2016.

[17] A. S. Ashour, M. A. El-Sayed, S. E. Waheed, S. Abdel-Khalek, "New method based on multi-threshold of edges detection in digital images", International Journal of Advanced Computer Science Applications (IJACSA), vol. 5, no. 2, 2014.

[18] X. Li, Y. Zhang, "Digital image edge detection based on LVQ neural network", IEEE 11th Conference on Industrial Electronics and Applications (ICIEA), 2016.

[19] H. T. Chang, J. Su, "Segmentation using codebook index statistics for vector quantization images", International Journal of Advanced Computer Science Applications (IJACSA), vol. 7, no. 12, 2016.

[20] Q. Boaming, J. liu, Y. Yufan, "An adaptive algorithm for grey image edge detection based on grey correlation analysis", 12th International Conference on Computational Intelligence and Security, 2016.

[21] R. C. Gonzalez, "Digital Image Processing", Pearson, Prentice Hall, 2008, 3rd ed.

[22] H. Scheffe, "The Analysis of Variance", Wiley, 1959.

[23] J. Neter, W. Wasserman, M. H. Kutner, "Applied linear statistical models", IRWIN, 2nd ed. 1985.

[24] R. Haberstroh, L. Kurz, "Line detection in noisy and structured backgrounds using Graeco-Latin squares", Computer Vision Graphics, Image Processing, Graph Models Image Processing, vol. 55, pp.161-179, March 1993.

[25] M. H. Benteftifa, L. Kurz, "Feature detection via linear contrast techniques", Pattern Recognition, vol. 26, pp. 1487-1497, 1993.

[26] J. Cheung, D. Ferris, L. Kurz, "On classification of multispectral image data', "IEEE Transactions on Image Processing, vol. 6, pp. 1456-1460, Oct 1997.

[27] J. F. Y. Cheung, C. W. Wicks, J. Genello, L. Kurz, "A statistical theory for optimal detection of moving objects in variable corrupted noise", IEEE Transactions on Image Processing, vol. 8, no. 12, 1999.

oDyRM: Optimized Dynamic Reusability Model for Enhanced Software Consistency

R. Selvarani
Professor, Dept. of CSE
ACED Alliance University
Bengaluru, India

P. Mangayarkarasi
Research Scholar
Visvesvaraya Technological University
Belgaum, India

***Abstract*—Accomplishment of optimization technique on Object Oriented design component is a very challenging task. The prior model DyRM has introduced a technique to perform modeling of design reusability under three real-time constraints. The proposed study extends the DyRM model by incorporating optimization using multilayered perception techniques in neural network. The system takes the similar input as is done by the prior DyRM, which is subjected to Levenberg-Marquardt optimization algorithm using multi-layer perceptron of configuration 4-24-2 to generate the optimal output of consistency factor. The paper discusses the underlying technique elaborately and presents the outcome that shows a good curve fits between experimental and predicted data. The model is therefore termed as optimized DyRM (oDyRM) to evaluate the consistency factor associated with the proposed model.**

Keywords*—Cost; Back propagation Algorithm; Design Reusability; Object Oriented Design; Optimization; Project management

I. Introduction

The study carried out in software reusability and software consistency, models and metrics suggests that it potentially benefits the clients from economic and performance perspectives. In the software industry the term 'reuse' is associated with cost efficiency for improving software development processes [1]. Various models for validating the preciseness of the design process depends on accuracy of results accomplished from a model that is directly proportional to the input data [2], [3], [4]. Accuracy is closely associated with reality. However, results may not be always accurate and hence sensitivity analysis is carried out. In order to gauge the level of accuracy and the factors affecting it, the study considers mathematical modelling for the purpose of optimization of design reusability. Mathematical modelling proved to be useful for validation and verification of the Software Reusability metrics. The other benefits of the software metrics are i) Development Benefits, ii) Maintenance, iii) Quantification of benefits and cost validation iv) and Use of economic models for validation. Economic models of reuse can help in making decisions concerning reuse and its applicability to address problems of uncertainty. The proposed study is an extension of the prior study where the enhancement is being carried out using optimization principle. The proposed system uses neural network to perform optimization and retrieve the consistency of the proposed software reusability model. Section 2 discusses about the related work followed by discussion of problem identification in Section -3. Section -4 discusses about proposed model followed by research methodology in Section-5. Implementation of proposed model is discussed in Section-6 followed by result discussion in Section-7. Finally, Section-8 makes concluding remarks.

II. Related Work

Many significant studies in the area of software engineering focus on reusability aspects as well as software consistency for study. The study introduced by Nair and Selvarani [5] presented a framework with the capability to compute the reusability index. The authors considered three of the Chidamber and Kemerer metrics viz. DIT (Depth of Inheritance Tree), RFC (Response for a Class) and WMC (Weighted Methods per Class). Same authors also carried out a complete analysis of the relationships that exist between internal quality attributes in terms of the complete suite of Chidamber and Kemerer metrics and the reusability index of software systems [6]. They presented a new regression technique for mapping the association between reusability and design metrics. Das et al. [7] studied about the mitigation techniques for the errors in the software modelling. They carried out the simulation study based on continuous time factor on case study of flight control software. Gargoor and Saleem [8] adopted swarm optimization technique along with neural network and exhibit better predictive capabilities to analyze software consistency issues. Strong et al. [9] adopted statistical methods to enhance the software consistency factor.

Emphasis on software consistency laid by Wason et al. [10] state the significance of automata-based approach. Anjum et al. [11] proposed a soft computing based technique using Poisson process to evaluate the software consistency. Similar direction of study also carried out by Yakonoyna et al. [12]. Sabbineni and Kurra [13] implemented a dynamic technique for the purpose of consistency allocation of software components. Sheakh [14] presented an enhanced algorithm to improve the performance of software consistency. However, the extent of the author's contribution is found to be poorly discussed with less evidence to prove its efficiencies. Kumar also carried out by Antony [17]. Fetaji et al. [18] and Singh et al. [19] also carried out empirical investigation towards improving reusability as well as software consistency on the object-oriented design components.

III. Problem Description

The identified problems after reviewing the existing research contribution towards software reusability are i)

Existing study emphasizes on code reusability and not design reusability, ii) Existing techniques of software reusability doesn't consider essential attributes of project management e.g. human resources, skill gap, requirement volatility, training, cost of new development, etc., iii) Less extent of optimization of the design reusability from the OO component design is found, iv) The studies using CK metric discussed by various authors are found with theoretical assumptions. Majority of the studies considering CK metrics manipulates the same formulations with minor concern to introduce practical scenarios, v) Few consideration or attempt to model real-time constraints are found in existing literatures, hence the outcomes of the model are more inclined to hypothetical figures and less possibilities of applicability with real-time requirements, and vi) Software reusability as well as software consistency is not found to be combine studies. Modern day software development methodologies encounter more dynamicity, uncertainties, and unforeseen possibilities of failures of projects. Such issues cannot be addressed by theoretical and hypothetical framework of software consistency, which is found to be less connected with software reusability in existing system. Hence, all the above problems are highly critical and invoke various issues while attempting novel modeling of software reusability management. There is a need of designing a model considering presences of various uncertainties are errors to closely check the efficiency level of the solution. Such critical emphasis was never found in the literatures till date and hence calls for addressing the same. The next section will present a solution towards this problem:

IV. PROPOSED SYSTEM

The prior model named as Dynamic Reusability Model (DyRM) created a fundamental base for establishing a relationship among the CK metrics (CBO, RFC, WMC, DIT, and NOC) and design reusability [20]. DyRM was basically designed for the purpose of evaluating the impact of design reusability in software engineering under three real-time constraints e.g. quantity, work schedule, and cost of new development. Owing to various possible uncertain scenarios (e.g. requirement volatility, change management, skill gap, attrition rate), there is a possibility that the outcome of DyRM model may be associated with significant errors that are hard to find. Hence, there is a need of an optimization technique to our prior DyRM model for achieving following benefits e.g. i) Inconsistency reduction in DyRM leading to better predictability of the reusability outcomes, ii) Parallel computation of multiple and heterogeneous constraints-based reusability estimation in software project developments, and iii) Predictive optimization with assured consistency and robustness in future use. Hence, the development of a novel predictive optimization technique over DyRM model for enhancing the reusability management of software projects to greater extent. It is also applicable to overcome all the unseen constraints to a large extent that are not considered in this model. Hence neural network multi-layered perceptron is applied for developing the proposed predictive optimization principle. The technique allows for multiple forms of input to the processor in the form of real time constraints, which after processing gives the output of consistency score and uniformity score. The proposed system oDyRM is implemented in a typical way as exhibited in Fig.1

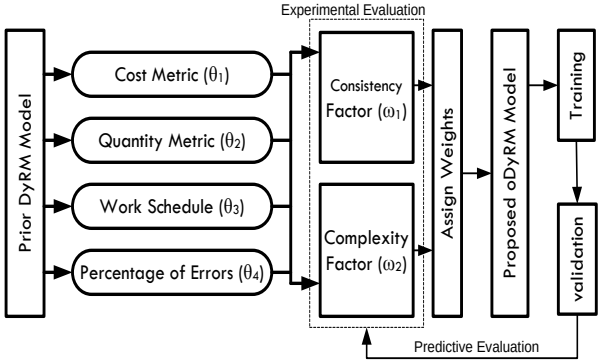


Fig. 1. Schema of Proposed Implementation (oDyRM)

V. RESEARCH METHODOLOGY

The proposed system adopts analytical research methodology, where the prime target to check for level of consistency for prior DyRM model after being subjected to predictive optimization. Similar assumptions and problem formulation discussed in design principles of DyRM model is continued in proposed system also. Apart from inclusion of older forms of 3 inputs of design attributes (or constraints) i.e. i) quantity, ii) work schedule, and iii) cost of new development, the proposed model also considers a new input attribute i.e. percentile of error for performing analysis for optimal consistency. The rationale behind selection of 4th design attribute of error is – DyRM chooses only three input attribute which definitely lowers the scope of design reusability value in sophisticated software projects. There is also a possibility of many other input attributes (or constraints) that equally impacts design reusability computation e.g. skill gap, requirement volatility, change management, etc. Such abstract parameters combine to have negative impact of design reusability computation and hence may eventually affect the optimization process. Therefore, we consider such parameters as percentile of error, whose values are defined between 1-4 depending on total numbers of inputs. The outcome of optimization is valued with respect to consistency factor and complexity factor. Fig.2 shows. The objective is achieved by developing a multi-layered perceptron for predicting the consistency factor in reusability model. Investigation was conducted to determine the strength of design metrics in the form of consistency after 100 iteration rounds testified with various parameters like quantity, work schedule, cost of new development, and percentage of errors. A total of 300 permutations of 4 input variables were developed. Out of these data sets, certain data sets (80% of total data) were used for training and the remaining data sets were used for validation. The input and output vectors have been normalized in the range (0, +1) using suitable normalization factors or scaling factors. The following input parameters were selected to predict the consistency factor.

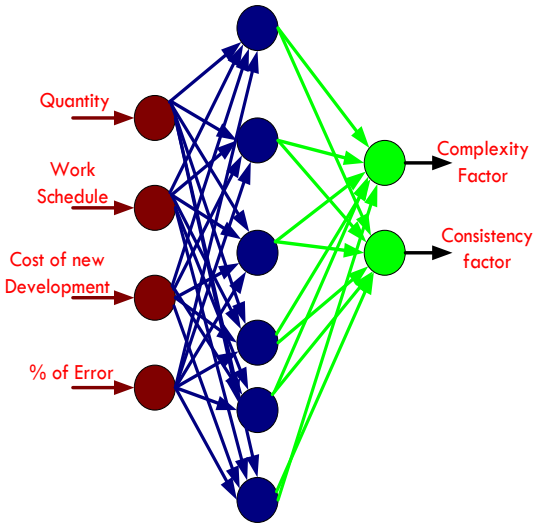


Fig. 2. Inputs and Outputs of oDyRM

- **Cost (θ_1)**: This metric estimates the cost for new development of the reusable design of projects from prior DyRM model. The values are further randomized within a scope to generate cost involved various projects.
- **Quantity (θ_2)**: This metric estimates the number of the projects with reusable designs (with higher values of θ_1) and further randomized statistically.
- **Work Schedule (θ_3)**: This metric evaluates the optimal time required to dispatch a particular projects (with higher values of θ_1 and$\theta_{2)}$.
- **Percentage of Errors (θ_4)**: This is the newly introduced metric that introduces random errors that may be possibly caused due to various extrinsic factors in software development methodologies.

Based on the input parameters selected, the proposed model was formulated as

$$IP_{oDyRM} = [\theta_1, \vartheta_2, \theta_3, \theta_4] \quad (1)$$

The following output parameters were selected to be predicted from the network model e.g.

- **Consistency Factor (ω_1)**: This factor statistically evaluates the extent of consistency after adopting inputs specified in eq.(1). The mathematical representation of novel consistency factor is,

$$\omega_1 = \sum \frac{IP_{oDyRM} - 0.1}{0.8} . \Delta wt \quad (2)$$

The above eq.(2) considers inputs from eq.(1), generated weight *wt* and considers the lower limit of 0.1 and higher limit of 0.8 in probability theory. The system chooses to consider 0.8 as accomplishing higher consistency factor of 0.9 and 1 is impractical assumption in probability theory.

- **Complexity Factor (ω_2)**: This factor checks the uniformity of the generated values after performing validation to check the consistency of the consistency factor (ω_1).

VI. Algorithm Implementation

The system mainly attempts to ensure better reusability management by enhancing consistency factor. The implementation of the proposed system starts by evaluating cost metric, quantity metric, work schedule metric, and percentage of error metric from prior DyRM model, which is statistically subjected to evaluation of consistency factor. As the enhancement of prior DyRM model is carried out using multi-layered perceptron, hence, it is important to assigned weight. Adopting the techniques of multi-layered perceptron, the proposed system has only three layers (input, hidden, and output layer). Hence, the configuration of the oDyRM model is: 4-24-2, where 4 is the number of inputs nodes, 24 is the number of hidden nodes and 2 is the number of output nodes. The calculations of the weight to be allocated is done as number of weights to be determined (4x24) + (24x2) =144. The computational algorithm used for the training is as follows:

Algorithm for Optimizing ω_1 andω_1

- **Input:** IPoDyRM, wt,
- **Output:** consistency adoptability factor (ω1), complexity factor (ω2)
- Initialize IPoDyRM, wt, Neurons
- Data Obj ← Read(Datafilefor Training)
- Compute, Min/ Max (Data Obj), → Normalization for Min/ Max ← output
- Set Epoch, Initialize Neural Network NLf() / Layers, Train, Test , error
- Evaluate consistency adoptability factor (ω1), complexity factor (ω2).

The algorithm is designed using the similar concept in multilayer perceptron, where the inputs are given as Cost Metric (θ_1), quantity metric (θ_2), work schedule metric (θ_3), and percentage of error metric (θ_4). The inputs are fed for processing and furnished an output as consistency adoptability factor (ω_1) and complexity factor (ω_2). The training was for 2000 iteration and checked for the curve-fitting using neural network in numerical computing simulation for both experimental and predicted data. The algorithm mainly took less than 3.5 seconds to execute and outcomes are discussed in next section.

VII. Results and Discussion

Firstly, the technique of accomplishing the data as well as processing the data for claiming the optimization in design reusability concept is discussed in this section. Case studies of two software projects of Enterprise Resource Planning (ERP) which are mainly open source are taken namely Apache OFBiz [21]. The OFBiz is configured on user machines. The classes were re-configured related to asset management, human resource, accounting, and inventory management etc. to have real-time environment of ERP application. A plug-in Metrics

1.3.6 [22] is used for extracting the CK metrics which is used in prior DyRM model i.e. CBO, RFC, WMC, DIT, and NOC. The DyRM model is used to estimate design reusability under multiple iterations along with new inclusion of 4th new input parameter i.e. percentage of error. The outcomes are recorded in comma delimited file to use as an input to Statistical analysis applications such as SPSS etc. The t-test and analysis of variance is performed in SPSS to observe the statistical outcome, which is further arranged in the form of 4 input parameters in numerical computation to carry out training and validation phase. The system considers the input from the SPSS where the empirical analysis for the 4 design parameters of the oDyRM model is considered (cost, quantity, work schedule, and percentage of error). Table 1 highlights the outcome after 2000 epochs of training period. This optimization process uses all forms of non-linear input data that may eventual lead to non-linear optimization problems. Hence, the training is carried out using Levenberg-Marquardt algorithm for solving non-linear squares problems.

TABLE. I. DETAILS OF SCALING FACTORS

Nature of vector	Parameter	Minimum Value	Maximum Value	Scale Factor
Input Vector	Cost Metric	0.3	0.5	0.7
	Quantity Metric	2	3	4
	Work Schedule	1	3	3
	Error	1	4	4
Output Vector	Consistency Factor	0.6	0.8	1.3
	Complexity Factor	4.9	13.8	18.7

The table highlights basically two types of information i.e. i) experimental data (before training) and ii) predicted data (after training). The experimental data is being calculated using SPSS, which after feed to the training module generates the predicted data in numerical computation. Both experimental and predicted data shows the higher rate of data consistency as a part of validation test in multi-layer perceptron based error reduction. The higher consistency factor (near to 1) accomplished by the training state confirms the robustness of the proposed optimization model using probability theory. In order to accomplish the optimal outcomes, the proposed system used a scale factor which is calculated as (MinVal+MaxVal)-1, which in other sense refers to the degrees of freedom to evaluate the outcomes. Table 1 gives the details of the input vectors and output vectors. The model considers neural network parameters as number of nodes, nodal properties corresponding to the input as well as output vectors along with hidden layers. The training process adjusts the epoch in run time as per the error. It was observed that there are 24 neurons present in hidden layer. Hence, the configuration chosen for the proposed model is 4-24-2. The proposed system uses backpropagation algorithm to evaluate its weight depending on the used gradient search technique after generation of the weight-factors in SPSS. We define the number of weight as 144 and scale the weights using SPSS. After the training is accomplished, it is found that expected outcomes as well as actual outcomes are highly matching with each other, as seen in Fig.3.

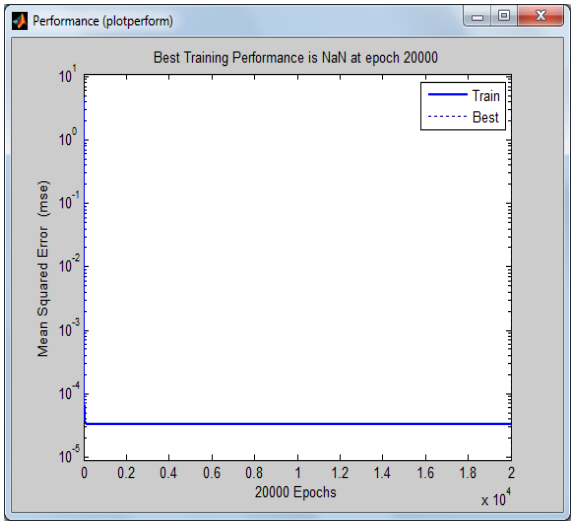


Fig. 3. Performance of oDyRM

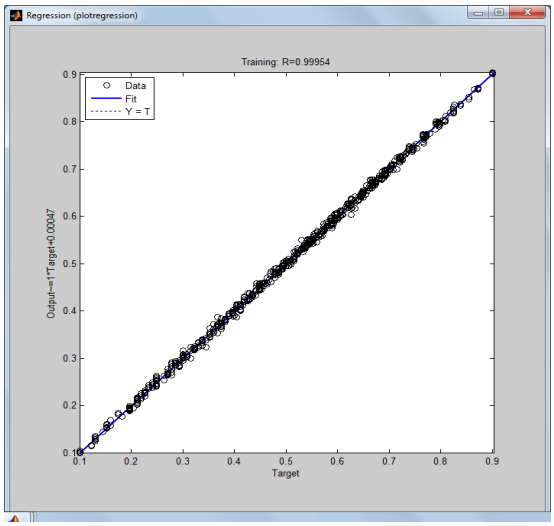


Fig. 4. Regression Analysis of Proposed eDyRM

Fig.3 highlights the regression analysis performed on proposed eDyRM model. The outcome is found to have the best fit with the training data as well as better uniformity in the error outcomes as linear and deterministic trend of error curve. The regression analysis is also performed for RMS value (Fig.4) with the increasing number of epoch. The duration of the epoch is from 100-2000.

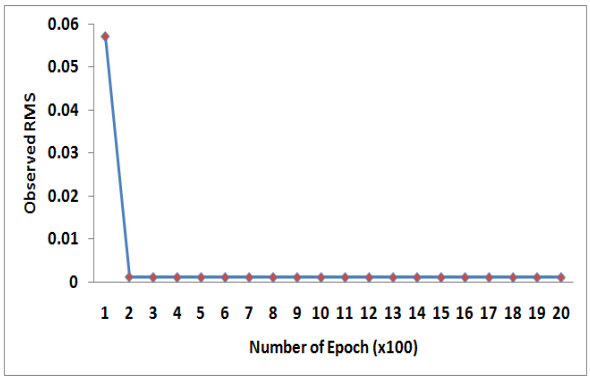


Fig. 5. Analysis of RMS

The optimization principle adopted in the proposed oDyRM model ensures the better solution towards enhancing normal gradient descent as well as drastic minimization of RMS values (Fig.5). Hence, the proposed technique has the faster convergence rate irrespective of the variances in the input data size. The proposed technique is also flexible for incorporating more empirical methods for testifying the reusability metrics pertaining to OO design optimization in software engineering. The extent of design complexity is always a matter of worry in when it comes to design reusability. Even with higher values of epoch (>2000), the outcomes were found with similar trend of consistency.

VIII. CONCLUSION

Design reusability is one of the prime concerns in software engineering especially when working on complex Object Oriented software components. The proposed paper has enhanced the prior model DyRM. The enhancement includes evaluating the consistency factors and consistency factor of proposed optimized model oDyRM using multilayered perceptron approach of neural network. The system takes the input of cost of new production to generate reusable design components, quantity of the projects to be delivered, flexibility in the work scheduling, and percentage of possible errors. The outcome of the study shows that the model is able to check for error, reduce it recursively till it gets best curve fitting for the trained and real-data. Hence, the proposed model can be used by any technical architect to evaluate the possible risk or gain to adopt a particular software development methodology to measure the level of effectiveness in design reusability and software consistency.

REFERENCES

[1] Wentzel, K.D.: Software Reuse - Facts and Myths. Proceedings of the 16th international conference on Software engineering. IEEE Computer Society, 267-268 (1994)

[2] Vanmali, M., Last, M., Kandel, A.: Using a Neural Network in the Software Testing Process. International Journal Of Intelligent Systems, 17, 45–62 (2002)

[3] Musilek, P., Meltzer, J.: Assessing Empirical Software Data With Mlp Neural Networks. ICS AS CR, (2005)

[4] Adebiyi, A., Arreymbi, J., and Imafidon, C.:Security Assessment of Software Design using Neural Network. International Journal of Advanced Research in Artificial Intelligence. 1, 4, (2012)

[5] Nair, T.R.G and Selvarani R.: Estimation of Software Reusability: An Engineering Approach. Association for Computing Machinery (ACM) – SIGSOFT. USA, 35, 1 (2010)

[6] Selvarani, R., Nair, T.R.G.: Software Reusability Estimation Model Using Metrics Governing Design Architecture, International Book: "Knowledge Engineering for Software Development Cycles: Support Technologies and Applications", Engineering Science Reference, IGI Publishing, USA (2009)

[7] Das, S., Dewanji, A., Sobti, A.: Software Reliability Modeling with Periodic Debugging Schedule. Indian Statistical Institute (2013)

[8] Gargoor, R. G. A., Saleem, N. N.:. Software Reliability Prediction Using Artificial Techniques. IJCSI International Journal of Computer Science Issues. 10 (4) (2013)

[9] Strong, K.: Using FMEA to Improve Software Reliability. In Pacific Northwest Software Quality Conference (PNSQC) (2013)

[10] Wason, R., Ahmed, P., & Rafiq, M.Q.: Automata-Based Software Reliability Model: The Key to Reliable Software. Int. J. of Software Engineering & Its Applications. 7(6), 111-126 (2013)

[11] Anjum, M., Haque, M.A., & Ahmad, N.: Analysis and ranking of software reliability models based on weighted criteria value. International Journal of Information Technology and Computer Science (IJITCS), 5(2), 1 (2013)

[12] Yakovyna, V., Fedasyuk, D.,.Nytrebych, O., Parfenyuk, I., Matselyukh, V.: Software Reliability Assessment Using High-Order Markov Chains. International Journal of Engineering Science Invention. 3(7), 1-6 (2014)

[13] Sabbineni, S., & Kurra, R.:Estimation of Reliability Allocation on Components Using a Dynamic Programming. International Journal of Computer Science Issues (IJCSI). 10(3) (2013)

[14] Sheakh, T.H.: An Improvised Algorithm for Improving Software Reliability. International Journal of Computer Applications. 79 (17) (2013)

[15] Kumar, D., Dinker, A. G.: Enhancement of Reliability in Object-Oriented Software Reliability Model. International Journal of Advanced Research in Computer Science and Software Engineering. Vol. 4 (2014)

[16] Kapila, H., & S.Singh.:Analysis of ck metrics to predict software fault-proneness using bayesian inference. International Journal of Computer Applications. 74 (2013)

[17] Antony, P.J.:Predicting Reliability of Software Using Thresholds of CK Metrics. International Journal of Advanced Networking & Applications. 4(6) (2013)

[18] Fetaji, B., Reci, N., & Fetaji, M.:Analysing and Devising a Model for Trustworthy Software. Recent Advances in Electrical and Computer Engineering. (Retrived 2015)

[19] Singh, P. K., Sangwan, O.P., Singh, A.P., and Pratap, A.: A Quantitative Evaluation of Reusability for Aspect Oriented Softwareusing Multi-criteria Decision Making Approach. World Applied Sciences Journal. 30 (12) 1966-1976 (2014)

[20] Selvarani, R., and Mangayarkarasi, P.:A Dynamic Optimization Technique for Redesigning OO Software for Reusability. SIGSOFT Softw. Eng. Notes. 40 (2) 1-6. (2015)

[21] "Apache Ofbiz®", http://ofbiz.apache.org/, (Retrived, 05th Jan, 2017)

[22] "Metrics 1.3.6-Getting Started", http://metrics.sourceforge.net/, (Retrived, 05th Jan, 2017)

A Review of Secure Authentication based e-Payment Protocol

Mr.B.Ratnakanth
Dep.of Computer Science and Systems Engineering
Andhra University
Visakhapatnam, India

Prof.P.S.Avadhani
Dep.of Computer Science and Systems Engineering
Andhra University
Visakhapatnam, India

***Abstract*—The growth of e-commerce platform is increasing rapidly and possesses a higher level of hazard compared to standard applications as well as it requires a more prominent level of safety.Additionally, the transaction and their data about clients are enormously sensitive, security production and privacy is exceptionally crucial. Consequently, the confirmation is generally vital towards security necessities as well as prevents the data from stolen and unauthorized person over the transaction of e-payment. At the same time, privacy strategies are essential to address the client data security. Because of this, the data protection and security ought to be viewed as a central part of e-business framework plan. Particularly enormous consideration is given to cash exchanges assurance. In the past decades, various methods were created to permit secure cash transaction using e-payment frameworks. This study will review and discuss the e-payment scheme. It uses various encryption algorithms and methods to accomplish data integrity, privacy, non-repudiation and authentication.**

Keywords—Security protocol; smart card; encryption technique; payment protocol; E-commerce

I. INTRODUCTION

The internet is raised the area for social communication as well as coordinated effort. Specifically, it chose source to online accessible e-services like e-banking, e-voting, e-government, e-commerce and so on. it permits organization as well as persons towards exchange ideas, communicate, services and trade goods as more productively [1]. In the developing countries, the internet has been warmly accepted and established for social interaction with information. However, the part of the internet in e-commerce (commerce/trade) is still extremely constrained. The e-commerce plays an essential part in financial improvement through decreasing the expense of services and products. Also, it is known as conducting business through online. Further, it has a capable to offer sell or buy the products, services, and information through online as well as used for other internet applications. The amount transaction is necessary for various trading action, and it should be reliable and safe between transaction parties. The e-payment method is the essential part in e-commerce platform[2]. The EPSs (Electronic payment system) enables the significant role afterward the client choice towards paying the services or products as well as carry payments from client to sellers in an effective way[3].Additionally, the requirement of payment through mobile phone have emerged for the growth of mobile electronic commerce [4]. Also, the EPSs system encourages medium and small size enterprises as well as permits them towards contend with giants in the same commercial center [5]. The web application wants a robust security that has a prominent attribute towards ensuring client secret information. Additionally, on the internet based e-payment framework security is the main problem. There are various web dangers that increase the risk as well as influence the security scheme on behalf of the electronic transaction. Mostly, this depends on individually recognizable proof numbers, keys to get to their own record data and passwords. This sort of validation framework can't confirm or verify the personality of the clients who he/she rights to be[6]. Over an insecure communication channel, the security of e-payment is a challenging task which incorporates numerous serious zones as a robust encryption method, trusted third party and protected communication channel towards maintaining the online database[7].

A. Need for the Study

The internet is the communication channel that permits more people or organizations towards converse each other without abundant efforts and charges. Recently the online hacker is widespread all over the world, and it makes main resource profits for criminals. Presently online misrepresentation is extremely famous everywhere throughout the world; it has turned into a noteworthy wellspring of income for offenders. The banks or budgetary foundations are extremely mindful in identifying and counteracting online fakes [8], [9].In spite of that electronic business is a developing marvel that upcoming improvement towards a substantial range, vulnerable through the absence of proper payment framework. Subsequently, the majority of the business to customer installments through internet recently performed by MasterCard’s and indeed problematically payment medium because of security, trust issues, and cost. Further, the requirement of new payment scheme obviously rises up out of the previous circumstance [10]–[12]. In order to resolve this situation, previous research, and development in internet based payment system proposed by online based e-payment system, a great extent of which has been put to utilize. This was conceivable because of the fortifying variables scheduled above, and in any case because of the reducing cost and availability of the assisting technology[13]. For overcoming these issues in the e-payment field, need to propose a conventional payment system for online business using an intermediate.

II. Related Work

Recently, the secure wireless transactionschemehas turned into a research extent towards payment system.Some of them have concentrated towards security and privacy issues alone such as Ray et al. [14], Kouta et al. [15], Cui et al. [16], Tiwari et al. [17],Raghuwanshi et al. [18], Abdellaoui and Pasquet[19], Mazumdar and Giri[20], Takyi and Gyaase[21],Lawal et al. [22],Khan et al. [5],Aigbe and Akpojaro[23],Kim et al. [24]. On the other hand, focused on secure mobile payment system[25]–[30]and some of them have focused Secure Electronic Transaction protocol[31]–[33].Those are discussed as follows.

A. Studies Related to Secure Mobile Payment System

Hu et al.[34]proposed a secure mobile micropayments using AMA method. By this method, the customer gets a services or goods from mercantile in various domains through mobile micropayments except disclosing his/her privacy. Additionally, without expanding correspondence overheads noticeable all around, the most computational exertion is moved to the wired system towards reducing the computational overheads on a smart card or mobile with restricted storage and computational capacity. Further, the execution of bottlenecks of the whole framework focuses the clearing, credential, and settlement.Subsequently, sometimes the credential center is also difficult issues in this payment system. In the cryptography based symmetric key approach, this is a common issue but in the future studies, this would be improved.

Tabandehjooy and Nazhand [25]proposed a secure and lightweight protocol for mobile payments. Further, they prevent their non-repudiation as well as authenticate a client using the complex key. In this cryptography method, each data exchange the information, and it makes integrity and more confidentiality. Also, it improves the trust payment. The results show that the protocol considerably improves the customer trust through complex key since not ever enter their secure data in the electronic payment system.

Sazzad et al.[26]proposedmobile banking system using SMS services and offered an opportunity to all customer of economic foundation towards utilizing the available facilities through netbanking. This scheme mainly used for all users and does nor require any refined internet connectivity or cellular phone in the mobile handset.Also, they provided main services such as amount transfer between authorized clients, balance inquiry, bill payment as well as save the precious time. The internet base secures communication between banking and mobile server that was beyond the client control and is handled critical problems. Also, employed audio based digital watermarking for providing higher security with voice authentication. By this secure handle system, the clients easily attract to e-banking, specifically in emerging countries.

Suryotrisongko and Setiawan[35] proposed a mobile payment scheme for the cooperative system. To improve the security, they used quick response encrypted message and two-factor authentication. However, this study needs to be validated and tested most thoroughly in security aspect. Furthermore, more massive experiments and survey need to be conducted in order to measure user's acceptance regarding this model.

Mwafise and Stapleton[36]discussed the institutional and socio-technical domain to attain technology adaptation. However, this study was restricted in that it didn't exploit results in different locales which will have been a rich contribution to either reinforcing a portion of the focuses created or in testing provincial contrasts of perspectives on innovation selection which could likewise be helpful in building up a strong model of the deterministic variables on innovation acceptance in emerging countries.

Isaac and Zeadally[27]put forth a proposal for a protocol that was lightweight in order to achieve on-line payments that were secure. This was in a limited eventuality where a direct mode of communication could be possible between the clients and merchant. The protocols that postulate rely on symmetric cryptographic techniques which function effectively on lower requirements of computation. Additionally, the Payment Gateway plays an active role in processing payments as it discharges and its role as a proxy that permits communication between the client and merchant. Even though the suggested protocol was devised for a mobile payment system for a limited eventuality, the security properties remain intact. The study of the performance indicates that the mobile payment protocol that is suggested demands lesser computation than that required for KSL and LMPP, which in turn results enhanced end-to-end performance and can be installed on mobile devices operating with minimum or reduced computational resources.

Kim et al. [28]studied the security threats, requirements of security, and various security modules as a user processes the payment through NFC-based mobile. Mobile payment offers convenience and efficiency. But, at this situation, the process of payment is initiated post-authentication by validating key information, such as card and personal information which is stored and handled on NFC-based USIM. Therefore, for the purpose of secure mobile payment, the components design and modules was crucial. Hence, in this study, for a secure mobile payment system through NFC, examined the threats to security, the requirements and the requirement of every module in payment procedures. This will undertake a study in the future about the protocol that devised and utilized superior and strong encryption algorithms as well as safe, secure validation mechanisms that can tackle security threats.

Ahamad et al. [29]suggested a Secure Mobile Payment Framework (SMPB) that relied on biometrics and worked through Universal Integrated Circuit Card (UICC) and Wireless Public Key Infrastructure (WPKI). Additionally, they carried out a comparative analysis between this method and works in the recent past and learned that this method was a better option in the context of achieving an end to end security.This mobile payment protocol which originates from Mobile Payment Application to Bank Server results in Fair Exchange, ensuring Authentication, Confidentiality, Non-Repudiation, and Integrity, pre-empts spending twice, spending above limits and money laundering, in addition, to being able to remain resilient duringreplay, MITM (Man in the Middle) and Impersonation attacks. A plan exists to validate

the mobile payment protocol by utilizing AVISPA and Scyther Tool in the anticipated circumstances.

Lomte et al. [30]put forth a proposal for a secure payment protocol, taking into consideration the limitations of cellular networks in developing nations. Moreover, this method fulfills the satisfaction quotient in terms of convenience and ease of use, two criteria of mobile users making small transactions. Additionally, it offers security for the transaction and non-repudiation property which is mandatory for macro payments. Even though the suggested technique was developed so as to be in harmony with the present GSM network, the modular design envisaged is future-ready, it will accept improvements in the future to mobile network technology and associated infrastructure, for instance, EMS and MMS, requiring minimal changes to the protocol structure. But, the exact nature of implementation of workflow will ultimately depend on the disposition of the user. Like a light motive, businesses with multichannel infrastructure need to unify harmoniously the level of security for m-payment and security architectures that are web-based for m-payment so as to safeguard their businesses and for the development of future-proof architectures.

Ting et al. [37]investigated the effects of subjective norm, attitude with perceived behavioral control on intention by mobile payment system amongst Chinese and Malays in Malaysia, through the use of underlying based planned behavior theory. They collected and tested the data by independent sample t-test with multiple linear regressions in SPSS. Because of the limits of SPSS, four separate models are needed towards executing the regressions. Furthermore, when performing bunch examination by ethnicity, the extension is delimited to Malays and Chinese. Thus, future studies can be directed utilizing Structural Equation Modeling (SEM) to better clarify expectation towards m-installment framework in a solitary auxiliary model.

B. Secure Electronic Transaction Protocol

Guan et al. [31]put forth an expandable SAFER (Secure Agent Fabrication, Evolution & Roaming)-based e-payment module that meets the requirements of commerce depending on agents. Secure Electronic Transaction (SET) and E-Cash protocols were selected to function as modes of payment. The well-defined interface also makes possible the inclusion of additional features in one single module, without affecting the reliability in other related modules. Further improvements in future of the system may see the inclusion of agent security measures. Additional research has been undertaken in this field by projects running paralleled, and the results obtained can be utilized to improve the present system. Additionally, different electronic payment schemes can be considered for implementation in the form of additional payment modules so as to build on the flexibility of this framework as well as convenient to users.

Wang and Varadharajan[32]posit a secure payment protocol that was agent-assisted in a manner that supports multiple payments, which utilizes Signcryption-Share scheme and Signature-Share scheme employing a Trusted Third Party (TTP). The protocol that has been suggested put forth the principle that every player with a role is aware of what is actually mandatory for him/her and followed similar to SET when the non-repudiation property gets improved. The dispatch agents selects flexibly and in a dynamic manner the merchant and affix a sign for the cardholder with the consent of the TTP, and all of this executed without having to reveal any information of a secret nature pertaining to the card to the merchants and the TTP. The information of offers is safeguarded from unrelated merchants. In order to minimize the risk of using the services of mobile agents, the reliability and integrity of merchants can be assessed well in advance. Nevertheless, this paper has to coordinate the suggested protocol into the PumaMart system for the purpose of enhancing performance – a B2C marketplace that is mediated by an agent [38], [39] executed over and above Java and IBM Aglets toolkits [40].

Sun [41]SET protocol is a system that is concentrated, comprising a request for the request, a payment verification, and the final payment through means. To begin with, the trade process is categorized as per the various transaction amounts, followed by the optimization of the lightweight transaction process, and testing of the SPIN model of theSET protocol.Eventually, the SET protocol is enhanced as per the outcome of simulation and testing.Addition, validation of the security protocols that select the encryption algorithm and the protocol agreement in combination with the encryption technology to ascertain the presence of the attacker at an increased agreement level.

Ismaili et al. [33]suggested a three faceted (3D) system of security, which includes 3-D Secure and 3D as methods of enhancing the security of e-commerce transaction. On the fundamentals of SSL, SET, 3D security schemes and the specific needs of electronic payment, a safe, secure and effective E-Payment protocol has been devised.This presents an additional level of protection for merchants and cardholders. Customers are requested to key in a separate password after the completion of checkout to ascertain if the person is indeed the legitimate card holder; the validation is carried out directly between the card issuer and cardholderutilizing the security certificate apart from the act of not utilizing the third party (Visa, Master Card). Nevertheless, this paper needs to treat analysis of security and performance as the lynchpin of the suggested protocol. Sfenrianto [42]examined the client intention towards use e-payment scheme. However, they specifically focused in Indonesia.

C. Studies Related to Smart Card

Mobahat [43]debated various protocols regarding validation and cryptography in low-cost RFID and carried out a comparison of the resultant output with a qualitative approach. This plan ensures that there is a chance to unearth similitudes and variations among protocols and corresponding solutions.This permits specialists or executives on the look out for a proper protocol or method that matches their requirements and priorities, could select one out of two among many; to put it in a different way, while investigating various types of attacks launched against schemes, either the administrator or customer using the technology could establish or refer to a comparison to determine the appropriate protocol as per the criteria or priorities; such as the value of data transmitted by RFID transceivers located in airfields or

wireless media, or the potential attacks that could occur as per the location of RFID devices. Nevertheless, during the course of this paper certain issues could not be explored, i.e. the privacy and security properties. However as per the references, the same could be investigated and considered in works and researches in the future.

Madhoun et al. [44]brought out a novel security protocol for payments and transactions through NFC which resolved the security vulnerabilities which were identified in the EMV protocol. This is based on cloud infrastructure to ascertain the authenticity of payment terminals and offer confirmation to smartphones. This ensures mutual verification, non-repudiation among NFC smartphone, NFC payment terminals, reliability, and the maintenance of confidentiality of information of private banking. They successfully examined in depth the accuracy of the protocol through the Scyther authentication tool that offers standardized proof to verify security protocols. They intend to offer a better solution, create a prototype and display its efficiency in in a real scenario.

Pal et al. [45]suggested a model which involves two client oriented features with four system oriented features in general. Further, they empirically evaluated this study and segmented the clients into two groups such as late and early adapters. Also found the features which affect the client intention to this system. The experimental solution shows this approach apparent helpfulness is the two robust interpreters for clients acceptance of NFC-based mobile payment scheme and ease of use. However, this payment system, the security problems with government support were not considered. Additionally, for every sample which took mostly use debit or credit card. In any case, this can turn out to be an inclination component as it is not known how individuals not utilizing such cards will respond to this new framework.

D. Studies Related To Security, Privacy, And Encryption Based Protocol

Ray et al. [14]suggested a protocol that could be termed sanguine wherein the trusted third party is used only when any partyconducts itself in a manner that is inappropriate or aborts prematurely.By utilizing this protocol, a reasonable amount of fairplay is achieved and disputes are resolved in an automatic manner within the boundaries of protocol. This additionally demonstrates the manner in which the function of the third party is spread out across many third parties; this improves the robust nature of the protocol. Additionally, this also indicates the manner in which a payment mechanism needs to be adopted; transacted electronically, offered discretion to the transaction of customers. Moreover, they intended to execute such systems in the future. In order to examine the protocol through conventional software specification and authentication tools such as FDR etc. Specifically, they intend to examine the input by the trusted third party and determine the nature in which the frequency of failure by the third party has a bearing or influence on the performance. This study will assist in identifying methods to manage the protocol in the most effective manner. Finally; they intend to execute the protocol. The intention is to rely on COTS components for implementing. Execution will offer us a divergent view on protocol and may necessitate addressing of new issues.

Koutaet al. [15]discussed various methods using multiple agents which have implemented towards offering security as well as proposed agent-based scheme for e-payment. This approach has a substantial overhead that was initiated through multiple agents. It needs there is no interface between the inventor when the mobile agents as send out. Additionally, this proposed scheme resolve the security threats that prime towards new notes through threshold signature method in mobile agents.

Cui et al. [16]designed a typical E-voting scheme relying on a blind signature or discrete channel of communication channel - it is difficult for them to resolve fresh issues such as a vacant ballot, fraud or cheating, vote collusion, etc. The list signature is an extension of group signature, which includes public detection. Therefore, the suggested scheme can fulfill the aspects of E-voting and will resolve fresh problems in a convenient manner. In comparison with the conventional schemes, it is easier to locate members who give more than a single ballot in the suggested scheme. The results of the experiment display the safety and performance of this method.

Tiwari et al. [17]suggested a solution that uses application-layer security for a wireless payment system that offers end-to-end verification and protection of data among wirelesses J2ME based clients and J2EE based servers. This study proposes a novel protocol for verification of web users based on multifactor authentication approach which is proven to be completely safe and convenient to execute. Also, propose a method for two-way verification protocol to validate both parties. This solution can effectively be executed within the controlled resources of a Java MIDP device, without the need for modifications to the fundamental protocols or wireless network infrastructure. They intend to concentrate on devising a novel and effective method to obtain TIC codes from financial organizations. TIC code installation on the cell phone of the users should also necessarily be a simple task to prevent frequent visits by a user to the bank or financial organizations. Server side TIC maintenance and a mechanism to manage it so as to fulfill the requirements of numerous users should also be put down as necessary work to be undertaken in the future.

Raghuwanshi et al. [18]suggested mathematical model towards validating the integrity of payment method as well as ordered details through online. This was based on the third party confirmation which receipts different messages from merchant and client. Also, integrity was verified. This scheme has the low-cost implementation and simple to use. However, this approach has involved the appraisement of the floating point. Furthermore, it is related to the integer computation or improves the efficiency as well as accurateness of calculation.

Abdellaoui and Pasquet[19]suggested a novel creation that will present itself in electronic payment transactions and termed as a payment service provider; it executes payment interactions for customers and merchants on the part of the private banking network. This novel payment scheme takes

care of the issues of trust and safety. In order to handle this problem ofpayment, the service provider needs to suggest a safe and foolproof solution that will be convenient to coordinate with the web application of the merchant and ensure that the client has a good experience. But, this paper has not concentrated on the methods to integrate the three components or entities in the payment method that includes all other parties - it will be of interest particularly in an innovative payment platform that relies on the combination of many modes of payment (such as credit card, gift card) and the new challenges faced by e-commerce.

Mazumdar and Giri [20]proposed encryption approach for the design of secure protocol using online e- payment system. The token message was rationalized through distributing bank. Further, they verified the sender and user payment details as well as consent of client and mercantile. It offers truthfulness, fairness, secrecy and confidentiality.

Takyi and Gyaase[21]suggested that Robust Electronic Payment Protocol (REPP) fares better when compared theoretically with live cardholder verification in terms of security, usability, verification of cardholder and execution. But, the REPP authenticates the merchant just as SET does. Hence, the results prove that REPP has a better capability to reduce the chances of fraud, easy to use, and convenient to execute in a real world scenario. This suggested protocol could serve as the perfect antithesis to defrauding activities that are being witnessed in e-commerce markets. But, they were debated only with reference to theoretical issues. Additionally, theprotocol will be executed to analyze and prove its strengths.

Lawal et al. [22]discussed Multifactor Authentication approaches for e-payment and banking services of banks. Similarly, assessed the various kind of security scheme as well as a developed measure of security to verify the client retrieved their financial services using online. Also, address the customer awareness program as well as risk-based evaluations were conducted. However, they only focused on marketable banks in Nigeria. Also need to focus towards evaluates the efficiency and reliability of the system using 2-factor authentication.

Khan et al. [5]present the potential for development of e-commerce, analyze it with a fresh protocol termed as "Dual-Network E-Payments Protocol" which was suggested in Pakistan. The protocol relies on a grouping of GSM and IP networks. It fulfills all the desired characteristics of an e-payment mechanism. The protocol is reasonably safe and secure, reliable and appropriate for developing local infrastructure. Dual Network e-Payments Protocol relies on internet & GSM networks. This protocol is devised on the basis of prevailing infrastructure in the country. It offers solutions that are cost effective. It fulfills all the requirements of the security of networks which includes verification, non-repudiation, integrity, and confidentiality. It is a safe online transaction system with a high potential for receiving the trust of customers and merchants.

Aigbe and Akpojaro[23]offer a comprehensive review of the various categories of electronic payment systems in the context of online payment processes, verification mechanism, and types of authentication. The paper proceeds to prove the application of various verification mechanisms and categories of the electronic payments systems that are emphasized. Eventually, the study disclosesthat electronic payment system with verification mechanisms that involve two or more aspects of authentication are inclined to be safer, with minimized chances of being vulnerable to fraud, and augment the confidence of users in utilizing electronic payment systems. Additionally, this paper requires a combination of the verification mechanisms discussed above, specifically, the three-factor verification model – that includes the biometric (finger-vein) to devise an improved algorithm for electronic payment systems where the capability to verify would exceed the prevalent applications for online payment.

Jesudoss et al. [46]suggested Payment Punishment approach with different models towards inspire the actuality expressive during cluster nodes as well as acknowledge the successful data exchange amongst nodes/clusters. Further, they compared the efficiency of this approach with QoS-OLSR protocol. In future, they planned to combine the intrusion avoidance and detection in the occurrence of mischievous nodes. Also, distribute the bandwidth amongst nodes character throughout service delivery.

Kim et al. [24]brought out an effective mutual verification that was based on ID with a crucial agreement protocol by resolving earlier issues; hence it was robust and safe against all identified attacks, specifically in the context of a privilegedinsider attack. They discovered that the suggestion of Islam and Biswas [47] and Qiet al. [48] was bogged down by an issue. Hence they examined areas that considered as contentious and suggested considering our protocol which resolves the issues and provides enhanced security against all recognized attacks. Therefore, the suggested protocol will offer improved security when compared with previous protocols. Additionally, if the study progresses to prove in practical terms the security model, there will arise a possibility to obtain a more significant and worthwhile result.

The present study has reviewed 62 articles, ofwhich 3 were from springer digital library, 17 from IEEE digital library, 12 from Elsevier, 6 from ACM, 7 from the specific journal and international journal and 12 from google scholar (Figure 1). Of these 30 studies that were evaluated the secure mobile payment system, Secure Electronic Transaction protocol, Smart card system, Security, privacy, and encryption based protocol. The pictorial representation of the previous method is discussed in figure 2. From the analysis, most of them have focused towards Security, privacy, and encryption based protocol.

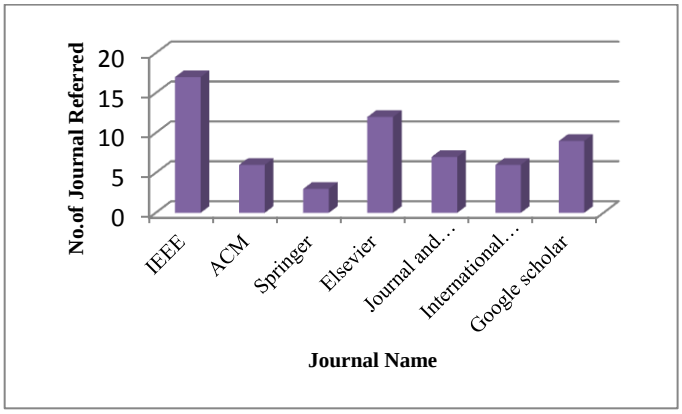


Fig. 1. Pictorial representation of number of research article referred

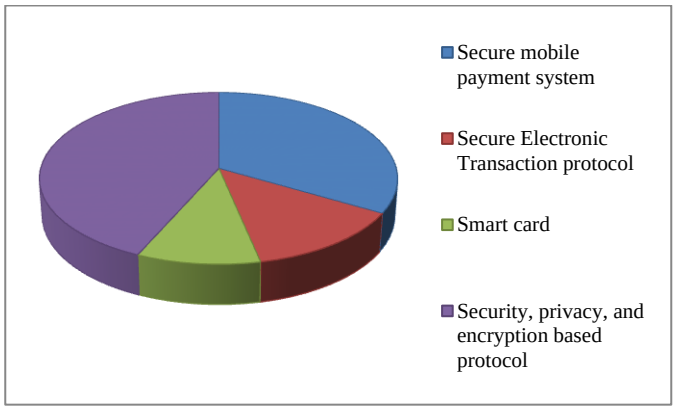


Fig. 2. Pictorial representation of existing method

III. Materials and Methods

A comprehensive literature search of the secure transaction based online payment system was conducted using a database such as Google search, Elsevier, IEEE, and springer digital library. The data searches were restricted to the period amongst 2004to 2016. From this search retrieved articles included terms associated with secure authentication protocol using mobile payment scheme, smart card, security and privacy based encryption. The search strategies applied are provided in the online supplement. Additionally, these reviewed databases examined the credentials of comprised articles with superiorconcentrations on existing related reviews. As seen in this review, the search retrieved a total of 58 potentially suitable articles to fulfilled inclusion criteria for this review. Here comprised unique investigation studies which were evaluated as well as developed transaction system in e-business, including summarization of the authentication and payment protocol. Further, omitted studies encountered following aspects: (1) Summarization of content outside the e-business; (2) merchant and client transaction without e-payment; and (3) not written in English might have lost schemes which précis in other languages.

IV. Results and Discussion

From the literature, the following problems are identified [49].In Bella et al. [50] the authors proved the dual signature method for payment authorization demand suggests the client on the sender side. In any case, this doesn't promise non-denial. Without a doubt, the investigation did by Herreweghen[51] SET does not convey any protected affirmation to the customer. Likewise in Bella et al. [50] assault against SET is depicted which is like their assault including the nearness of a terrible Payment Gateway, who plots with an awful Merchant to hurt the Cardholder. In Kessler and Neumann [52] and Neumann proposed a confirmation rationale amplifying the rationale AUTLOG utilized for demonstrating the responsibility as a part of the electronic trade and afterward utilize that rationale for checking SET and finished up as secure. Bolignano [53] depicted a confirmation strategy for breaking down the installment conventions by a method for evidence in modular rationale. A contextual analysis has been done on C-SET, a variation of SET. In Lu and Smolka[54] proposed a disentangled rendition of SET checked with FDR, a model checker taking into account the dialect CSP. Their examination infers that the improved form is secure. However,Panti et al. [55]proposed two assaults on that variant, in spite of the fact that these assaults can't be performed on SET itself. The security of an e-installment technique is essential for all gatherings required in exchange; however, security alone does not ensure accomplishment in the commercial center. An online payment framework should likewise be advantageous. The openness necessity is by and large ignored through security designers whose point is to make the framework as secure as could be expected under the circumstances. Hu et al. [34], Wang and Varadharajan[32], Wang et al. [38], [39], Kim et al. [24], Ismaili et al. [33], Lomte et al. [30], Wu et al. [56] and Ting et al. [37]proposed secure mobile micropayments for e-payment system. However, the performance of security requirements, modules and threads need to be improved by secure authentication and high strength encryption scheme which could be blocked security issues. Also, need to focus towards improving the system performance. Moreover, some protection and security properties were excluded by explored and could be considered and took after as future inquires about Mobahat [43].

On the other hand, some of them focused on lightweight protocol [25], [41]. However, this based on secure payments through online in a limited situation where direct communication amongst customer and mercantile is not possible. Further, the privacy and security properties were not included [57].Previously few of them have focused towards e-payment technique, yet just a couple are being utilized effectively. CyberCash, depends on card-based payment[57]. The e-payment not an exactly effective method as credit-card methods [11]. SET is another payment-card based protocol [58]. Further, it not designed as particularly for online payment and secure socket layer [59] based e-payment techniques as extensively used. However, a combination of this technique is possible. Few of them focused on RFID based transceivers authentication. However, the privacy and security properties were not included [43]. Previously, several methods show how to design protocols were future found out to have security breaches [60]. Thus the authentication of secure protocols is key as it can recognize defects which lead towards protocol disappointment. So need to verify this model using online based security protocols. Hu et al. [34],Tab and ehjooy and Nazhand[25] and Isaac and Zeadall[27] they used

cryptographic method for the transaction. However, the cryptographic-based encryptiondoes not assurance secure operation of the protocol, even if it is a precise method.

Furthermore, few of them have focused smart card-based secure protocol [44], [61], [62]. However, this study needs to improve the proposed solution towards developing a prototype as well as illustrate its efficiency in real time application. Ray et al. [2005] proposed an optimistic protocol in which the trusted third party is invoked only if any party misbehaves or prematurely aborts.However, this study needs to analyze the protocol through formal software specification and verification tools like FDR. In particular, to study the load at the trusted third party and how the frequency failure at third party affects the performance. On the other hand, focused secure payment system using mathematical mode Raghuwanshi et al. [18]. However, this needs to improve the accuracy and efficiency of the calculation. Abdellaoui and Pasquet[19] proposed online payment service provider that accomplishes payment interactions on behalf of the customer and the merchant on the private banking network side. However, this study does not focus on how to integrate three entities (client, merchant,andPSP) in the payment system that involves other parties.

V. CONCLUSION

In this study explored the safe authentication based online payment system. From the literature, found no one stated the best way towards secure multicast sessions in e-business operational environment as well as securing e-trade payment for multicast services. To examining the accessible of e-payment protocols and limits of their applicability towards online payment system in e-business environment were studied. This protocols and framework are strong substance towards utilizing the unicast e-business platform. In any case, none of this frameworkhas expressed could be applied towards transmitting online payment for multicast e-business environment where the versatility of the foundation will be deliberated as the prime target. To provide the security of client can able to buy the desired items using security methods. This can certify the security of payment system, thus make an incredible solution for e-business by an intermediate in the online payment system. In order to overcome issues of Secure Authentication based e-Payment Protocol, have planned to propose a stylized transaction for online commerce using an intermediary in the e-payment field. This proposed model of intermediary not only settles payments, but it also takes care of such needs as confirming seller and buyer identities, authenticating and verifying ordering and payment information and other transactional requirements lacking in virtual interactions.

- Initially, technical, business and user requirements should be considered for a payment system presenting an interoperable, modular, integrated, and extensible with payment architecture that provides the potentials for deploying security extensions.
- Secondly, according to the specified system requirements, a financial system evaluated and interactions between internal components and external components of the system.
- The third step will identify potentials of the payment model of the system for security enhancement along with preserving system behavior including protocols, services, transactions, and message structure.
- Next, an interface has been designed which is used to interact with the adopted financial system.
- Identifying potentials points of interactions between mobile applications and backbone system in applying security constraints was the starting point to employ security arrangements.
- According to all information related to evaluating system security potentials, security requirement specifications will be determined, so that the system can proceed persistently along with predicted security circumstances.

REFERENCES

[1] J. Tan, K. Tyler, and A. Manica, “Business-to-business adoption of eCommerce in China,” Inf. Manag., vol. 44, no. 3, pp. 332–351, 2007.

[2] R. Kalakota and A. Whinston, Electronic commerce: a manager’s guide. Boston: Addison-Wesley, 1997.

[3] P. M. A. Ribbers and E. V. Heck, “Introducing electronic auction systems in the Dutch flower industry - a comparison of two initiatives,” Wirtschaftsinformatik, vol. 4, no. 3, pp. 223–231, 2004.

[4] K. C. Laudon and C. G. Traver, E-commerce: business, technology, society. London: Addison Wesley, 2002.

[5] W. A. Khan, S. Yousaf, N. A. Mian, and Z. Nawaz, “E-commerce in Pakistan: Growth potentials and e-payment solutions,” in Proceedings - 11th International Conference on Frontiers of Information Technology, FIT 2013, 2013, pp. 247–252.

[6] A. Tiwari, “A Multifactor Security Protocol for Wireless Payment-Secure Web Authentication using Mobile Devices,” Indian Institute of Information Technology, Allahabad, 2007.

[7] H. Gupta and V. K. Sharma, “Role of Multiple Encryption in Secure Electronic Transaction,” Int. J. Netw. Secure. Its Appl., vol. 3, no. 6, pp. 89–96, 2011.

[8] Entrust, “White paper : Enhanced Online Banking Security , Zero Touch Multi-Factor Authentication,” 2016. .

[9] State Services Commission, Guidance on Multi-factor Authentication. Wellington: State Services Commission, 2006.

[10] D. C. Lynch and L. Lundquist, Digital money: the new era of Internet commerce. Chichester: Wiley, 1996.

[11] P. Wayner, Digital cash: commerce on the Net, 2nd ed. London: AP Professional, 1997.

[12] R. Guttmann, Cybercash: the coming era of electronic money. Basingstoke: Palgrave, 2003.

[13] D. Abrazhevich, Electronic Payment Systems: a User-Centered Perspective and Interaction Design. Eindhoven, The Netherlands: Technische Universiteit Eindhoven, 2004.

[14] I. Ray, I. Ray, and N. Natarajan, “An anonymous and failure resilient fair-exchange e-commerce protocol,” Decis. Support Syst., vol. 39, no. 3, pp. 267–292, May 2005.

[15] M. M. Kouta, M. M. Abou Rizka, and A. M. Elmisery, “Secure e-Payment using Multi-agent architecture,” in Proceedings - International Computer Software and Applications Conference, 2006, vol. 2, pp. 315–320.

[16] G. H. Cui, L. Su, M. X. Yang, and Y. Wang, “A secure E-voting system based on list signature for large scale,” in First International Conference on Communications and Networking in China, ChinaCom ’06, 2007, pp. 1–5.

[17] A. Tiwari, S. Sanyal, A. Abraham, S. J. Knapskog, and S. Sanyal, “A Multi-Factor Security Protocol for Wireless Payment - Secure Web Authentication using Mobile Devices,” in International Conference on

Applied Computing Proceedings of the IADIS International Conference on Applied Computing, 2007, pp. 160–167.

[18] S. Raghuwanshi, R. K. Pateria, and R. P. Singh, "A new protocol model for verification of payment order information integrity in online E-payment system," in 2009 World Congress on Nature and Biologically Inspired Computing, NABIC 2009 - Proceedings, 2009, pp. 1665–1668.

[19] R. Abdellaoui and M. Pasquet, "Secure communication for internet payment in heterogeneous networks," in Proceedings - International Conference on Advanced Information Networking and Applications, AINA, 2010, pp. 1085–1092.

[20] A. Mazumdar and D. Giri, "On-line Electronic Payment System using signcryption," Procedia Technol., vol. 6, pp. 930–938, 2012.

[21] A. Takyi and P. O. Gyaase, "Enhancing Security of Online Payments: A Conceptual Model for a Robust E-Payment Protocol for E-Commerce," in Contemporary Research on E-business Technology and Strategy: International Conference, iCETS 2012, 2012, pp. 232–239.

[22] O. B. Lawal, A. Ibitola, and O. B. Longe, "Internet Banking Authentication Methods in Nigeria Commercial Banks," African J. Comput. ICT, vol. 6, no. 1, pp. 208–215, 2013.

[23] P. Aigbe and J. Akpojaro, "Analysis of Security Issues in Electronic Payment Systems," Int. J. Comput. Appl., vol. 108, no. 10, pp. 10–15, 2014.

[24] S. Y. Kim, H. Kim, and D. H. Lee, "An Efficient ID-Based Mutual Authentication Secure against Privileged-Insider Attack," in 2015 5th International Conference on IT Convergence and Security (ICITCS), 2015, pp. 1–4.

[25] A. A. Tabandehjooy and N. Nazhand, "A lightweight and secure protocol for mobile payments via wireless internet in M-commerce," in IC4E 2010 - 2010 International Conference on e-Education, e-Business, e-Management and e-Learning, 2010, pp. 495–498.

[26] A. B. M. R. Sazzad, S. B. Alam, M. N. Sakib, C. Shahnaz, and S. A. Fattah, "Secured cellular banking protocols using virtual internet with digital watermarking," in 2010 IEEE International Symposium on Signal Processing and Information Technology, ISSPIT 2010, 2011, pp. 424–427.

[27] J. T. Isaac and S. Zeadally, "An Anonymous Secure Payment Protocol in a Payment Gateway Centric Model," Procedia Comput. Sci., vol. 10, pp. 758–765, 2012.

[28] E. Kim, Y. S. Lee, S. Y. Lee, J. W. Choi, and M. S. Jung, "A study on the information protection modules for secure mobile payments," in 2013 International Conference on IT Convergence and Security, ICITCS 2013, 2013, pp. 1–2.

[29] S. S. Ahamad, V. N. Sastry, and M. Nair, "A Biometric based Secure Mobile Payment Framework," in Proceedings - 4th IEEE International Conference on Computer and Communication Technology, ICCCT 2013, 2013, pp. 239–246.

[30] V. Lomte, S. Deshmukh, S. Jadhav, and V. Munde, "A Secure M-Payment Protocol for Mobile Devices," Int. J. Emerg. Res. Manag. &Technology, vol. 3, no. 4, pp. 75–79, 2014.

[31] S.-U. Guan, S. . Tan, and F. Hua, "A Modularized Electronic Payment System for Agent-based E-commerce," 2004.

[32] Y. Wang and V. Varadharajan, "A Mobile Autonomous Agent-based Secure Payment Protocol Supporting Multiple Payments," in IEEE/WIC/ACM International Conference on Intelligent Agent Technology, 2006, pp. 88–94.

[33] H. El Ismaili, H. Houmani, and H. Madroumi, "A Secure Electronic Transaction Payment Protocol Design and Implementation," Int. J. Adv. Comput. Sci. Appl., vol. 5, no. 5, pp. 172–180, 2014.

[34] Z.-Y. Hu, Y.-W. Liu, X. Hu, and J.-H. Li, "Anonymous micropayments authentication (AMA) in mobile data networks," in IEEE INFOCOM 2004, 2004, vol. 1, pp. 46–53.

[35] H. Suryotrisongko and B. Setiawan, "A Novel Mobile Payment Scheme based on Secure Quick Response Payment with Minimal Infrastructure for Cooperative Enterprise in Developing Countries," Procedia - Soc. Behav. Sci., vol. 65, no. 1, pp. 906–912, 2012.

[36] A. M. Mwafise and L. Stapleton, "Determinants of user adoption of mobile electronic payment systems for microfinance institutions in developing countries: Case study cameroon," IFAC Proc. Vol., vol. 45, no. 10, pp. 38–43, 2012.

[37] H. Ting, Y. Yacob, L. Liew, and W. M. Lau, "Intention to Use Mobile Payment System: A Case of Developing Market by Ethnicity," Procedia - Soc. Behav. Sci., vol. 224, no. 1, pp. 368–375, 2016.

[38] Y. Wang, K.-L. Tan, and J. Ren, "PumaMart: a parallel and autonomous agents based internet marketplace," Electron. Commer. Res. Appl., vol. 3, no. 3, pp. 294–310, Sep. 2004.

[39] Y. Wang, K.-L. Tan, and J. Ren, "A Study of Building Internet Marketplaces on the Basis of Mobile Agents for Parallel Processing," World Wide Web, vol. 5, no. 1, pp. 41–66, 2002.

[40] D. Lange and O. Mitsuru, Programming and Deploying Java Mobile Agents with Aglets. Boston: Addison-Wesley Pub Co, 1998.

[41] A. Sun, "Optimization Study for Lightweight Set Protocol," in 2012 International Conference on Industrial Control and Electronics Engineering, 2012, pp. 1206–1209.

[42] J. Sfenrianto, "A Model of Factors Influencing Consumer's Intention to Use E-payment System in Indonesia," Procedia Comput. Sci., vol. 59, no. 1, pp. 214–220, 2015.

[43] H. Mobahat, "Authentication and lightweight cryptography in low cost RFID," in ICSTE 2010 - 2010 2nd International Conference on Software Technology and Engineering, Proceedings, 2010, vol. 2, pp. V2-123-V2-129.

[44] N. El Madhoun, F. Guenane, and G. Pujolle, "A cloud-based secure authentication protocol for contactless-NFC payment," in 2015 IEEE 4th International Conference on Cloud Networking, CloudNet 2015, 2015, pp. 328–330.

[45] D. Pal, V. Vanijja, and B. Papasratorn, "An Empirical Analysis towards the Adoption of NFC Mobile Payment System by the End User," Procedia Comput. Sci., vol. 69, no. 1, pp. 13–25, 2015.

[46] A. Jesudoss, S. V. Kasmir Raja, and A. Sulaiman, "Stimulating truth-telling and cooperation among nodes in VANETs through payment and punishment scheme," Ad Hoc Networks, vol. 24, no. PA, pp. 250–253, 2015.

[47] S. H. Islam and G. P. Biswas, "An improved ID-based client authentication with key agreement protocol on ECC for mobile client-server environments," Theor. Appl. Informatics, vol. 24, no. 4, p. 293–312, 2012.

[48] Y. Qi, C. Tang, M. Xu, and B. Guo, "An identity-based mutual authentication with key agreement scheme for mobile client-server environment," in Communications Security Conference (CSC 2014), 2014, 2014, pp. 1–5.

[49] S. Brlek, S. Hamadou, and J. Mullins, "A flaw in the electronic commerce protocol SET," Inf. Process. Lett., vol. 97, no. 3, pp. 104–108, Feb. 2006.

[50] G. Bella, F. Massacci, and L. Paulson, "The verification of an industrial payment protocol: the SET purchase phase," in Proc. 9th ACM Conf. on Computer and Comm. Security, 2002, pp. 12–20.

[51] E. Van Herreweghen, "Non-repudiation in SET: open issues, in: Proc. 4th Conf. on Financial Cryptography," in Lecture Notes in Computer Science, 2001, vol. 1962, pp. 140–156.

[52] V. Kessler and H. Neumann, "A sound logic for analysing electronic commerce protocols," in 5th European Symposium on Research in Computer Security Louvain-la-Neuve, 1998, pp. 345–360.

[53] D. Bolignano, "Towards the formal verification of electronic commerce protocols," in Proceedings 10th Computer Security Foundations Workshop, 1997, pp. 133–146.

[54] S. Lu and S. A. Smolka, "Model checking the secure electronic transaction (SET) protocol," in MASCOTS '99. Proceedings of the Seventh International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems, 1999, pp. 358–364.

[55] M. Panti, L. Spalazzi, S. Tacconi, and S. Valenti, "Automatic verification of security in payment protocols for electronic commerce," in Proc. 4th Internat. Conf. on Enterprise Inform. Systems (ICEIS'02), 2002, pp. 968–974.

[56] J. Wu, C. Liu, and D. Gardner, "A Study of Anonymous Purchasing Based on Mobile Payment System," Procedia Comput. Sci., vol. 83, pp. 685–689, 2016.

[57] A. Levi and C. K. Koc, "CONSEPP: CONvenient and secure electronic payment protocol based on X9.59," in Proceedings - Annual Computer Security Applications Conference, ACSAC, 2001, vol. 2001–Janua, pp. 286–295.

[58] MasterCard, SET Secure Electronic Transaction Specification (Book 1: Business Description). New York (NY): MasterCard Inc., 1997.

[59] A. O. Freier, P. Karlton, and P. C. Kocher, The SSL Protocol Version 3. Mountain View, CA: Netscape Communications Corp., 1996.

[60] S. Muhammad, Z. Furqan, and R. K. Guha, "Understanding the intruder through attacks on cryptographic protocols," in 44th Annual ACM Southeast Conference, ACMSE 2006, 2006, vol. 2006, pp. 667–672.

[61] M. Badra and R. B. Badra, "A Lightweight Security Protocol for NFC-based Mobile Payments," Procedia Comput. Sci., vol. 83, no. 1, pp. 705–711, 2016.

[62] T. T. T. Pham and J. C. Ho, "The effects of product-related, personal-related factors and attractiveness of alternatives on consumer adoption of NFC-based mobile payments," Technol. Soc., vol. 43, no. 1, pp. 159–172, 2015.

Design of Frequency Reconfigurable Multiband Meander Antenna Using Varactor Diode for Wireless Communication

I.ROUISSI
URCRFS, FST
University of Tunis El Manar, 1002
Tunis, Tunisia

J.M.FLOC'H
IETR, INSA Rennes
20 avenue Buttes de coësmes, 35043
Rennes, France

H.TRABELSI
URCRFS, FST
University of Tunis El Manar, 1002
Tunis, Tunisia

***Abstract*—A compact multiband frequency reconfigurable meander antenna proposed for wireless communication systems is designed and described in this paper. A folded structure has been chosen due its good tradeoff between size, bandwidth and efficiency. The reference antenna is based on a meander patch structure radiating at F1=1, F2=1.94, F3=2.6GHz and optimized to be integrated in the Printed Circuit Board (PCB). In order to sweep the resonance frequencies, a chip capacitor was inserted between the meander patch printed on the top layer and a floating ground plane on the back one, in first case. Than in second case, by inserting a varactor diode to tuning electronically the resonance frequencies over wide bands. The measured results agree with simulations and good radiation properties were obtained. Realized prototype and related results are indeed presented and discussed.**

Keywords*—Frequency reconfigurable meander antenna; chip capacitor; Varactor diode; Size reduction; tuning range; wireless communication

I. INTRODUCTION

Wireless communication have been grown strongly and developed significantly in recent years. These new communication systems use multiple standards, e.g. (Radio, TNT, GSM850, GSM1800, GSM1900, WiFi, WLAN, LTE,…), operating on many frequency band for various applications. This evolution opens the field to the development of frequencies reconfigurable antenna. This type of antenna presents a promising alternative to multiband and ultra large band (ULB) because of its ability to adjust its operating frequency band with sometimes-dynamic agility.

Frequency agility can be achieved using several techniques such as PIN diode, varactor diode, MEMS, liquid cristal… Several prototypes have been studied to define frequency reconfigurability. For example, a frequency reconfigurable multiple-input-multiple-output (MIMO) monopole antenna was designed in [1], using a pin diode and a wide frequency tuning ranging from 1.88 to 2.64GHz was achieved. In [2], a compact folded monopole with a transparent dielectric loading was presented. Also, a tuning range from 419 to 883MHz was achieved, by controlling the voltage of varactor diode. A reconfigurable microstrip rectangular loop antenna is discussed in [3]. The physical perimeter of the loop antennas is electronically modified by controlling the states of the MEMS switches. The resonance frequency shifts from 1.16 to 2.08GHz. Reference [4], demonstrated a monopole antenna achieving a frequency tenability from 1.7 to 3.5GHz, by using a movable metallized plate inside a microfluidic channel.

However, many engineers are interested in the electronic components for easy integration, high reliability and small size. The MEMS present some inconvenient including high activation voltage, higher cost and lower reliability. Compared with an MEMS switch, the varactor diode has better reliability, faster switching speed and lower applied voltage. This is the main reason why many researchers and industrial are interested in the varactor diode instead of MEMS switches for frequency tuning.

Compact communication systems working in these standards required a compact size [5], low cost and multiband antenna [6-8]. Meander technologies are the most investigated for the design of reconfigurable structures [9], which allows small size designs, by increasing the electric length of the patch, and wide band operations [10]. Several meander reconfigurable antenna designs were reported in the literature for wireless applications.

A microstrip monopole reconfigurable antenna is presented in [11]. The antenna size of $25\times12mm^2$, operates at two bands (2.3- 2.4GHz) and (5.15- 5.35GHz). The frequency tuning was produced by inserting a PIN diode and a varactor diode into a meander. The PIN diode is used for switching wireless services and the varactor diode is used for tuning agility. In [12], an electrically small antenna (ESA) based on the meander antenna structure is proposed. This antenna with size of $23.5\times43mm^2$ operates in the 800MHz band of LTE. The antenna is of higher size with respect to the other structure.

In [13], a frequency reconfigurable planar monopole antenna for cognitive radio has been designed for wide tuning range. The radiator antenna composed of meandering element and a U-shape was connected together using a PIN diode. Depending on the state of the PIN diode ON or OFF, the antenna resonates eithers at 2.39 and 2.96GHz. More agility from 2.69 to 3GHz was achieved by placing a varactor diode on the radiator antenna. The designs proposed in [11-13] were meander structure, covering high frequency band above 2GHz and, hence are not suitable candidates for low-frequency mobile standards and wireless handheld devices.

In this paper, a frequency reconfigurable meander antenna is presented for wireless communications systems. The structure is based on a meandering line. The frequency agility was achieved by embedding firstly, a chip capacitor then a varactor diode. The advantage given in this structure resides in its small size 12×15mm² with multiband behavior (4 bands). Compared to the existing designs, large tuning range is obtained from 0.7 to 2.87GHz when inserting a single varactor diode. The planar structure and its operation at low-frequency band below 1GHz are the key features of the proposed design. The distinguishing feature of the proposed antenna is that can be easily integrated into a Printed Circuit Board PCB useful for mobile phone.

II. Antenna Geometry

The geometry of the proposed reconfigurable antenna is displayed in Fig.1. It is printed on 150×70mm² commercially available FR4 dielectric substrate with a permittivity of 4.4 and a thickness of 1.6mm. The meander patch was printed on the top layer coupled with a floating ground plane placed on the back layer.

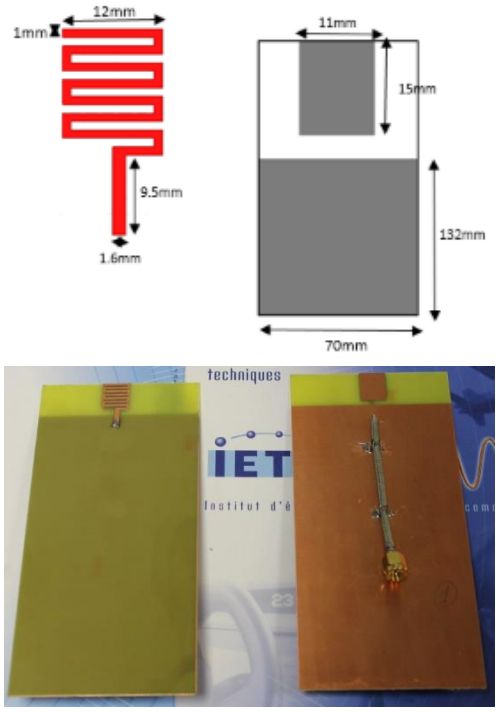


Fig. 1. The meander antenna

Adding a floating ground plane under the meander patch allows changing the effective permittivity of the antenna. Thus, an increase of the electrical length of the meandered line antenna structure and hence results in reducing the overall size of the antenna. In fact, the meander behaves as an inductive structure. The variation in the spacing between the meander lines varies the coupling between them, and therefore the generation of the resonance frequencies. Hence, the multiband behavior.

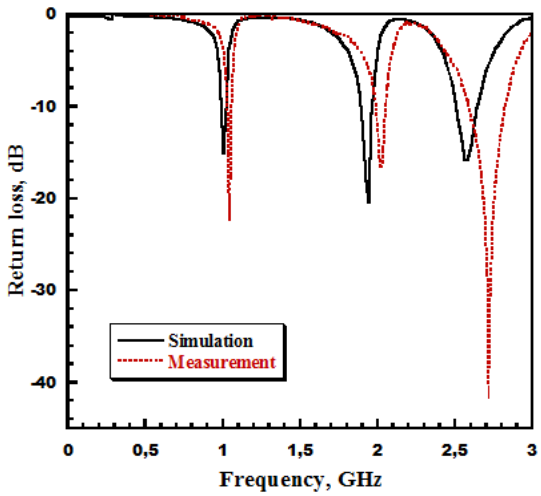


Fig. 2. Simulated and measured return loss S_{11} of the meander antenna

The simulated and measured return loss of the optimized antenna are shown in Fig.2. The planar compact antenna operates in simulation at F_1=1GHz, F_2=1.94GHz, F_3=2.6GHz and in measured results F_1=1.04GHz, F_2=2.01GHz, F_3=2.71GHz. It is noted that good agreement between simulated and measured results are obtained.

3D pattern of the antenna was displayed for the operating frequencies in the 1.04, 2.01 and 2.71GHz, presented in Fig.3.

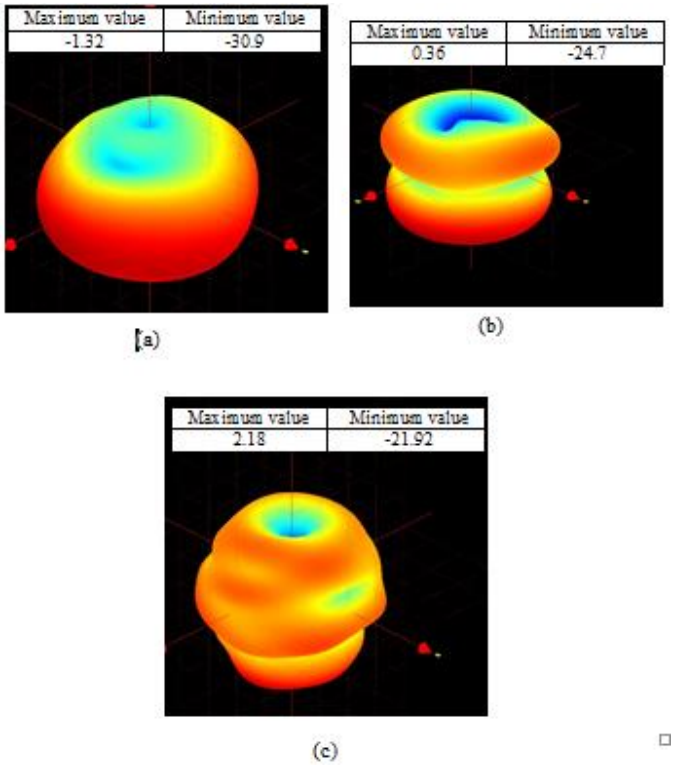


Fig. 3. Measured 3D radiation pattern at: (a) F1=1.04, (b) F2=2.01 and (c) F3=2.71GHz

As shown, this pattern is almost circularly symmetric around x-axis and has a null value for φ=0° at all the resonant frequencies. It can be seen in F_1=1.04GHz that the proposed structure presents omnidirectional radiation pattern in the yz-plane with a peak gain of -1.32dB, confirming the appearance monopoly antenna. In the 2.01GHz and in the xz-plane, the antenna has a peak gain of 0.36dB. The pattern in φ=0° and 90° has a null value. In 2.71GHz, the proposed antenna presents quasi-omnidirectional radiation pattern in the yz-plane with a maximum gain of 2.18GHz.

III. FREQUECNY AGILITY ANTENNA

We are interested in this part, to investigate the frequency agility of the proposed meander antenna. First, the study was done using a single lumped capacitor to identify the suitable position to tune the resonance frequency. Then by embedding a varactor diode to achieve a large tuning frequency range.

A. *Lumed capacitor*

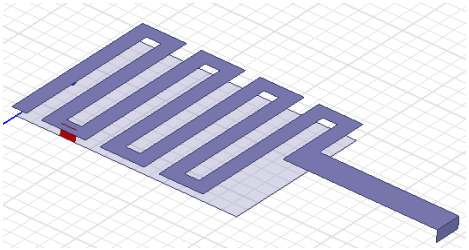


Fig. 4. The meander antenna loaded with a chip capacitor

In order to have a frequency agility property, a chip capacitor was loaded between the meander patch on the top side and the floating ground plane on the backside. The proposed structure was depicted in Fig.4. When the capacitance values varied from 1.3 to 5.2pF, we obtained a tuning range of 100MHz[0.94GHZ-0.84GHz]for the first frequency F_1, 40MHz [1.9GHz-1.86GHz] for F_2, 260MHz [2.46GHz-2.2GHz] for F_3 and 200MHz [2.98GHz-2.78GHz] for F_4. Simulated return loss is presented in Fig.5. Detailed results are summarized in Table I.

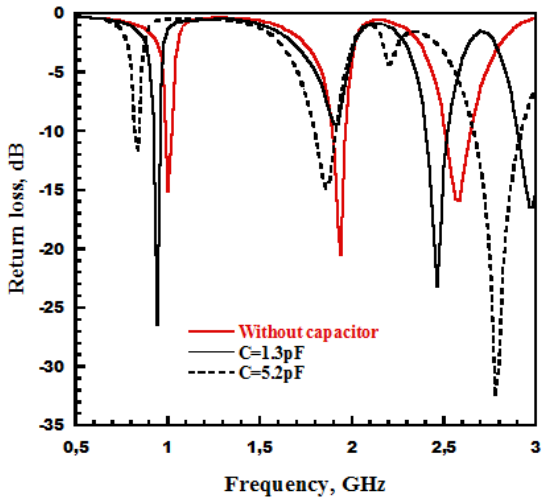


Fig. 5. The meander antenna loaded with a chip capacitor

TABLE. I. SIMULATED RESONANCE FREQUENCY

Capacitor C	F_1(GHz)	F_2(GHz)	F_3(GHz)	F_4(GHz)
Without capacitor	1	1.94	2.6	-
1.3	0.94	1.9	2.46	2.98
5.2	0.84	1.86	2.2	2.78

The current density of the unloaded meander antenna at all the three frequencies concentrated along the feed line and the first meandered line. The strong current density couples the high current on the floating ground plane. The capacitor was inserted at the end meander line where there is a low current density, caused the ability to draw the current, both generated frequencies can be shifted to lower frequencies while increasing the capacitances values. Fig.6 shown the current density of the antenna when the capacitance value is C=5.2pF.

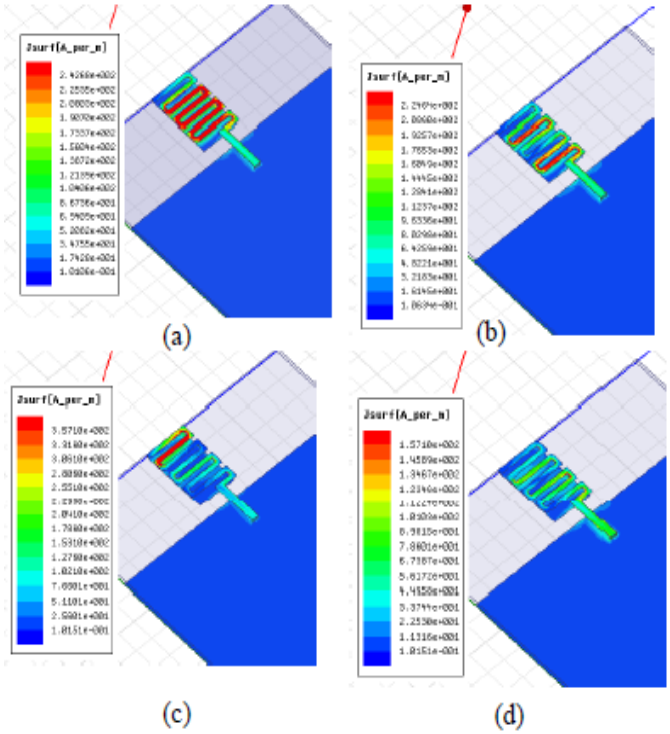


Fig. 6. Simulated current density of the meander antenna for C=5.2pF at: (a) F1= 0.84, (b) F2= 1.86, (c) F3= 2.2 and (d) F4= 2.78GHz

The representation of the current density explains well the radiation property of the meander antenna for all the resonance frequencies when c=5.2pF. 3D radiation pattern has a hallow along the ox-axis. We note that we have the same radiation behavior of unloaded antenna. The simulated obtained gain presented in Fig 7, is about -0.97dBi (0.84GHz), 3.9dBi (1.86GHz), -3.4dBi (2.2GHz) and 4.12dBi (2.78GHz).

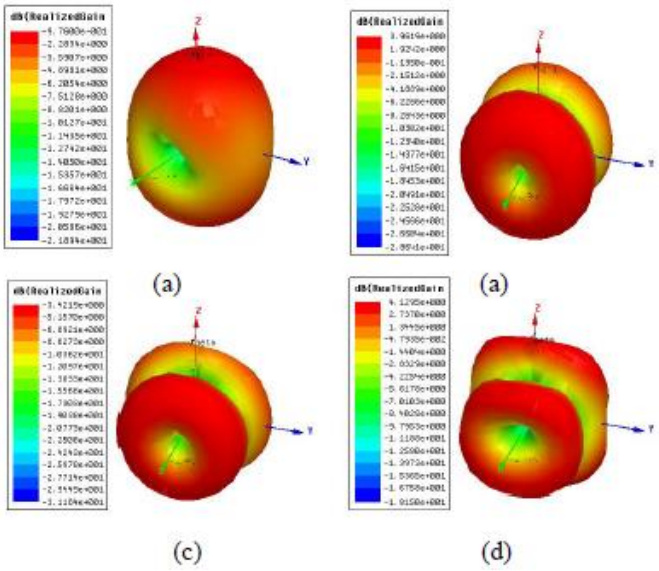


Fig. 7. Simulated 3D radiation pattern of the meander antenna for C=5.2pF at: (a) F_1= 0.84, (b) F_2= 1.86, (c) F_3= 2.2 and (d) F_4= 2.78GHz

B. *Varactor diode*

In order to electronically reconfigure the proposed antenna in a specific bands and achieving a large tuning frequencies range, a BB833 varactor diode from Infineon was replaced the chip capacité between the meander and the ground plane. The diode was reverse biased so requires adding a resistance, referred potential diode. It was placed under the meander, connected to the ground plane. A low pass filter was added to the antenna structure, which aims to eliminate the interference between the continuous and nonlinear. Therefore, eliminate the impact of the polarization of the diode of the antenna radiation behavior. The prototype frequency reconfigurable antenna was shown in Fig.8.

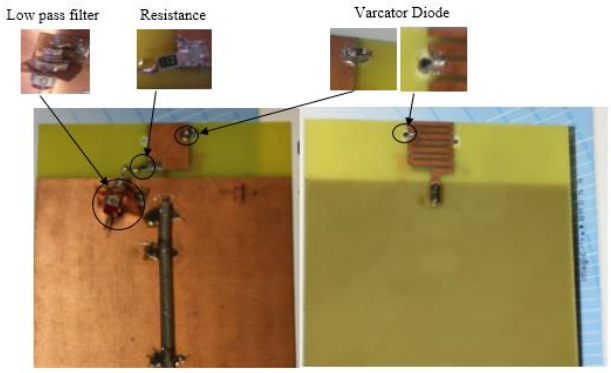


Fig. 8. Prototype of the frequency reconfigurable antenna

The measured return loss of the proposed antenna where the applied reverse voltage V rises from V=0V to V=17V (capacitor value C decreases), are shown in Figs.9. Hence, the effective electric length changes leading to the shift of antenna bandwidth to higher frequencies. We can note that, when applying a reverse voltage from 0 to 8V, only the two frequencies F1 and F2 shift to higher frequencies with a good input impedance matching. Frequencies F3 and F4 stay unchanged. Obtained tuning measured frequencies range from 0.7 to 0.98GHz for F1 and from 1.46 to 2.02 for frequency F2. When the applied voltage increases from 9 to 17V we observed that, the frequency F2 is coincident with F3 fixed at 2.02GHz and F1 stay unchanged with obtained frequency of 0.97GHz. Frequency F4 changes slightly to higher frequencies. A new frequency F5 appears with get tuning measured frequency ranges from 2.23 to 2.47GHz. This slight variation in the resonant frequencies is due to the small variation of the capacitances values of the varactor diode.

We can note that a compact multiband antenna using varactor diode was achieved with an important tuning ranges and good input impedance. All obtained measured frequencies are summarized in Table II.

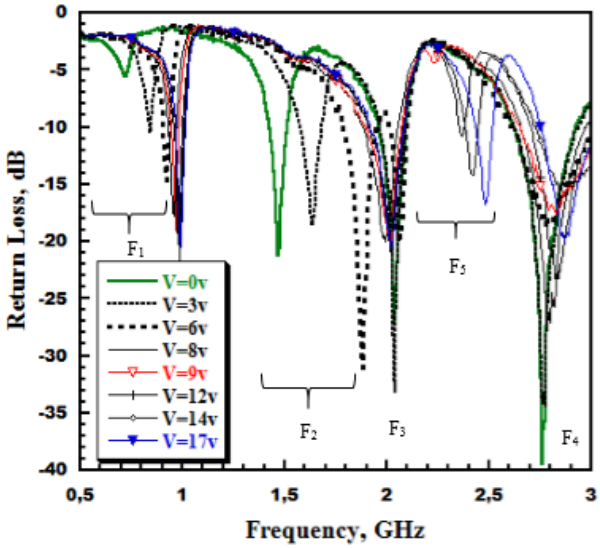


Fig. 9. Measured return loss S11 for V=0 to 17V

A comparison between simulation and measured return loss when V=6V (C=2.7pF) are presented in Fig.10.

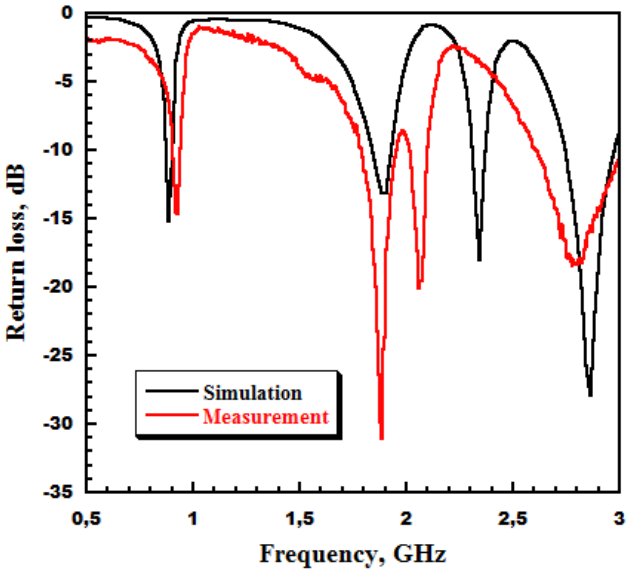


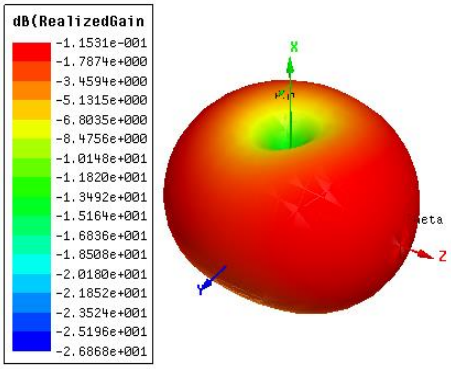
Fig. 10. Simulated and measured return loss S11 of the meander antenna for V=6V

Measured reflection coefficient agrees well with simulated results. The shift found at higher frequencies may be assigned to the fact that the varactor diode model used in simulation was simplified (losses were not taken into consideration). It is noted that the capacity value is that given by the constructor and may be different from the real value.

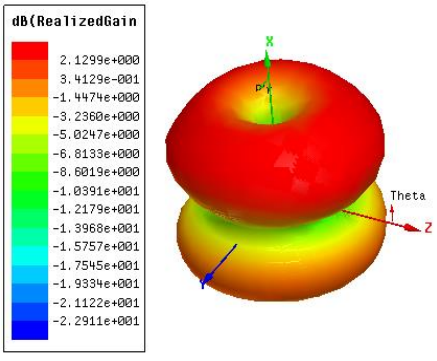
TABLE. II. MEASUREMENT RESONANCE FREQUENCY

V(v)	C (pF)	F_1 (GHz)	F_2 (GHz)	F_3 (GHz)	F_4 (GHz)	F_5 (GHz)
0	-	0.7	1.46	2.03	2.75	-
3	**5.3**	0.84	1.63	2.04	2.77	-
6	**2.7**	0.91	1.88	2.05	2.77	-
8	**1.8**	0.95	1.9	1.98	2.78	-
9	**1.5**	0.97	1.98	2.01	2.81	2.23
12	**0.8**	0.97	2.01	2.02	2.82	2.36
14	**0.76**	0.98	2.02	2.02	2.84	2.4
17	**0.7**	0.98	2.03	2.03	2.87	2.47

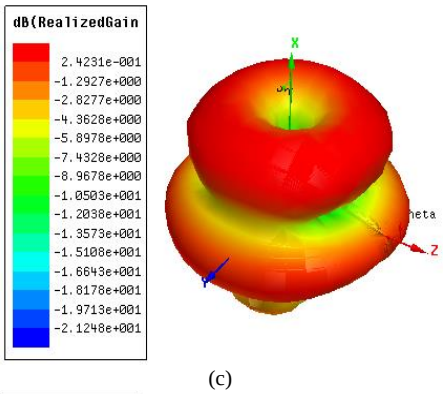
The 3D radiation patterns for the proposed antenna when V=6V, are shown in Figs.11 and 12. There is a good agreement between the simulation and the measured radiation patterns. We note that with the introduction of a varactor diode, we keep the same radiation patterns with a simple structure. Detailed of simulated and measured gains are presented in Table III.



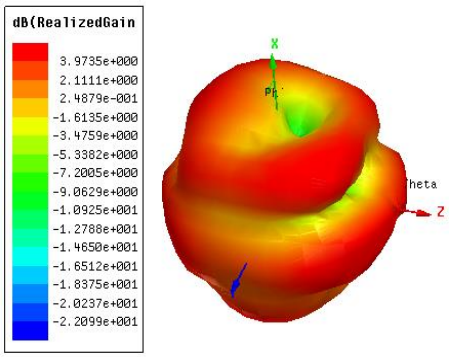
(a)



(b)



(c)



(d)

Fig. 11. Simulated 3D radiation pattern of the meander antenna for V=6V at: (a) F1= 0.88, (b) F2= 1.9, (c) F3= 2.34 and (d) F4= 2.86GHz

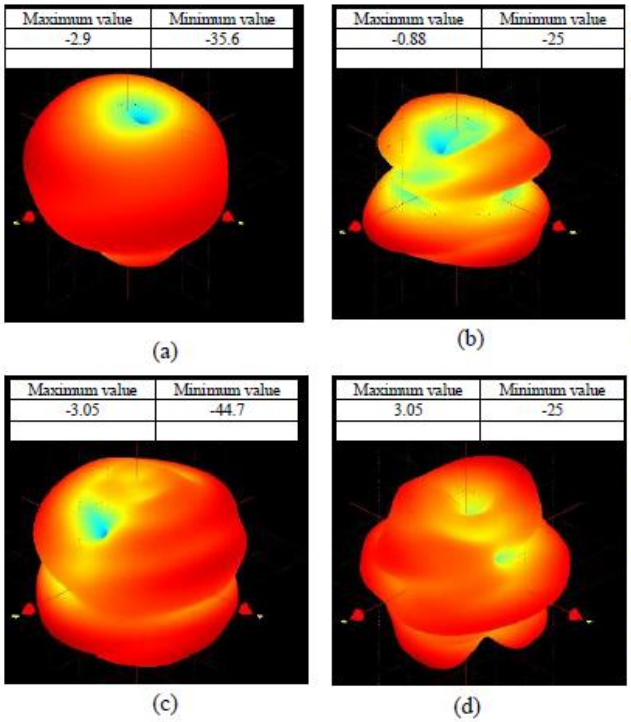


Fig. 12. Simulated 3D radiation pattern of the meander antenna for V=6V at: (a) F1= 0.88, (b) F2= 1.9, (c) F3= 2.34 and (d) F4= 2.86GHz

TABLE. III. SIMULATED AND MEASURED GAIN FOR V=6V

Frequency (GHz)		Gain (dBi)	
Simulated	**Measured**	**Simulated**	**Measured**
0.88	0.91	-0.11	-2.9
1.9	1.88	2.12	-0.88
2.34	2.05	0.24	-3.05
2.86	2.77	3.97	3.05

IV. CONCLUSION

In this paper, frequency reconfigurable meander patch antenna with compact size was studied. The resonance frequency can be controlled by inserting firstly a chip capacitor then by embedding a varactor diode. Wide tuning frequency ranges was achieved with stable radiation properties. The performance of the antenna such as, return loss and 3D radiation pattern are measured. The proposed multiband antenna can be easily integrated into PCB and it seems useful to meet wireless communication system needs. Further studies are planned to improve the performance of the antenna in terms of gain and efficiency, using the micro fluidic structure.

REFERENCES

[1] Y. Cao, S.W. Cheung , and T.I. Yuk, "Frequency reconfigurable multiple-input-multiple-output monopole antenna with wide-continuous tuning range," IET Microwaves, Antennas & Propagation, 2016, Vol. 10, pp. 1322-1331

[2] L. Xing and Y. Huang, "A Transparent Dielectric Loaded Reconfigurable Antenna With a Wide Tuning Range," IEEE Antennas and Wireless Propagation Letters, 2016, Vol. 15, pp. 1630-1633

[3] L. Zhou, S.K. Sharma , and S.K. Kassegne, "Reconfigurable Microstrip Rectangular Loop Antennas Using RF MEMS Swithces," Mirowave and Optical Technology Letters, 2008, Vol. 50, No.1

[4] A. Dey , "Microfluidically Controlled Freaquency Tunable Monopole Antenna for High Power Applications," IEEE Antennas and Wireless Propagation Letters, 2016, Vol. 15, pp. 226-229

[5] H. Chen, C.Y. Sim , C.H. Tsai and C. Kuo, "Compact Circularly Polarized Meandered-Loop Antenna for UHF-Band RFID Tag," IEEE Antennas and Wireless Propagation Letters, 2016, Vol. 15, pp. 1602-1605

[6] P. Bartwal, A.K. Gautam ,A. K. Singh, B.K.Kanaujia and K. Rambabu, "Design of Compact multi-band menader-line antenna for global positionning system/wireless local area network/worldwide interoperability for microwave access band applications in laptops/tablets," IET Microwaves, Antennas & Propagation, 2016, Vol. 10, pp. 1618-1624

[7] A.Khaleghi, "Dual Band Meander Line Antenna for Wireless LAN Communication," IEEE Transactions on Antennas and Propagation, 2007, Vol.55,PP 1004-1009

[8] A. Loutridis, M. John and M.J. Ammann, "Folded menader line antenna for wireless M-Bus in the VHF and UHF bands," Electronics Letters, 2015, Vol.51, No.15, pp. 1138-1140

[9] C.Y.Desmond, T.Y.Han and Y.J.Liao, "A Freaquency Reconfigurable Half Annular Ring Slot Antenna Design," IEEE Transaction on Antennas and Propagation, 2014, Vol.62,PP. 3428-3431

[10] N. Ibrahim, M.Elamin, T.Abdul Rahman, and A.Y. Abdul Rahman, "New Adjustable slot meander patch antenna for 4G handheld devices," IEEE Antennas Wireless Propagation Letters, 2013, Vol. 12, pp. 1077-1080

[11] C.W.Jung, Y.J.Kim, Y.E.Kim and F.De Flaviis, "Macro-micro frequency tuning antenna for reconfigurable wireless communication system," Electronics Letters, 2007, Vol.43, No.4

[12] Mohammad S. Sharawi, Yanal S. Faouri, and Sheikh S. Iqbal, "Design of an electrically small meander antenna for LTE mobile terminals in the 800 MHz band", IEEE GCC Conference and Exhibition, February 19-22, 2011 Dubai

[13] Y.Cao,, S.W. Cheung , X.L.Sun, and T.I.Yuk, "Frequency reconfigurable monopole antenna with wide tuning range for cognitive radio," Mirowave and Optical Technology Letters, 2014, Vol. 56, No.1

Improving the Control Strategy of a Standalone PV Pumping System by Fuzzy Logic Technique

Houssem CHAOUALI (*), Hichem OTHMANI, Dhafer MEZGHANI, Abdelkader MAMI
UR-LAPER, Faculty of Sciences of Tunis,
University of Tunis El Manar
2092 Tunis, Tunisia

***Abstract*—This work aims to develop an accurate model of an existing Photovoltaic Pumping System (PvPS) which is composed of an Ebara Pra-0.50T Asynchronous Moto-Pump (AMP) fed by Kaneka GSA-60 photovoltaic panels via a Moeller DV-51 speed drive. The developed model is then used to compare the performance of the system with its original control strategy based on classical indirect vector control strategy using PI speed controller and the proposed new control strategy based on Fuzzy Logic control technique for speed control and MPPT system. The obtained results of comparative simulations, induced in different dramatic variation of working conditions, show that the developed control strategy brought major enhancements in system performance.**

***Keywords*—Photovoltaic Pumping System; Asynchronous Moto-Pump; PI Speed Controller; Fuzzy Logic Control Technique; MPPT Tracking System; Simulation**

I. INTRODUCTION

Air pollution problems and its disastrous consequences mainly caused by increasing consumption of conventional energy sources such as gas, oil and coal, has encouraged scientific society towards developing environmentally friendly energy sources which are mainly extracted from renewable energy sources such as the sun, water, wind… etc.

Among these new energy sources, and thanks to its continuous technological progress and manufacturing cost reduction, Photovoltaic (PV) energy presents one of best choices from different existing renewable energy sources.

PV technologies are widely used for numerous and various types of applications [1,2]. For some countries, especially where agriculture is an important economic engine such as Thailand [3], pumping water using PV generator (PVG) is a practical solution for rural development where water demand for irrigation and domestic use, is increasing. PV pumping importance is related to the fact that these rural areas are generally without electricity supply sources. [4]

In these cases, different types of PV pumping systems have rapidly replaced traditional pumps such as diesel and gasoline pumps especially that these developed new technologies help to avoid the need for maintenance personnel and fuel supply problems [5]. Despite these different advantages, the generated PV power is strongly dependent on weather conditions especially solar irradiation which contributes constantly in rapid variation of I-V and P-V characteristics of the PVGs [6].

This fact might explain the totally degraded performance that was found in early stage of PV pumping application in comparison with same pumping systems once supplied with a constant voltage source. Trying to fix this problem, later studies suggested that using a DC/DC converter as an adaptation between the load and the source guaranties major improvements in the used motor characteristics as well as the generated power from the PVG [7].

Therefore, researches have been focusing on developing different algorithms and techniques to be used in computing the optimal duty cycle of the DC-DC converter to ensure the better tracking of the Maximum Power Point (MPP) continuously in spite of irradiation variations [8].

In another hand, it is known that different types of disturbances, intern or extern, heavily influence the dynamic performance of several industrial systems such as power systems [9]. In this context, different types of controllers have been introduced over the years. But unlike conventional controllers, Fuzzy Logic (FL) controllers have proved better efficiency in different industrial processes thanks to its robustness. Thus, this reality made us think to apply an upgrade, by using the FL technique, of the existing control strategy of a Photovoltaic pumping system. [10]

This paper gives in a first place a general overview of the system where its different blocks are presented and modeled separately:

- The model of the PV generator which is composed by Kaneka GSA-60 PV panels.
- The model of the 3 phased digital speed drive inverter which is a Moeller DV51 type.
- The model of the 3 phased asynchronous machine and the trained centrifugal pump which both form the studied moto-pump type Ebara Pra-0.50T.

Then, the next section presents is reserved for presenting the actual control strategy and the upgraded one. In this part, we give a general presentation of the Fuzzy Logic technique and how we deployed different FLCs in the system in order to improve its control strategy.

In the last section, we present several results comparing the behaviour of the pumping system with and without the developed FL control strategy and the major proved enhancements in performance.

II. General Overview of the Studied System

A. Description of the studied system

The system we are working on can be simply presented, as shown in Fig. 1, by a 3 phased Asynchronous Moto-Pump (AMP) fed by a DC source, in our case a GPV, via a 3 phased DC-AC inverter speed drive which includes an integrated speed control system.

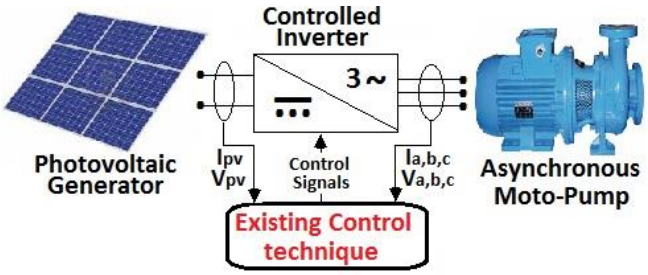


Fig. 1. The block diagram of the studied system

In the next subsections, we are giving a general presentation of every part of this system along with its different features and the different equations used to develop the simulation model.

1) Kaneka GSA-60 Photovoltaic Generator (GPV)

The GPV that our research lab disposes is composed of 5 Kaneka GSA-60 panels, mounted in series, as shown in Fig.2. and situated on the roof of physics department of the Faculty of Sciences of Tunis (El Manar - Tunisisa).

Fig. 2. Kaneka GSA-60 photovoltaic generator

The used model in simulation is developed according to the features given in Table1 of a single PV module type Kaneka GSA-60.

TABLE I. Kaneka GSA-60 Single Module Features

Parameter	Value
Pmpp	60 W
Vmpp	67 V
Impp	0.9 A
Voc	92 V
Isc	1.19 A

Figure 3presents both I-V and P-V characteristic curves of 1 single module, and Figure 4 presents the same curves of the GPV. These curves are obtained by maintaining the temperature at 25°C and varying the solar irradiance in order to show its influence on the generated current.

2) The Moeller 3 phased Inverter

The fact that, the GPV is a DC generator and the Asynchronous machine presents a 3 phased AC load, which requires the use of a 3phased inverter.

In another hand, to ensure the speed control of the machine, consecutively the pumping performance, we use a 3 phased speed drive.

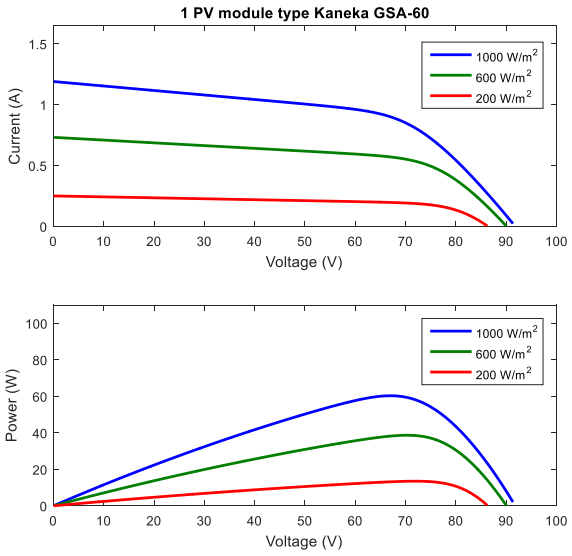


Fig. 3. The influence of solar irradiation variation on I-V and P-V characteristics of a single Kaneka GSA-60 module

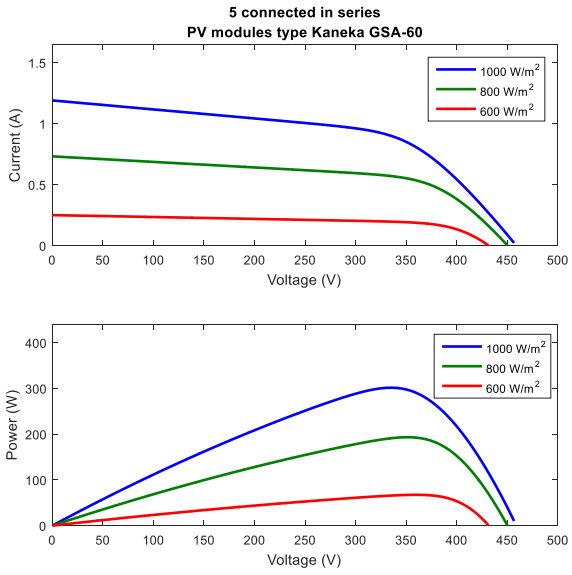


Fig. 4. The influence of solar irradiation variation on I-V and P-V characteristics of 5 Kaneka GSA-60 PV modules connected in series

This speed drive, shown in Figure 5. is a Moeller type DV51-2.2 Kw. His main features are presented in Table 2.

TABLE II. MAIN FEATURES OF THE USED MOELLER INVERTER

Parameter	Value
Max. Power	2.2 kW
Input Voltage	AC : 230V / DC : 400 V
Output Voltage	230 V/ 400 V
Control technique	Vector Control with PI speed regulation

Fig. 5. The 3~ Moeller DV51 Inverter

3) The Asynchronous Moto-Pump (AMP)

The Asynchronous Moto-Pump, shown in Figure 6. is composed of an electrical 3 phased asynchronous motor driving a centrifugal pump.

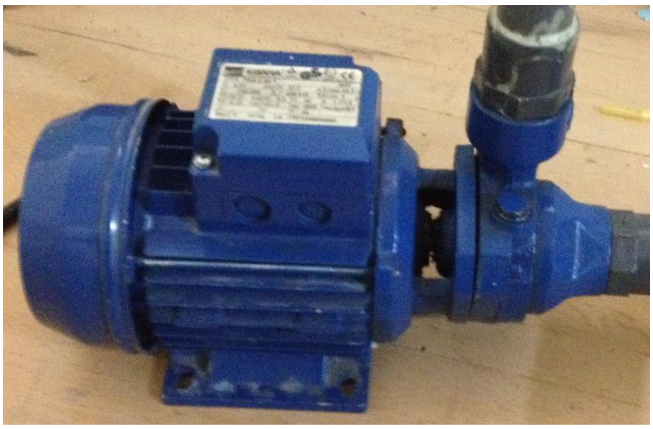


Fig. 6. 3 phased AMP type Ebara-PRA50

The studied AMP is an EBARA type where its reference is PRA50. Its different parameters given by the constructor in the Data Plate are presented by Table 3.

TABLE III. EBARA PRA-0.50T MAIN FEATURES

Parameter	Value
Power	0.37 Kw
Voltage	3~ 240V
Nominal Current	1.8A
Frequency	50 Hz
P	2
Cos ρ	0.8
Maximum Speed	2850 rpm ≈ 300 rad/s
Maximum Flow rate	45 L/mn

B. Modeling the existing system

1) Model of the GPV

In this part, we are presenting the general modeling equations of the GPV and showing the obtained different characteristic curves of the developed model, depending on solar irradiation variation.

Based on the conventional equivalent electrical circuit given in numerous references, we developed the model of our Kaneka generator using (1), (2), (3) and (4). [11,12]

$$I_{pv} = I_{ph} - I_D - I_{Sh} \tag{1}$$

With:

$$I_{ph} = I_{cc}\frac{E}{E_r} + k_{isc}\left(T - T_r\right)\frac{E}{E_r} \tag{2}$$

$$I_D = I_s\left[\exp\left(\frac{V_{pv}}{V_T}\right) - 1\right] \tag{3}$$

$$I_{Sh} = \frac{V_{pv} + R_s I_{pv}}{R_{sh}} \tag{4}$$

2) Model of the 3 phased inverter

Figure 7 presents the equivalent electrical scheme of the power circuit, responsible for the commutation, which we used to develop the simulation model. As shown in the figure and mentioned in numerous references such as [13,14], a 3 phased inverter is generally composed of 3 arms, each one is composed of 2 electronic switches, in this case we used a MOSFET transistor and a parallel diode for each switch.

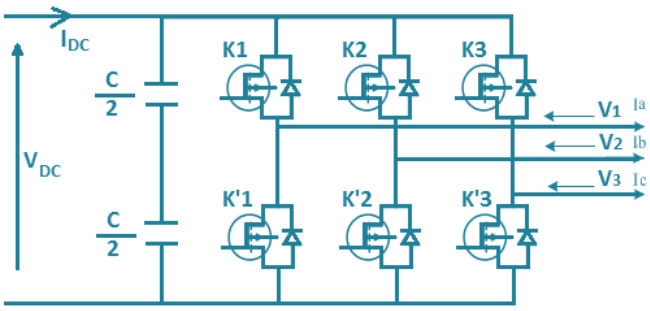


Fig. 7. General equivalent circuit of the 3~ Inverter

Based on this electrical scheme, the different voltages can be expressed by (5) and the relation between the input and the 3 output currents is given by (6). [15]

$$\begin{bmatrix} V_1 \\ V_2 \\ V_3 \end{bmatrix} = \frac{V_{DC}}{3} \begin{bmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{bmatrix} \begin{bmatrix} K_1 \\ K_2 \\ K_3 \end{bmatrix} \quad (5)$$

$$I_{DC} = K_1 I_a + K_2 I_b + K_3 I_c \quad (6)$$

3) Model of the AMP

The next two subsections present each part of the AMP independently and modelling equations are presented.

a) Model of the Asynchronous Motor

Different AM voltages at the stator and the rotor windings of the asynchronous machine can be modelled by equations (10) and (11) as shown in [16]:

$$\left[V_{si}\right] = \left[R_s\right].\left[I_{si}\right] + \frac{d\left[\phi_{si}\right]}{dt} \quad (7)$$

$$\left[V_{ri}\right] = \left[R_r\right].\left[I_{ri}\right] + \frac{d\left[\phi_{ri}\right]}{dt} \quad (8)$$

Where

s refers to stator, r refers to rotor and i refers to the winding.

Also, we can model the flux in both stator and rotor by the system of equations (9).

$$\begin{cases} \left[\phi_s\right] = \left[L_{ss}\right].\left[I_s\right] + \left[L_{sr}\right].\left[I_r\right] \\ \left[\phi_r\right] = \left[L_{rs}\right].\left[I_s\right] + \left[L_{rr}\right].\left[I_r\right] \end{cases} \quad (9)$$

where

$\left[\mathrm{L_{ss}}\right]$ is the matrix of stator inductances and given by (10)

$\left[\mathrm{L_{rr}}\right]$ is the matrix of rotor inductances and given by (11)

$\left[\mathrm{L_{sr}}\right]$ and $\left[\mathrm{L_{rs}}\right]$ are the matrix of mutual inductances between the rotor and stator and given by (12)

$$\left[\mathrm{L_{ss}}\right] = \begin{bmatrix} l_s & M_s & M_s \\ M_s & l_s & M_s \\ M_s & M_s & l_s \end{bmatrix} \quad (10)$$

$$\left[\mathrm{L_{rr}}\right] = \begin{bmatrix} l_r & M_r & M_r \\ M_r & l_r & M_r \\ M_r & M_r & l_r \end{bmatrix} \quad (11)$$

$$\left[L_{sr}\right] = \left[L_{rs}\right]^T = M_{sr}. \begin{bmatrix} \cos(\psi') & \cos\left(\psi' + \frac{2\pi}{3}\right) & \cos\left(\psi' - \frac{2\pi}{3}\right) \\ \cos\left(\psi' - \frac{2\pi}{3}\right) & \cos(\psi') & \cos\left(\psi' + \frac{2\pi}{3}\right) \\ \cos\left(\psi' + \frac{2\pi}{3}\right) & \cos\left(\psi' - \frac{2\pi}{3}\right) & \cos(\psi') \end{bmatrix} \quad (12)$$

with

l_s , l_r : stator and rotor proper inductances.

M_s , M_r : Stator, respectively rotor, mutual inductance between two of its windings.

M_{sr} : maximal mutual Inductance between one winding of the stator and another one of the rotor.

the electrical position of the machine is given by (13)

$$\psi' = p\psi \quad (13)$$

where

ψ : rotor real position (mechanical angle).

ψ' : rotor electrical position.

p : numberof pairs of pole in the machine.

And finally, the mechanical model of the machine can be expressed like shown in equation (14):

$$T_{em} - T_L - f\Omega = J\frac{d\Omega}{dt} \quad (14)$$

where

f : Viscous friction coefficient of the machine.

J : Inertia moment of the rotating masses.

T_{em} : Electromagnetic torque.

T_L : Load torque.

Ω : Rotor speed.

b) Model of the Centrifugal Pump

The centrifugal pump have a proportional resistive torque (C_r) to the square of its rotational speed (Ω). This aerodynamic relation is given by (15) as shown in [17].

$$T_L(\Omega) = C_2\Omega^2 \quad (15)$$

where

C_2 is the torque constant of the pump.

The mechanical losses applied at the AMP shaft are presented by the set of torques $C_{fv}(\Omega)$ *which is* proportional to the speed.

$C_{fv}(\Omega)$ is described by (16) as shown in [18].

$$C_{fv} = C_1\Omega \quad (16)$$

where

C_1 is the coefficient of viscous friction.

To the previously presented torques, we add the acceleration torque $\frac{Jd\Omega}{dt}$. So, the electromagnetic torque expression given in (14) is now described by (17).

$$T_{em} = C_2\Omega^2 + C_1\Omega + J\frac{d\Omega}{dt} \quad (17)$$

Where

C_1 is the same coefficient f .

III. Proposed Improvements Based on Fuzzy Logic Control Technique

A. General overview on applied F.L controllers on the system

Since 1965, after being proposed by Lotfi Zadeh, Fuzzy logic technique has become very common tool in developing intelligent controllers that have replaced conventional techniques and introduced many improvements in decision making with complex systems. [19]

In this context, we tried to enhance our system by integrating FL controllers, as shown in Figure 8. by upgrading the existing speed control system with an FL controller and adding a controlled DC-DC controller where it's duty cycle is computed by an MPPT algorithm based on FL technique.

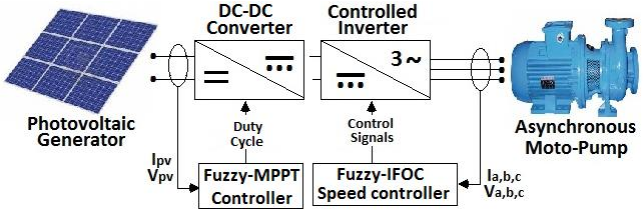


Fig. 8. Block diagram of the new control strategy

Generally, a Fuzzy Logic Controller works by applying 3 main steps on the input data in order to take the decision, the output. Figure 9 presents the working principle of an FLC and its different phases which are: The Fuzzification phase, applying predefined rules by the Inference Engine and finally the defuzzification phase. [20,21]

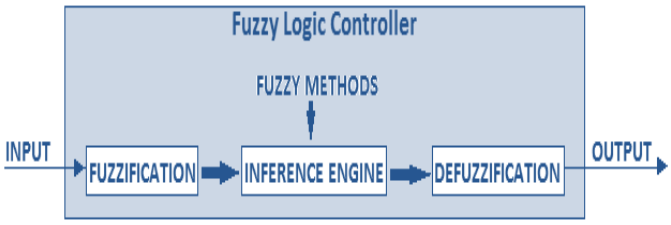


Fig. 9. F.L.C working principle

The first step (Fuzzification) consists of transforming the numeric values on linguistic values that will be treated by the Inference engine (the second phase of FLC) which contains different inference rules, based on predefined fuzzy methods and experimental knowledge, to define a logical connection between the input and output variables by applying. After that, the computed fuzzy value is treated by the last part of the FLC, Defuzzification phase, in order to determine the final numeric value of the output solution.

In our work, we have chosen to use the Mamdani method for the Inference engine step which consists of applying Max-Min relation between the membership degrees of the treated values as explained in [22] and the Gravity Center method, as given in [23], for the defuzzification phase that converts the fuzzy value into numeric value by applying (18).

$$Value_{output} = \frac{\int_i \mu_i . VF_i}{\int_i \mu_i} \quad (18)$$

where:

- μ_i is the membership degree of the fuzzy value.
- VF_i is the fuzzy value.
- $Value_{output}$ is the computed output numeric value.

B. Improvements applied on speed control system

In this part, we present the different blocks we developed for our controller-

For the Fuzzification phase, based on the work presented in [24], we have divided the first input variable "E", which is the error between the speed reference and the measured one as given in (19), in seven fuzzy variables as shown in Figure 10. The second input variable "dE", which is the variation of the error given by (20), in three fuzzy variables as shown in figure 11. Finally, the output variable "Tem*",which is the computed torque reference, in seven fuzzy variables as shown in figure 12. Different figures show that we used a number of mixed trapezoid and triangle sets.

Finally, the different rules defined for the inference engine are presented in Table 4.

$$E(k) = \Omega^* - \Omega \quad (19)$$

$$dE = E(k) - E(k-1) \quad (20)$$

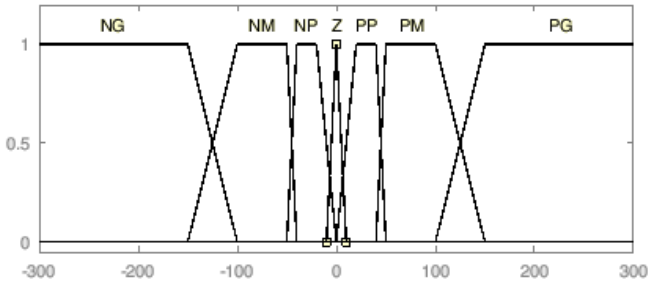


Fig. 10. Membership Functions of the input E

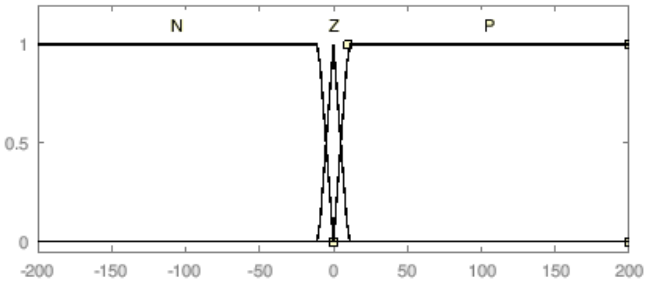


Fig. 11. Membership Functions of the input dE

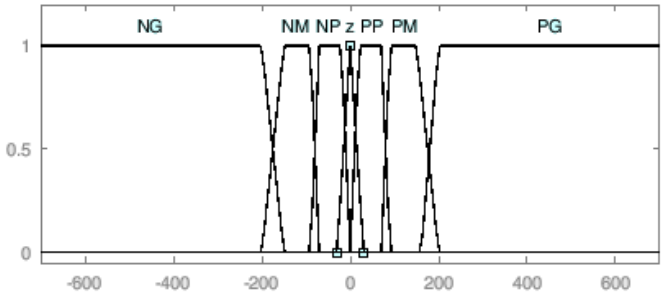


Fig. 12. Membership Functions of the output Tem*

TABLE IV. INFERENCE RULES OF FL SPEED CONTROLLER

E \ dE	N	Z	P
NG	NG	NG	NG
NM	NG	NM	NG
NP	NM	NP	NM
Z	NP	Z	PP
PP	PM	PP	PM
PM	PG	PM	PG
PG	PG	PG	PG

C. Developed Fuzzy-MPPT system

1) The DC-DC converter

As explained in [25], a boost converter is chosen based on its characteristics allowing the required 340 V minimum as DC input of the inverter to obtain the needed 240 V in its AC output and also allowing an almost permanent tracking of the desired MPP.

Figure 13 presents the equivalent electrical circuit of the BOOST DC-DC converter that we have adopted for our work.

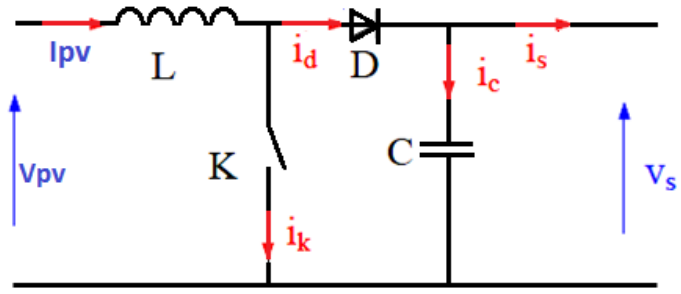


Fig. 13. General structure of BOOST converter

Using the BOOST Converter classical structure, we have modeled our converter based on the model given by (21) and (22).

$$I_S = (1-\alpha) I_{PV} \tag{21}$$

$$V_S = \frac{V_{PV}}{(1-\alpha)} \tag{22}$$

2) FL-MPPT Control algorithm

We fixed two input variables basing on the work given in [26]. These two variables are E, given by (23), and dE, given by (24).

$$E(k) = \frac{Ppv(k) - Ppv(k-1)}{Vpv(k) - Vpv(k-1)} \tag{23}$$

$$dE = E(k) - E(k-1) \tag{24}$$

For the fuzzification, we have divided the first input variable E in three fuzzy variables, the second input variable dE in two fuzzy variables and the output variable which is the computed duty cycle variable that we named D, in seven fuzzy variables.

The different defined membership functions are given respectively by Figures 14, 15 and 16. As shown, we used a combination of trapezoid and triangle functions for these fuzzy sets. And finally, the different rules defined for the inference engine are presented in Table 4.

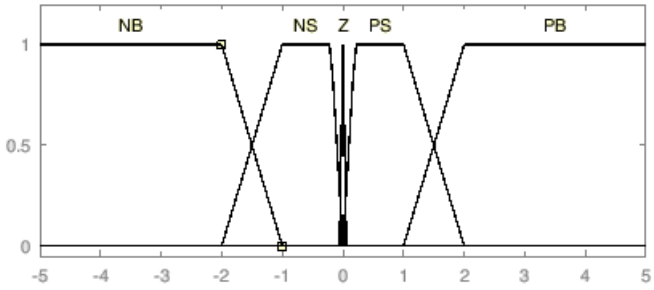


Fig. 14. Membership Functions of the input E

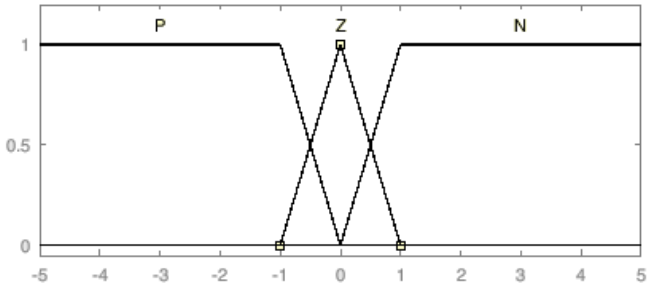


Fig. 15. Membership Functions of the input dE

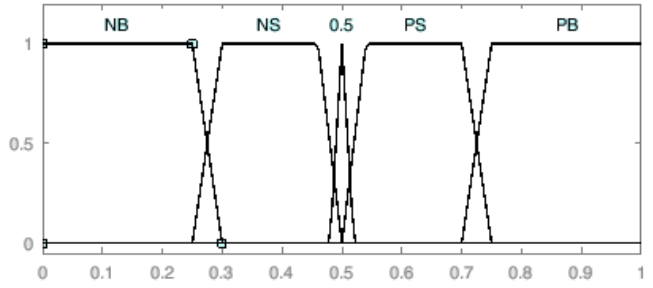


Fig. 16. Membership Functions of the output D

Table 5 presents the nine defined rules for the inference engine. For this step, we conserved the Mamdani method for computing the output fuzzy value.

TABLE V. INFERENCE RULES FOR THE FL-MPPT CONTROLLER

dE E	P	Z	N
NB	**NB**	**NB**	**NB**
NS	**NS**	**NS**	**NS**
Z	**PB**	**0.5**	**NB**
PS	**PS**	**PS**	**PS**
PB	**PB**	**PB**	**PB**

The gravity center method is used for the last operation, the Defuzzification step. And by applying this method on the obtained fuzzy value, the final output of the FLC, the duty cycle D, is finally generated.

IV. RESULTS AND INTERPRETATION

A. Comparing Speed controllers

In a first step, we tried to show the advantage of the added FLC in speed regulation of the system. For that, we simply used a DC voltage source, as previously shown in Figure 9, instead of the PVG and established different scenarios of simulation to compare our AMP performance with the originally preinstalled speed control and the developed one.

These different scenarios are testing in a first place the precision of the regulation process and the speed of response, taking account of both the climbing time and the reference establishing time. Then, we are comparing the robustness of the two regulators by introducing different disturbances on the AMP in the middle of its normal functioning.

The used model of the AMP and the PI speed regulator is given in Table 6.

TABLE VI. AMP AND PI CONTROLLER MODEL PARAMETERS

Parameter	**Value**
Resistance R_s	24.6 Ω
Resistance R_r	16.1 Ω
Inductance L_m	1.46 H
Inductance L_s	0.03 H
Inductance L_r	0.02 H
Pair of poles number	2
Shaft inertia J	$6.5\ 10^{-3}$ Kg m^2
Nominal power	0.37 KW
Voltage	220 V
Frequency	50 Hz
Friction factor F	$1.75\ 10^{-3}$ N.m.s
Kp	13
Ki	1
Flux reference	0.4 Wb

Figures 17,18 and 19 present the different measurements obtained by investigating different working cases given as next:

- Figure 17: a constant speed reference is applied (= 100 rad/s).

- Figure 18: a sudden variable speed reference is applied at t=3s from 80 rad/s to 160 rad/s.

- Figure 19: a constant speed reference is applied and a sudden variation in load torque is introduced at t=2.5s.

Each figure presents different curves where :

- Figures 17.a, 18.a and 19.a: Comparisons between the speed responses with both controllers.

- Figures 17.b, 18.b and 19.b: Comparisons between the variation of the pumped water flow rate.

- Figures 17.c, 18.c and 19.c: Show the generated control signal, which is the Torque reference, by the two controllers.

- Figures 17.d, 18.d and 19.d: show the difference between the measured stator currents with FL and PI controllers.

B. Elaborated FL-MPPT controller

In this part, we are showing the important improvements obtained by both the FL speed controller and the developed fuzzy logic MPPT control algorithm acting through the boost converter.

Table 7 shows the PVG and the boost converter models used in this study. Figure 20 shows the results in term of speed and flow when a constant speed reference (= 150 rad/s) is applied while the irradiance and temperature are maintained constant (Irradiance=1000W/m^2 and Temperature = 25°C). Figure 21 shows the variation of speed and flow while introducing sudden variations in irradiance values:

- At t = 5s: sudden variation of irradiance from 1000W/m^2 to 200W/m^2.

- At t = 10s: sudden variation of irradiance from 200W/m^2 to 800W/m^2.

TABLE VII. PVG AND BOOST MODEL PARAMETERS

GPV Parameters	**Values**
Parallel Strings	1
Series-connected Strings	5
Series Resistance R_s	24.6 Ω
Shunt Resistance R_{sh}	16.1 Ω
Light-Generated Current I_L	1.265 A
Diode Ideality Factor	2.19
Diode Saturation Current I_D	$2.25*10^{-12}$ A
Boost Parameters	**Values**
Inductance L	$0.65*10^{-3}$ H
Capacitance C	$100*10^{-5}$ F
Diode : Ron, Lon, Vf	0.01Ω , 0H, 0.8V

C. Results Discussion

Figures 17, 18 and 19 shows that both controllers pursue perfectly the reference and present a short climbing time (= 0.2s). Meanwhile, the PI controller shows a 10% overshooting while the FL controller shows almost 0% overshooting. In term of robustness, FL controller better reaction towards different disturbances applied on the system while PI controller shows longer time in stabilizing the system after load torque and reference sudden variations.

Figure 20 shows that, without an MPPT regulation system, the speed and consecutively the flow of the AMP are unable to reach the proposed reference and the optimal working performances of the AMP even after more than 5 seconds. By introducing the FL-MPPT controller, the AMP was able to reach the asked working performance in less than 2 seconds. This proves that the performances of the PvPS are enhanced thanks to the additional controller which is also noticed in

Figure 21 that shows the results acquired when using an FL-MPPT regulator are much better than the acquired results without it especially in facing the sudden variation of available sun irradiance.

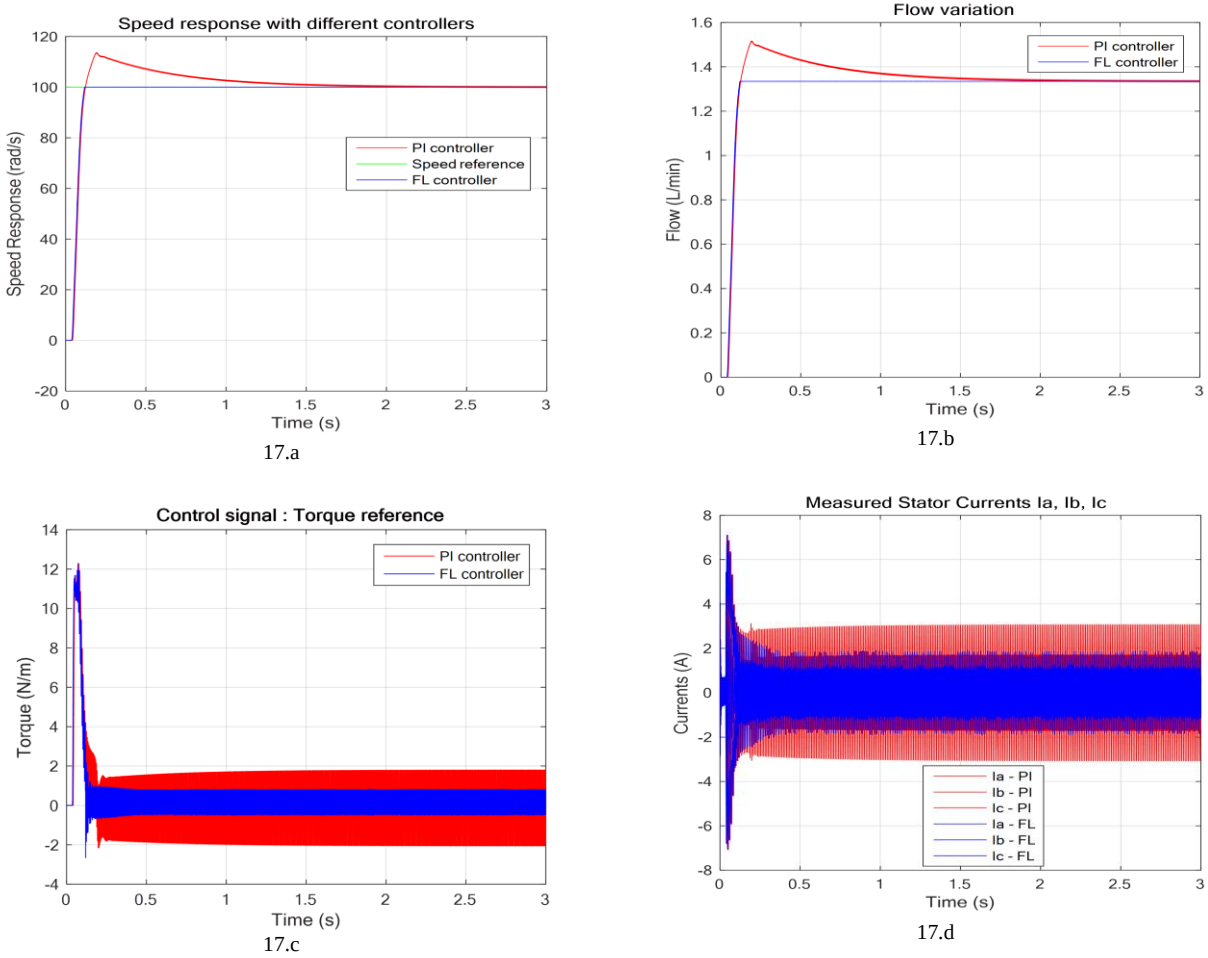
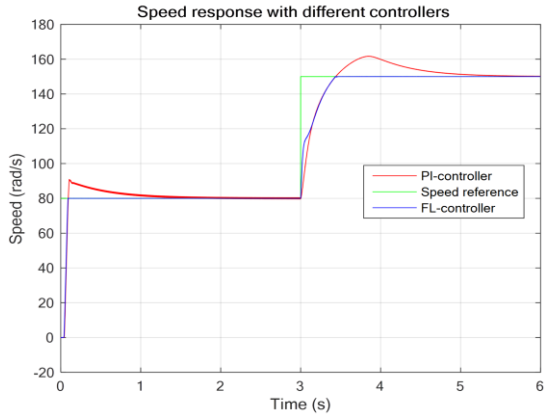
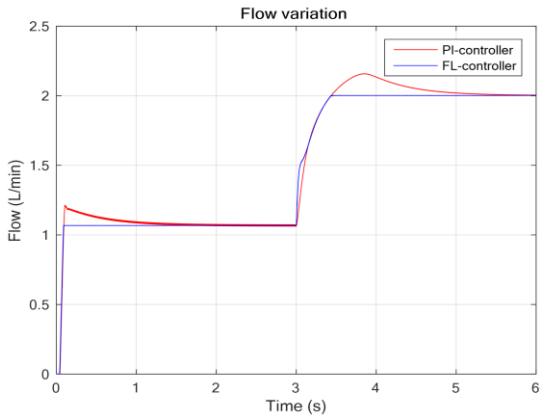


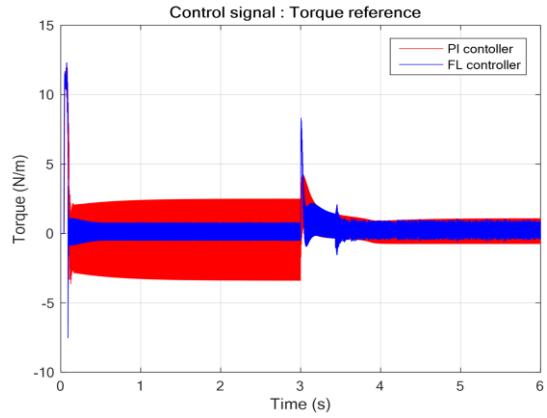
Fig. 17. Different measurements with constant speed reference



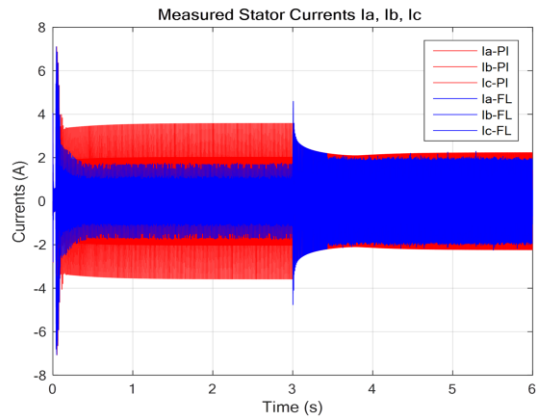
18.a



18.b

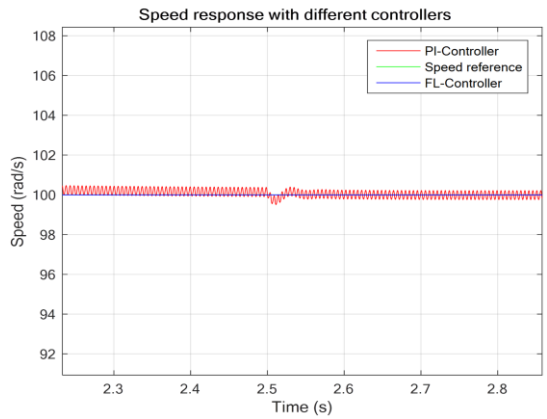


18.c

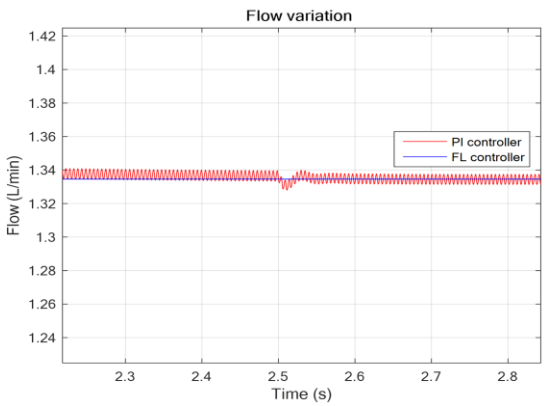


18.d

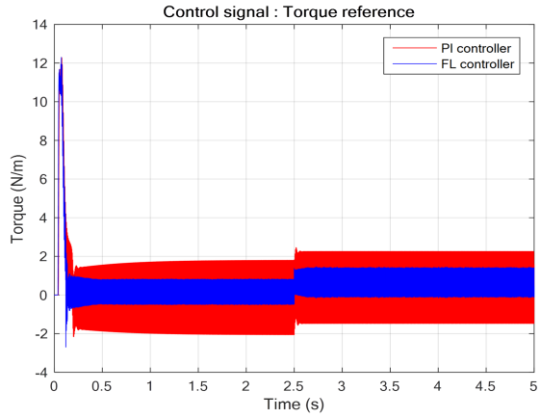
Fig. 18. Different measurements with variable speed reference



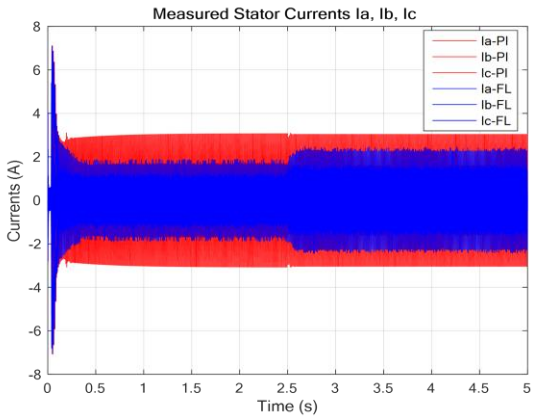
19.a



19.b

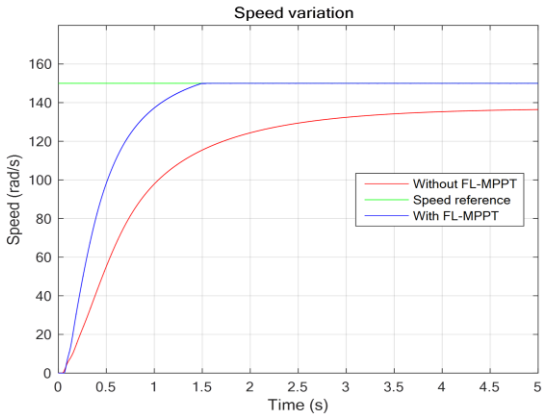


19.c

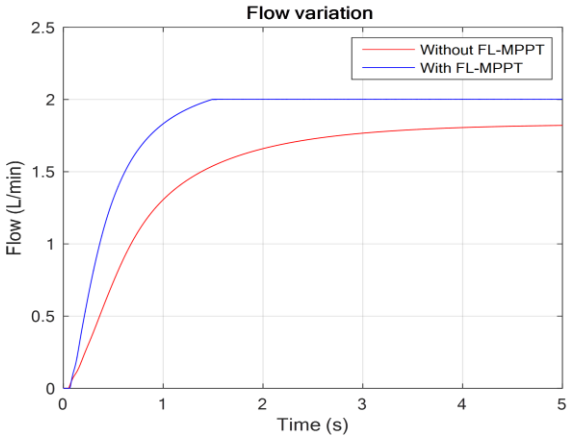


19.d

Fig. 19. Different measurements with variable load torque

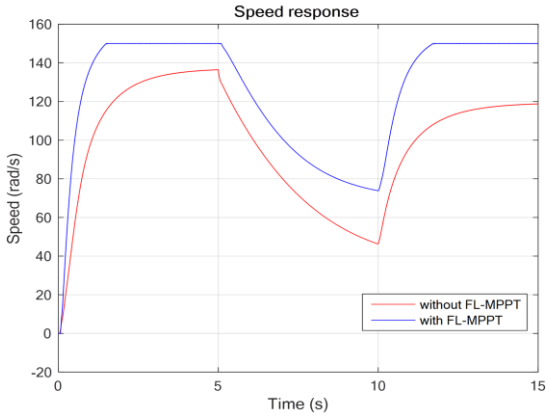


20.a

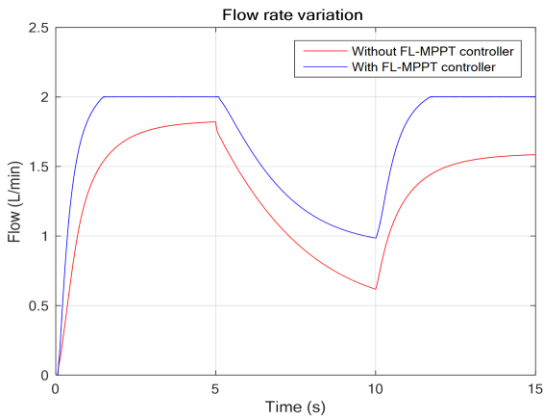


20.b

Fig. 20. Comparing Speed and Flow measurements for a stable solar irradiation



21.a



21.b

Fig. 21. Speed and Flow measurements for sudden variations in solar irradiation

V. CONCLUSION

The different results obtained by simulating the developed speed control system in a first place showed the important advantages of using an FL controller instead of a conventional PI controller in term of time of response, stability and robustness against different disturbances. In a second place, the implementation of a DC-DC converter controlled by an MPPT algorithm based on FL technique shows an upgraded

performance of the PV Pumping system in different scenarios of solar irradiance variations.

This work will be followed by practical investigation by implementing the different algorithms on embedded boards such us FPGAs, STM32s and ARDUINOs.

REFERENCES

[1] E. Román, R. Alonso, P. Ibañez, S. Elorduizapatarietxe, D. Goitia. "Intelligent PV Module for Grid-Connected PV Systems", IEEE Transactions on Industrial Electronics, Vol. 53, Issue 04, pp. 1066-1073, August 2006. [Online] Available: http://doi.org/10.1109/TIE.2006.878327

[2] N. A. Ahmed, A. K. Al-Othman, M. R. Al Rashidi, "Development Of An Efficient Utility Interactive Combined Wind/Photovoltaic/Fuel Cell Power System With MPPT And DC Bus Voltage Regulation", Electric Power Systems Research, Vol. 81, Issue 05, pp. 1096–1106, May 2011. [Online] Available: http://dx.doi.org/10.1016/j.epsr.2010.12.015

[3] V. Vongmanee, "The Photovoltaic Pumping System Using A Variable Speed Single Phase Induction Motor Drive Controlled By Field Oriented Principle", The 2004 IEEE Asia-Pacific Conference on Circuits and Systems, December 6-9, 2004.

[4] A. A. Ghoneim, "Design optimization of photovoltaic powered water pumping systems", Energy Conversion and Management, Vol. 47, Issues 11-12, pp. 1449–1463, July 2006. [Online] Available: http://dx.doi.org/10.1016/j.enconman.2005.08.015

[5] Y. Bakelli, A. H. Arab, B. Azoui, "Optimal sizing of photovoltaic pumping system with water tank storage using LPSP concept", Solar Energy, Vol. 85, Issue 02, pp.288–294, February 2011. [Online] Available: http://dx.doi.org/10.1016/j.solener.2010.11.023

[6] R. Faranda, S. Leva, "Energy Comparison of MPPT Techniques for PV Systems", WSEAS Transactions on Power Systems, Vol. 3, Issue 06, pp. 446-455, June 2008.

[7] A. Terki , A. Moussi, A. Betka, N. Terki, "An Improved Efficiency Of Fuzzy Logic Control Of PMBLDC For PV Pumping System", Applied Mathematical Modelling, Vol. 36, Issue 02, pp. 934–944, March 2012. [Online] Available: http://dx.doi.org/10.1016/j.apm.2011.07.042

[8] G. J. Yu , Y. S. Jung, J. Y. Choi , G. S. Kim, "A Novel Two-Mode MPPT Control Algorithm Based On Comparative Study Of Existing Algorithms", Solar Energy, Vol. 76, Issue 04, pp. 455–463, April 2004. [Online] Available: http://dx.doi.org/10.1016/j.solener.2003.08.038

[9] E. Ǧam , İ. Kocaarslan, "Load Frequency Control In Two Area Power Systems Using Fuzzy Logic Controller", Energy Conversion and Management, Vol. 46, Issue 02, pp. 233–243, January 2005. [Online] Available: http://dx.doi.org/10.1016/j.enconman.2004.02.022

[10] S. P. Binguac, "On The Compatibility Of Adaptive Controllers (Published Conference Proceedings style)," in Proc. 4th Annu. Allerton Conf. Circuits and Systems Theory, New York, 1994, pp. 8–16.

[11] D. Mezghani, H. Othmani, F. Sassi, A. Mami, G. Dauphin-Tanguy, "A New Optimum Frequency Controller of Hybrid Pumping System: Bond Graph Modeling-Simulation and Practice with ARDUINO Board", International Journal of Advanced Computer Science and Applications, Vol. 8, Issue 01, pp. 78-87, 2017. [Online] Available: http://dx.doi.org/10.14569/IJACSA.2017.080112

[12] H. Chaouali, H. Othmani, D. Mezghani, H. Jouini, A. Mami, "Fuzzy logic control scheme for a 3 phased asynchronous machine fed by Kaneka GSA-60 PV panels", 7th International IEEE Renewable Energy Congress (IREC), Tunisia 2016, DOI: https://doi.org/10.1109/IREC.2016.7478893

[13] H. Othmani, F. Sassi, D. Mezghani, A. Mami, "Comparative Study between Fuzzy Logic Control and Sliding Mode Control for Optimizing the Speed Deportment of a Three Phase Induction Motor", International Review of Automatic Control, Vol. 9, Issue 03, pp. 175-181, May 2016. [Online] Available: https://doi.org/10.15866/ireaco.v9i3.9269

[14] D. Mezghanni , R. Andoulsi, A. Mami, G. Dauphin -Tanguy, "Bond Graph Modeling Of A Photovoltaic System Feeding An Induction Motor-Pump", Simulation Modelling Practice and Theory, Vol. 15, Issue 10, pp. 1224-1238, November 2007. [Online] Available: http://dx.doi.org/10.1016/j.simpat.2007.08.003

[15] D. Kairous, B. Belmadani, "Robust Fuzzy-Second Order Sliding Mode based Direct Power Control for Voltage Source Converter", International Journal of Advanced Computer Science and Applications, Vol. 6, Issue 08, pp. 167-175, 2015. [Online] Available: http://dx.doi.org/10.14569/IJACSA.2015.060823

[16] I. Birou, V. Maier, S: Pavel, C. Rusu, "Indirect Vector Control of an Induction Motor with Fuzzy-Logic based Speed Controller", Advances in Electrical and Computer Engineering, Vol. 10, Issue 01, pp. 116-120, February 2010. [Online] Available: https://doi.org/10.4316/AECE.2010.01021

[17] D. Mezghani, M. A. Jaballah, A. Mami, " A new design vector control of pumping photovoltaic system: Tests and measurements", Europeen journal of scientific reseach, vol 61, Issue 04, pp. 493-507, October 2011.

[18] D. Mezghani, A. Mami, " Input-Output Linearizing Control Of Pumping Photovoltaic System: Tests And Measurements By Micro-Controller Stm32", International journal of advances in engineering and thechnology, vol 4, Issue 02, pp. 25-37, September 2012. [Online] Available: http://doi.org/10.7323/ijaet/V4_iss2

[19] N. A. Gounden, S. A. Peter, H. Nallandula, S. Krithiga, "Fuzzy logic controller with MPPT using line-commutated inverter for three-phase grid-connected photovoltaic systems", Renewable Energy, Vol. 34, Issue 03, pp. 909–915, March 2009. [Online] Available: http://dx.doi.org/10.1016/j.renene.2008.05.039

[20] H. Iqbal, M. Babar, "An Approach for Analyzing ISO/IEC 25010 Product Quality Requirements based on Fuzzy Logic and Likert Scale for Decision Support Systems", International Journal of Advanced Computer Science and Applications, Vol. 7, Issue 12, pp. 245-260, 2016. [Online] Available: http://dx.doi.org/10.14569/IJACSA.2016.071232

[21] A. Varghese, J. P. Sreedhar, S. Kolamban, S. Nayaki, "Outcome based Assessment using Fuzzy Logic", International Journal of Advanced Computer Science and Applications, Vol. 8, Issue 01, pp. 103-106, 2017. [Online] Available: http://dx.doi.org/10.14569/IJACSA.2017.080115

[22] F. Culić, D. Matić, B. Dumnić, V. Vasić, "Optimal Fuzzy Controller Tuned by TV-PSO for Induction Motor Speed Control", Advances in Electrical and Computer Engineering, Vol. 11, Issue 01, pp. 49-54, February 2011. [Online] Available: http://doi.org/10.4316/AECE.2011.01008

[23] K. Laaroussi, M. Zelmat, M. Rouff, "Implementation of a Fuzzy Logic System to Tune a PI Controller Applied to an Induction Motor", Advances in Electrical and Computer Engineering, Vol 9, Issue 03, pp. 107-113, October 2009. [Online] Available: http://doi.org/10.4316/AECE.2009.03019

[24] M. Tuna, C. B. Fidan, S. Kocabey, S. Görgülü, "Effective and Reliable Speed Control of Permanent Magnet DC (PMDC) Motor under Variable Loads", Journal of Electrical Engineering and Technology (JEET), Vol. 10, Issue 05, pp. 2170-2178, September 2015. [Online] Available: http://dx.doi.org/10.5370/JEET.2015.10.5.2170

[25] Geoffrey R. Walker and Paul C. Sernia, "Cascaded DC–DC Converter Connection of Photovoltaic Modules", IEEE Transactions on Power Electronics, VOL. 19, Issue 04, pp. 1130-1139, July 2004. [Online] Available: http://doi.org/10.1109/TPEL.2004.830090

[26] Ahmad Al Nabulsi and Rached Dhaouadi , "Efficiency Optimization of a DSP-Based Standalone PV System Using Fuzzy Logic and Dual-MPPT Control", IEEE Transactions On Industrial Informatics, VOL. 8, Issue 03, August 2012. [Online] Available: http://doi.org/10.1109/TII.2012.2192282

Enhanced Security for Data Sharing in Multi Cloud Storage (SDSMC)

Dr. K. Subramanian
Assistant Professor
P.G and Research Department of Computer Science
H.H The Rajah's College
Pudukkottai

F.Leo John
Research Scholar
P.G and Research Department of Computer Science
J.J College of Arts and Science (Autonomous)
Pudukkottai

***Abstract*—Multiple Cloud storage has become one of the essential services of cloud computing. This Multi-Cloud storage models allow users to store sliced encrypted data in various cloud drives. Thus, it provides support for various cloud storage services using the single interface rather than using single cloud storage services. Cloud security goal primarily focuses on issues that relate to information privacy and security aspects of cloud computing. This latest data storage service and data moderation prototype focus on malicious insider's access on stored data, protection from malicious files, removal of centralized distribution of data storage and removal of outdated files or downloaded files frequently. Data owner does not necessarily need to worry about the future of the data stored in the Multi-Cloud server may be extracted or depraved. The other is ingress control of data. The proposed method ensures the file or data cannot get access without the knowledge or permission of the owner. Thus, this research aims at offering an architecture which reduces malicious insiders and file threats with an algorithm that improves data sharing security in Multi- Cloud storage services. This technique will offer a secure environment whereby the data owner can store and retrieve data from Multi-Cloud Environment without file merging conflicts and prevents insider attacks to obtain meaningful information. The experimental results indicate that the suggested model is suitable for decision making process for the data owners in the better adoption of multi-cloud storage service for sharing their information securely.**

Keywords*—Malicious Insiders; privacy; Index based Data slicing; Malicious Files; Multi-Cloud Storage; Data Sharing

I. INTRODUCTION

Multi-Cloud is the utilization of various computing services in a single heterogeneous architecture. Multi- Cloud Storage means the utilization of various cloud storage services using a single web interface rather than the defaults provided by the cloud storage vendors in a single heterogeneous architecture. Multi-Cloud data systems have the capacity to enhance data sharing and this aspect will be significantly of great help to data users. It enables data owners to share their data in the cloud. In any cloud computing model, security is regarded as the most crucial aspect due to the sensitivity and delicacy of the user's information or data stored in a cloud. Presently, every Organization is pushing its IT department to scale up their data sharing systems. Most cloud services are not free and possess different sizes. For instance, Single Cloud Storage falls among the services with storage limitation which makes it disadvantageous in comparison to multi-cloud storage. The main advantage of using multi cloud storage is performance and higher security for data sharing. In the single cloud storage data remains on the centralized storage which can be easily accessed by the malicious insiders. Companies should start considering working with more than one cloud provider at a time - for cost savings, performance, disaster recovery and other reasons. Most business organizations share most of their data with either their clients or suppliers and consider data sharing as a priority [1]. Through data sharing, higher productivity levels are reached. With several users from various organizations contributing to the cloud data, cost and time spent would be less compared to the traditional ways of manually sending and sharing data, which often led to the creation of out-of-date and redundant documents [1].

Although many cryptographic data slicing methods [2], [3], [4] have been proposed as the main problem arises in the insider's access to stored data. Insiders are the trusted secondary admin or managers who maintains the third party server with the same authorization as the admin. Since the third party servers or infrastructure has been used to store any sensitive information. Administrators and third parties manage the infrastructure as they have remote access to the servers; if administrators or third party managers are malicious then they gain access to the user's data. The other threat is unlike the single cloud storage, retrieval of the sliced files from the multi-cloud server is not an easy procedure. In addition, malicious files can be easily uploaded in all the existing approaches in single cloud storage and multi-cloud storage. The lesser focus has been applied in designing the multi cloud architecture when malicious files are uploaded. The only existed solution is the integration antivirus tool from the third party or cloud provider which creates customer to wait for a longer time while uploading the files.

The remainder of the paper is formed as follows. Section 2 describes the overview of the related work in the field. Section 3 discusses the proposed System model. Section 4 describes the overview of architecture, components and its operating

www.multcloud.com
a sample interface for multicloud storage
Need for Multi-Cloud http://www.huffingtonpost.com/young-entrepreneur-council/the-cloud-and-your-busine_b_13751184.html.
Bank Data set File size take n fromhttps://www.chicagofed.org/banking/financial-institution-reports/commercial-bank-data-complete-2001-2010).
Top Threats Group, "The Treacherous 12 Cloud Computing Top Threats in 2016",http://www.cloudsecurityalliance.org

activity with algorithms. Section 5 explains the experimental solutions, and Section 6 Concludes the report and future work.

II. Related Works

Privacy and security for cloud storage are generally a wide area of research. Numerous academic interrogations have been conducted to identify the potential security issues about this subject. It is important to note that sharing files over cloud platform possess numerous vulnerabilities that can lead to unauthorized access. The attackers of cloud have varied intensions or goals which leads to the poor image of the cloud providers once the goal is achieved. In the view of [2] an architecture has been proposed for sharing health care records in multi-cloud storage using Attribute Based Encryption (ABE) and cryptographic secret sharing. Multi-Cloud proxy splits the encrypted record and stores it in the Multi-Cloud. The main drawback in this approaches are group sharing requires huge computation and long waiting time, since file indexing is not used ambiguous information results in file retrieval process. Since the CP-ABE is provided by third party malicious insider may have easy access to the data. File size more than 50 MBs increase the customer's waiting time. The experiments are performed using a highly configured machine hence it is cost consuming in real time. Malicious files are also easily uploaded by the third party authority or role based managers to corrupt the entire scheme. All the tasks are not automated i.e to upload a file client must create a signed medical record using CP-ABE Scheme. Cloud provider's splits the data and transfers data from multi-cloud proxy to cloud data sources.

In order to enhance the secure data sharing in the multi-cloud storage [3] proposed architecture with an Advanced Encryption Standard Algorithm (AES) which seeks to provide better cloud storage decision making for the customers. But insider attacks, colluding attacks, data integrity, data intruder and malicious files have not been focused.

To protect the data from malicious insiders [4] introduced a Secure Data Sharing in Clouds methodology which uses third party server to store a part of the encryption key and other part is maintained by the user. If the revoked user and third party server colludes data can be retrieved from the cloud. Similarly if the malicious cloud admin and third party server colludes data can be retrieved. This method uses single cloud storage and hence centralized distribution of sensitive data is not recommended for the customers. Larger files of 100 MB reduce the performance of this method and makes customer to wait for a longer time since uploading and encryption process are done consecutively.

[5] Introduced a proxy re-encryption scheme for secure data sharing in cloud but private key gets fully exposed when revoked user and proxy colludes. In addition the entire file is stored in single cloud storage which has low security and efficiency.

The reconstruction of data from multi-cloud requires an effective procedure to merge all the files without changing the meaningful information. In [6] very much similar approach has been proposed but does not guarantee the security for Meta table and failed to encrypt the video and other large files. Once the Meta table information is lost, retrieval process will be a tedious work.

In [7] Secure Scalable and Efficient Multi-owner data sharing scheme has been proposed. This scheme integrates Identity Based Encryption and asymmetric group agreement to enable group-oriented access control for data owners in a many-to-many sharing pattern. However the key generation process is carried out by the third party as a separate process and encryption and decryption process is carried out as another process which is burden to the data owner to wait for the completion of the whole process. Malicious files protection has not been guaranteed. Centralized distribution of data storage has not been much promising to the customers to share their data. Identity based encryption supports only small data of 50MB. Key escrow problem arises in Identity based scheme.

The work of [8] introduced a secure file sharing in multi-cloud using Shamir's secret sharing scheme and base 64 encoding in their algorithm. Malicious insider's attacks have been prevented by this scheme. However, indexing of files has not been used so that in the retrieval process receiver has to select all the shares to encode and reconstruct the file which is burden to the receiver. In addition malicious files are not prevented and automation of all the tasks in this scheme has not been focused which reduces the overall efficiency of this scheme.

Many similar approaches has been proposed but failed to implement an effective architecture and working procedure for the secure data sharing using the Multi Cloud storage providers. The existing above approaches does not guarantee the automation of file slicing, encryption, decryption and retrieval process. Existing research also does not focus on the merging file conflicts in the retrieval process, malicious files, colluding provider attacks, insider attacks, removal of centralized distribution of data and key management while sharing the data in Multi-Cloud Storage. Similarly all the existing architectures of single cloud storage and Multi-Cloud Storage follows the same pattern that is file uploading, encryption and slicing without index. If an encryption process is done before slicing very large files or video files cannot be uploaded securely and in addition it may also result to wait the customer for a longer time. Malicious files can also be easily uploaded which causes damages to the multi cloud server in the existing approaches. Further Malicious files [9] are detected in providers environment or by using third parties only after damage is caused. The proposed model is designed in such a way when the malicious files gets uploaded it first affects the owner's machine.

In order to address the above challenges this paper presents an effective architecture framework with a standard algorithm which would enable to enhance the secure data sharing through index based cryptographic data slicing and retrieval of file without file merging conflicts from the Multi Cloud storage. It also ensures the protection of data from malicious insiders and malicious files while uploading the file.

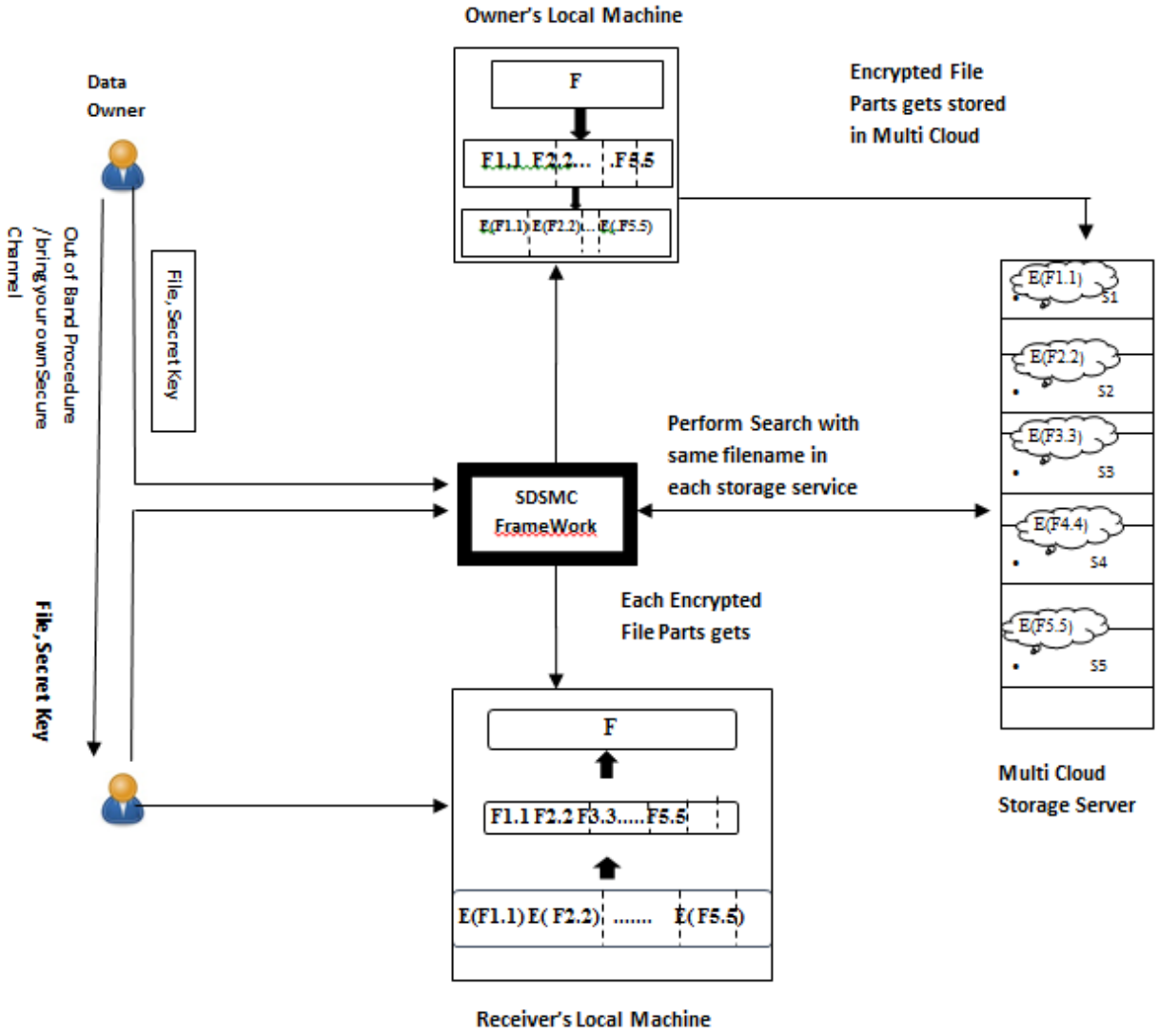


Fig. 1. SDSMC Architecture

III. PROPOSED SYSTEM MODEL

The Overview of Secure Data Sharing Multi Cloud (SDSMC) is shown in Figure-1 and the details are provided in Section 4.

The proposed methodology guarantees the file slicing with index based parts gets encrypted and stored on the Multi-Cloud. This method ensures the file cannot get access without the knowledge or permission of the owner. Data owner uploads the file through the proposed framework interface. The framework uploads the file in the local machine. The framework splits the file with its indexes assigned and encrypts each part of the file using the secret or private key provided by the owner. Each part of the encrypted file gets stored in the owner's machine and then transferred to the multi-cloud server. The receiver sends the decryption request to the owner or the owner can share the required credentials through Bring Your Own Secure Channel (BYOC) or out of band procedure. The receiver enters the credentials through the framework interface. The framework retrieve the file parts and each parts get decrypted, merged and stored the receiver's machine. The major contributions, as described in this report are as follows. The unique feature of this system is to protect the data access from malicious insiders and to protect the datacenters information from malicious files .In addition it also has provision the index based cryptographic data slicing in Multi-Cloud storage services to reduce the file merging conflicts and on demand cost for the customers. It make clients better and fair opportunities for decision making process to choose multi-cloud storage services for secure sharing of data based on trust. The proposed work guarantees that file slicing is based on the number of storage services. More than four cloud storage services are used for confidentiality and none of the Cloud Storage Service Providers can retrieve meaningful information from the pieces of information stored on its servers, without getting some more bits of data from other storage service providers.

In our approach, it is to be presumed that all participating storage cloud service providers, such as Drop Box Google Drive or other CPs, have a common interest securing the infrastructure and data against external, third party

adversaries. Hence the establishment of common and cooperative security mechanisms will be viable, even though many practical and procedural challenges could arise when putting through them in concrete usage scenarios .This work acknowledges those challenges, but consider them out of the scope of our current work.

A. SDSMC Framework

The Secure Data Sharing in Multi Cloud (SDSMC) framework is a web application and it has been described with the overall system flow and various procedures. File uploading, index based file slicing, file encryption, file distribution, file decryption, file retrieval and merging of files, file deletion and Unicode conversion are the automated process performed by the SDSMC framework when using the interface while uploading or downloading a file.

a) ***File Uploading****:* Data owner browse the file from local machine and uploads the file using SDSMC framework interface. This framework uses client resources to upload the file. It means file gets uploaded in the local machine.

b) ***Indexed Based File Slicing:*** This is the process of dividing the uploaded file into two or more parts with respective indices. In this process file slicing is based on the number of storage providers available in the multi-cloud server. At least five storage providers must take part in data sharing and data retrieval process in the proposed approach. This process happens in the owner's local machine.

c) ***File Encryption:*** This is the process of converting a readable file in to unreadable format. This framework encrypts all the index based sliced files using Advanced Encryption Standard (AES) algorithm. Although many existing approaches uses AES it has two draw backs. First it is a weak cipher and the second 128 and 256 bits key make the turnaround time higher which affects the turnaround time process and makes client to wait for a longer time. To overcome the above said limitations slicing is used to make it strong cipher and user defined secret key is used to reduce the turnaround time.

d) ***File Distribution:*** The process of sending the encrypted files along with their indices to different cloud storage providers available in the multi cloud server.

e) ***File Retrieval:*** It is the reversal process of file distribution and file slicing. It is also known as file reconstruction. In this framework the retrieval process starts with submitting the filename without extension. This framework searches the specified filename in each and every cloud storage in multi-cloud server.

f) ***File Decryption:*** Every filename from the multi-cloud server which is associated with specific filename submitted gets decrypted sequentially and stored in the local receiver's machine.

g) ***File Merging****:* This is the process of joining the files with respective indices and gets stored in the receiver's local machine.

h) ***File Deletion****:* This framework performs the automatic removal of files from multi-cloud server and file merging parts in the receiver's machine after the completion of retrieval process.

The idea is about using multiple private clouds simultaneously to deter the risk of disclosure, process tampering and above all, data manipulation in a malicious manner.

IV. Architecture Overview

Figure-1 describes a high level, a standard architecture for a multi-cloud storage service. In the Figure-1 F1.1, F2.2,.. F5.5 denotes the slice file parts name with its index. Similarly E(F1.1),E(F2.2)…E(F5.5) denotes the encrypted sliced parts with its indexing. S1, S2, S3, S4 and S5 are various storage service providers At its core the architecture consists of the following components:

Data Owner: The owner uploads the file with private or secret key. Data Owner acknowledges the request sent by the receiver and sent the details required for the decryption process through the out of band procedure or Bring your own secure channel (BYOSC). In addition the data owner maintains the authorized user's list and keys. Data owner performs the third party duties.

Key Management: There are three options to manage the keys in cloud storage. They are provider's data center, third party server and customer premises. To enhance flexibility and enable sharing of a file to another spacer, it is beneficial to induce the private key at the owner's premise in this approach, as in amazon S3 storage has an enabling option to manage the owner keys.

Multi Cloud Server: It consists of various trusted storage service providers like Cloud A, Cloud B, Cloud C. It stores the encrypted parts of the sliced file from the SDSMC framework to the specific storage service. In this approach minimum five trusted storage service providers are used.

Data Receiver: The receiver will act as a secondary user or sub user. Once the required details are obtained from the owner file can be downloaded.

Owner's Local Machine: All the operations file uploading, indexed slicing and encryption process uses owner's storage device and then encrypted parts are moved towards multi-cloud storage server. This process ensures or guarantee the data owner, uploaded data is highly secured and in addition if malicious files or virus files are uploaded owner's machine will be the priority of those attacks. This is the biggest advantage of our proposed framework and architecture since no additional local server or third parties infrastructure or services are used

Receiver's Local machine: After the successful search operation of the proposed framework, encrypted file parts are downloaded, decrypted and merged in the receiver's device.

TABLE I. NOTATION AND DESCRIPTION

Acronym	Description
F/FN	User's File Name to be uploaded/protected
F.1, F.2..Fn	Sliced parts of the file without encryption
E(F.1),E(F.2)…….E(Fn)	Sliced parts of the Encrypted File
SK	Secret Key

Algorithm-1 SDSMC File Splitting and Encryption

Input: Any file(.xpt, .dicm, video etc.), secret key
Output: Encrypted FilesE (F.1), E (F.2), E (F.3), E (F.4), and E (F.5)

Step 1:
Uploads a file (F) and give user defined secret key (SK)

Step 2:
Find the size of a file (SF)
Step 3:
Slice or Divide the size of a file (SF) by the service providers integrated with Multi Cloud.
Step 4:
Index based files (F.0, F.1, F.2, F.3 and F.4) are created with the same file name and get stored in the owner's local machine.
Step 5:
Pass the user defined secret key (SK) to the Unicode Encoding Object to initialize a key(K) and Vector (IV) which can be used to protect repetition pattern in encrypted files.
Step 6:
Encrypt Each part of the sliced file E (F.1), E (F.2), E (F.3), E (F.4), and E (F.5) from local server and store in the Multi Cloud server.
Step 7:
End

Algorithm -1 explains the application data or file is sliced and transmitted to distinct clouds based on the number of storage services. Files are the most used forms of data storage. The file is uploaded by the user to the Multi Cloud server. The uploaded file gets sliced into five parts with respective indices had been assigned and each part is encrypted using AES encryption algorithm. Five encrypted files are stored in the Multi Cloud Server with respective storage services.

Algorithm-2SDSMC File Decryption and Merging

Input:
File Name without Extension(.xpt, .dicm, video etc.), Secret key (SK)
Output: Decrypted File parts and Merged To get File(F)
Perform:
Step 1:
Get the File Name (FN) and Secret Key (SK) from the data owner or File owner by making request to the processor
Step 2:
Enter or Pass that File Name (FN) and secret Key (SK)
Step 3:
Perform a search with the filename associated in each Multi Cloud storage service provider directory (F.0, F.1, F.2, F.3 and F.4) and obtain the path of the encrypted files E (F.1), E (F.2), E (F.3), E (F.4) and E(F.5).
Step 4:
Pass the user defined secret key (SK) to the Unicode Encoding Object to initialize a key (K) and a vector(V) which can be used to create symmetric Decryptor object.
Step 5:
Merge each part of the decrypted files F1, F2, F3, F4,and F5 from Multi Cloud storage service provider to obtain the original file F.
Step 6:
Auto removal of all decrypted and encrypted parts of the files stored in the respective services.
Step 7:

End

Algorithm -2 describes the reverse process of encryption in which authorized receiver using the framework interface passes the file name and secret key obtained from the data owner. The framework start searching the filename associated in the multi-cloud server and then decrypts the file slices sequentially based on the indices and store the decrypted parts in the receiver's locations and finally merges the file based on indices. The merged file is downloaded at the receivers end. After the retrieval process decrypted and encrypted parts of the files are removed from the multi-cloud server and receiver's machine.

V. IMPLEMENTATION

The Secure Data Sharing in Multi Cloud (SDSMC) methodology is proposed to provide following benefits to the outsourced data:

- Confidentiality and secure distributed data sharing in clouds
- Provide protection from colluding service provider attacks

- Removal of centralized distribution of file storage.
- Automation of all the process such as file uploading, file slicing and indexing, encryption, decryption and merging.
- The file is stored on minimum of five storage service providers
- Self-protection of malicious files
- Insider attackers are not able to retrieve meaningful information.
- Removing of file merging conflicts in the retrieval process

A. Experimental Setup

The proposed methodology involves the creation of five private cloud storage services. There is no federated system is available to evaluate performance of the technique. The proposed Secure Data Sharing in Multi Cloud (SDSMC) methodology has been implemented in Visual Studio 2010 Asp.Net with C#. It consists of two entities Multi Cloud Storage Server and Users. The functionality or procedure required by the user is implemented as a client application that connects with Multi Cloud Server to receive the services. The SDSMC web application splits the uploaded file into n pieces based on number of storage services. Each file part has been assigned with indices and encrypted using Advanced Encryption Standard (AES) algorithm to be stored in the respective storage services. All the cryptographic operations are implemented using .net libraries. File name and secret key management gets rectified when it is maintained at the Data Owner premises. As discussed in section IV when malicious files are uploaded it automatically affects the owner's machine. Once the owner receives the request from the receiver or sub user, owner will send the details through the trusted secure channel or Bring Your Own Secure Channel (BYOSC) or out of band procedure for the decryption process. The receiver decrypts all the parts of the file using the details given by the owner and merges in to a File with meaningful information.

Files or Records can be varied in size and format depending on the data contained, which can be plain text or photographic images or even video files. The file sizes used in the first set of experiments are 52MB, 214MB, 345MB, 437 and 552 MB. The experiments are carried out using the following datasets to evaluate our proposed methodology. They are YouTube datasets for video files, Statistical Analysis System (SAS) Commercial Bank Data files with .xpt format containing the variables currently reported on the Report of Condition and Income plus structure and geographical variables (https://www.chicagofed.org/banking/financial-institution-reports/commercial-bank-data-complete-2001-2010) and .DCIM healthcare image datasets (http://www.osirix-viewer.com/datasets). In our methodology five private cloud storages are used for performance evaluation. Both Data Owner and Private Clouds were operated on a Windows 7 Professional 64 bit machine. The machine uses an Intel® Core (TM) 2 Duo CPU T6500 that runs at 2.10 GHz with 4 GB of DDR3 RAM. Retrieval of meaningful information is not possible for malicious insiders. It ensures the data confidentiality for the Data Owners.

B. Numerical Security Analysis

The high level assessment of this multi-cloud approach is performed on the security features such as privacy, insider attacks, confidentiality, secret keys, and data integrity. Table-2 shows the percentage of security obtained in the proposed SDSMC approach. Three models Cipher Text policy Attribute Based Encryption (CP-ABE), Secure Data Sharing in Clouds (SedaSC) and proposed Secure Data Sharing in Multi-Cloud (SDSMC) are allowed in the private clouds for the specific period of time.

TABLE II. COMPARISON OF SECURITY IN VARIOUS APPROACH

S.No	Security Features	SDSMC	CP-ABE [3]	SeDaSC [5]
1	Privacy	80%	60%	40%
2.	Insider Attacks	100%	80%	80%
3.	Confidentiality	90%	30%	30%
4	Secret Keys	60%	60%	60%
5.	Data Integrity	80%	20%	20%

100% means High secure Data sharing in Multi-Cloud Storage.

1) Security Discussions

a) Privacy

The three models were allowed in the multi- cloud for a specific period. The mode of testing was based on the ability, of at least 5 unauthorized persons to go beyond the first step of accessibility. Single cloud was accessible up to the second step by 3 people. The privacy percentage was obtained as follows:

5persons = 100% lack of privacy

3 =? Therefore; 3/5 x 100 = 60%

100% – 60% = 40%

Hence Single cloud was obtained to be 40% privacy.

Multi Cloud was accessible up to the second step by 2 people.

5=100%

2=? Therefore; 2/5 x 100=40%

100%-40%=60%

Hence Multi-cloud was obtained to be 40% privacy.

SDSMC on the other hand was accessed up to the first step by only one person. Mathematically;

5 = 100%

1 =? Therefore; 1/5 x 100 = 20%

100% - 20% = 80%

SDSMC had privacy percentage of 80%.

b) Insider Attacks

This was tested by intentionally allowing the insiders to be aware of the existence of the model. It was checking on the

discipline of the insider and their intensions. Maximum of ten attacks were considered for a period. SeDaSC and CP-ABE approach was attacked successfully twice, but attacks on SDSMC were not successful. Mathematically Single cloud percentage in this case was as shown below;

10 attacks = 100% insecurity

2 =? Therefore; 2/10 x 100% = 20%

100% – 20% = 80%.

SDSMC had zero attacks hence it had 100%, which is the highest quality.

c) Confidentiality

The quality of this feature depended on the number of persons with the secret keys at the first point of access of each model. SDSMC were only known by one person (owner) while cloud to cloud secret keys were known by three users. The higher number of persons with keys for single and multi-cloud lowered its confidentiality as computed as follows. The model is 100% confidential if no one knows the key. 0.9 represents the value of confidentiality if one person knows the key, therefore;

1 = 100%

0.9 =? Hence 0.9 x 100

= 90% confidentiality for SDSMC

If 0.9 = 1 person, then 3 persons = 1/3 x 0.9 = 0.3

1 = 100%

0.3 =? Therefore; 0.3 x 100 = 30% confidentiality for cloud to cloud

d) Secret Keys

Five people were selected randomly who were to guess the first three consecutive keys. 2 people successfully guessed the first two consecutive digits of SDSMC secret keys of first logging. Single and Multi-cloud also had 2 people. Mathematically this was expressed as shown:

5 = 100%

2 =? Therefore; 2/5 x 100 = 40%

100% – 40% = 60%

e) Data Integrity

Five data were allowed into both models. These were managed for a specific period by technicians of both models. Their integrity was later confirmed in case of any corruption. One of the SDSMC data was slightly altered and single and multi-cloud had 4 of its data altered. Mathematically this was expressed as shown:

5 = 100%

1 =? Therefore; 1/5 x 100 = 20%

100% - 20% = 80% for SDSMC

5 = 100%

4 =? Therefore; 4/5 x 100 = 80%

100% - 80% = 20% for cloud to cloud

From the table-II the proposed SDSMC approach has obtained the highest percentage of security in data sharing when compared with other approaches.

C. Performance Analysis

The results obtained from our technique indicate that all processing steps of our architecture can be accomplished with good performance. However, it's more important data owner's waiting time should be minimal for larger file size (500 MB).Since the current implementation performs all operations in memory CPU processing power and memory resources are also concern in performing this technique. It is therefore favorable to operate the proposed technique in firm Multi Cloud Server Environment.

The first set of experiment is carried out using you tube dataset. Table III shows the time taken to complete entire index based file slicing and merging process for the YouTube dataset. Table-IV shows the turnaround times for encryption and decryption process based on the file size of same You Tube Dataset. It is to be noted that file gets uploaded in the local server before the file slicing process started. File slicing or splitting is the process of dividing the files and creating indices for the files based on the number of storage providers.

TABLE III. TIME TAKEN TO COMPLETE SLICING AND MERGING PROCESS FOR YOUTUBE DATA

S.No	File Type	File Size (MB)	Time for Slicing (SECS)	Time for Merging (SECS)
1	.mp4 video	52	0.281	1
2	.mkv video	214	10	2
3	.mkv video	345	22	5
4	.mkv video	437	28	5
5	.mkv video	550	32	7

File slicing computation time is to be observed because it is done before the encryption process and file merging computation time is done after the decryption process. The slicing and merging time increases gradually with respect to the file size. It is to be noted that merging time is very less when compared to slicing time. This is due to the file uploading time is merged with file slicing time.The slicing involves the operation to evaluate the total file size divided by the number of storage services. It will give the constant file size for each storage services. Based on the constant file size, each part of the file has to be created with indices in the respective storage service.

TABLE IV. TIME TAKEN TO COMPLETE ENCRYPTION AND DECRYPTION PROCESS USING AES FOR YOU TUBE DATASET

S.No	File Type	File Size(MB)	Time for Encryption Process(secs)	Time for Decryption Process(secs)
1	.mp4 video	52	3	4
2	.mkv video	214	10	13
3	.mkv video	345	17	18
4	.mkv video	437	21	28
5	.mkv video	552	29	32

It is observed that proposed algorithm shows that encryption and decryption turnaround time has almost taken the same time to complete their process. The above table -IV also proves that the proposed scheme is well suited for non-organizational outsourced data.

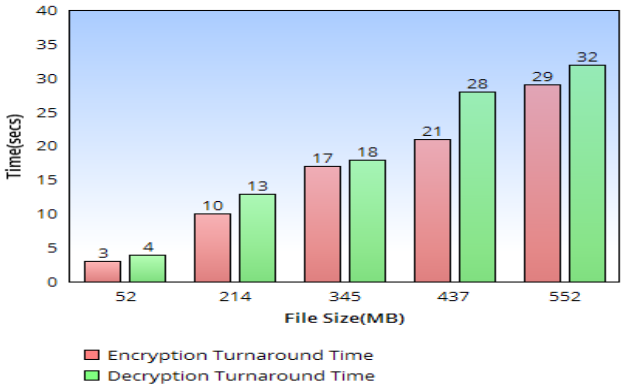


Fig. 2. Encryption and Decryption Turnaround Performance for YouTube Dataset

Above Figure-2 shows the results of YouTube Dataset encryption and decryption turnaround time. Some files have decryption turnaround time more than 7 seconds difference because other process might use the memory resources.The second set of experiments is carried out using commercial bank datasets. The file sizes used are 141,189,234,267 and 337 MB. The same process has been used as in the first set of experiment for the file slicing and merging process. Table-V shows the slice and merging time for bank data. Table -VI shows the encryption and decryption turnaround time for the Bank Data set.Figure-3 shows the results of Encryption Process Time and Decryption Process Time obtained for the Commercial Bank datasets.

TABLE V. TIME TAKEN TO COMPLETE SLICE AND MERGE PROCESS FOR COMMERCIAL BANK DATA

S.No	File Type	File Size (MB)	Slice Time (SECS)	Merge Time (SECS)
1	Call0407.xpt	141	06	02
2	Call0406.xpt	189	10	02
3	Call0209.xpt	234	13	2
4	Call0106.xpt	267	14	3
5	Call0206.xpt	337	24	4

TABLE VI. TIME TAKEN TO COMPLETE ENCRYPTION AND DECRYPTION PROCESS USING AES FOR COMMERCIAL BANK DATASET

S.No	File Type	File Size(MB)	Encryption Time(SECS)	Decryption Time (SECS)
1	Call0407.xpt	141	06	09
2	Call0406.xpt	189	09	12
3	Call0209.xpt	234	11	14
4	Call0106.xpt	267	13	16
5	Call0206.xpt	337	16	20

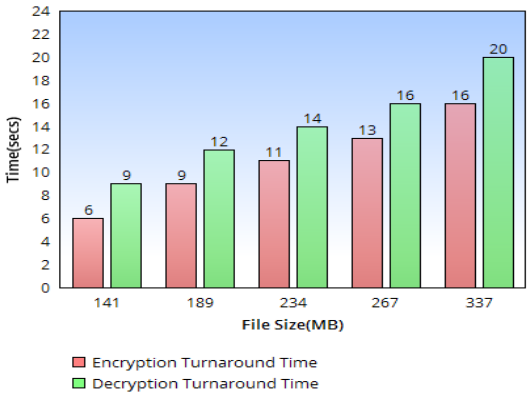


Fig. 3. Encryption and Decryption Turnaround Performance for Bank Dataset

The third set of experiments is carried out using health care data sets. The file here consists of medical records which can be plain text, photographic images or video files. Similar to the first and second experiments the same procedure has been followed in Table-VII and Table-VIII. Table-VII provides the slice and merge time for health care data set. Similarly Table-VIII shows the encryption and decryption turnaround time for healthcare dataset.

TABLE VII. TIME TAKEN TO COMPLETE SLICE AND MERGE PROCESS FOR HEALTH CARE DATASET

S.No	File Type	File Size (MB)	Slice Time (SECS)	Merge Time (SECS)
1	Corstd1.avi	26.3	0.311	0.355
2	Corstd2.avi	36.4	01	0.502
3	Corstd3.avi	79.3	01	01
4	Corstd4.avi	91.3	02	01
5	Corstd5.avi	108	02	01

Above dataset is obtained as DICOM image samples from osirix-viewer.com website. These image samples are converted to .avi files since they are very small in size and used for this research work. The above table-VII shows the various file sizes with slice time and merge time. Whenever file gets sliced indexed is already assigned or in other words file slicing means indexed based file slicing must be assumed throughout this work.

TABLE VIII. TIME TAKEN TO COMPLETE ENCRYPTION AND DECRYPTION PROCESS USING AES FOR HEALTH CARE DATASET

S.No	File Type	File Size (MB)	Encryption Time (SECS)	Decryption Time (SECS)
1	Corstd1.avi	26.3	01	01
2	Corstd2.avi	36.4	02	02
3	Corstd3.avi	79.3	04	05
4	Corstd4.avi	91.3	05	06
5	Corstd5.avi	108	06	07

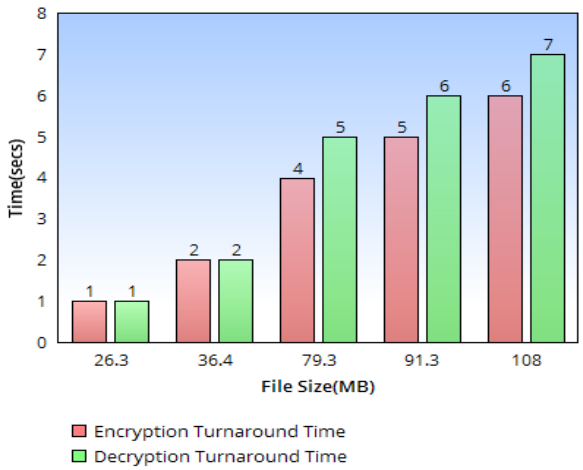


Fig. 4. Encryption and Decryption Turnaround Performance for Health Care Dataset

TABLE IX. COMPARISON OF TURNAROUND TIME WITH DIFFERENT SCHEMES

S.No	File size (MB)	Existing Single and Multi-cloud Storage Schemes						Proposed Scheme	
		[13] CL-PRE		[2]CP-ABE		[5]SeDaSC		SDSMC	
		EPT	DPT	EPT	DPT	EPT	DPT	EPT	DPT
1	1	1	2	0.9	0.9	1	1	0.2	0.2
2	10	13	9	2	2	6	6	1.4	1.6
3	50	53	33	3.4	3.9	9	10	2.4	2.8
4	100	99	57	5.6	5.8	17	20	4	4.8
5	500	369	215	39	40	33	39	26	28.6
6	552	-	-	-	-	-	-	29.2	34.2

EPT-Encryption Process TimeDPT-Decryption Process Time From Table-IX Schemes [13],[2],[5] results are based on single cloud while SDSMC is based on Multi Cloud. The graph has been constructed from the above table for the comparison of Encryption Process Time (EPT) and Decryption Process Time (DPT).

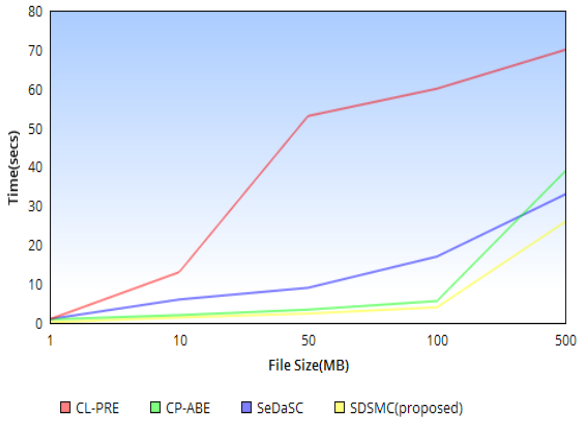


Fig. 5. Comparison of Encryption Process Time

Above figure-5 shows the turnaround performance time of various approaches. It is to be noted that proposed scheme has obtained lesser time seconds for the various file sizes. The consumers waiting time to complete the encryption process has been greatly reduced in the proposed scheme especially for the large file sizes (Mb).

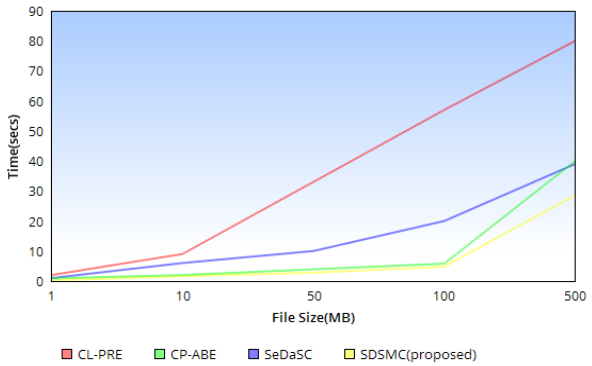


Fig. 6. Comparison of Decryption Process Time

Similarly Figure-6 shows the proposed SDSMC method has far better decryption turnaround time with other existing approaches. In the above table-IX SDSMC column values are obtained from YouTube Dataset, Commercial Bank Dataset and Health care Dataset. The comparison table shows the turnaround times presented in other schemes such as Certificate less Proxy Re-Encryption Scheme (CL-PRE), Cipher TextAttribute Based Encryption Scheme (CP-ABE), Secure Data Sharing in Clouds (SeDaSC) .Although CP-ABE values are very closely related to the proposed approach the share creation and share recovery turnaround times are very high and in addition this scheme uses various software for all the process so automation has not applied as in SDSMC. This scheme (CP-ABE) does not guarantee the malicious insider and file threats and uses high processing machine to obtain the results. Since the files are varied in size and format our methodology supports all types of files which can be used in an organization as well as non-organization for social aspects. Table IX shows the experimental evaluation of existing and the proposed (SDSMC). The experimental results indicate that all processing steps of our proposed architecture can be accomplished with good performance. From the table one can understood that the proposed approach is doing well in terms of time.

In general when the size of file increases time also gets increased but the other security limitations such as privacy, data confidentiality, data integrity and availability of data are far better than single cloud. Similarly when the size of the file, parts of the file and the number of providers increases then the overall performance time decreases because of the parallel execution of all the task at the same time in the proposed SDSMC Multi-Cloud Storage. In the proposed work threshold size of the file is 552 Mb and the minimum threshold number of the storage providers is five. Since the Multi-Cloud Storage is a subscription service the higher the size of the file the higher will be the cost to be paid by the user.

VI. FUTURE WORK

Although the proposed model ensures the protection of data sharing from malicious insiders and files there is a possibility of leakage of key without the owner's knowledge when the framework interface gets accessed from the public networks. When the data owner tries to upload the more files key management becomes cumbersome. To rectify above problems system a public key hybrid crypto system is needed. To enhance the trust of the customers file slicing parts can be defined by the owner itself is the other future directions of our proposed model.

VII. CONCLUSION

The proposed methodology is a Multi Cloud Storage security scheme for organizational as well as non-organizational aspects. Since the various data sets have been used to operate on the SDSMC model and reaches the higher security when compared with other models. The proposed architecture reduces the malicious insider threats and the proposed procedure ensures the providers resource protection from the malicious files. The SDSMC supports all type of files including video files can be encrypted based on the index based cryptographic technique. In the retrieval of the files a standard procedure is used which reduces on demand cost and the conflicts in the merging process. The experimental results justifies the efficiency of the proposed algorithm. The numerical results justifies the data sharing security of the proposed model.

REFERENCES

[1] DananThilakanathan, ShipingChen,Surya Nepal and Rafael A.Calvo "Secure Data Sharing in the Cloud". In Security, Privacy and Trust in Cloud Systems,Springer Berlin Heidelberg,2015,(pp. 45-72).

[2] Benjamin Fabian, Tatiana Ermakova,PhilippJunghanns "Collaborative and secure sharing of healthcare data in multi-clouds". Information Systems, Volume 48 Issue C, 2015,pp 132-150

[3] Balasaraswathi, V. R., &Manikandan, S. (2014)." Enhanced security for multi-cloud storage using cryptographic data splitting with dynamic approach". In Advanced Communication, International Conference onControl and Computing Technologies (ICACCCT), 2014 on (pp. 1190-1194). IEEE.

[4] Mazhar Ali, RevathiDhamotharan, ErajKhan,SameeU.Khan,AthanasiosV.Vasilakos,KeqinLi,Albert.Y.Zomaya "SeDaSC: Secure Data Sharing in Clouds", Systems Journal, IEEE, volume :PP, Issue:99,2015,pp 1-10.

[5] Wang Liang-liang,ChenKe-fei,Mao Xian-ping,Wang Yong-tao "Efficient and Provably-Secure Certificateless Proxy Re-encryption Scheme for Secure Cloud Data Sharing" Journal of Shanghai Jiaotong University Volume 19, issue 4,2014 pp 398-405.

[6] PengXu, XiaqiLiu,ZhenguoSheng,XuanShan,KaiShuang "SSDS-MC: Slice-based Secure Data Storage in MultiCloud Environment" 11th EAI International Conference on Heterogeneous Networking for Quality, Reliability, Security and Robustness (QSHINE) , 2015,pp 304-309.

[7] ShunganZhou.RuiyingDu,JingChen,HuaDeng,JianShen,Huanguo Zhang "SSEM: Secure, Scalable and Efficient multi-owner data sharing in clouds", China Communications IEEE ,Volume 13,issue 8, 2016,pp 231-243.

[8] Ibrahim Abdullah Althamary, TalalMousaAlkharobi "Secure File Sharing in Multi-Cloud using Shamir's Secret Sharing Scheme",Transactions on Network and communications Vol 4 issue 6, 2016,pp53-67.

[9] Safaa Salam Hatem, Maged H.Wafy,Mahmoud M.EI-Khouly "Malware Detection in cloud Computing",International Journal of Advanced Science and Computer Science Applications,Vol 5 No 2014.

[10] MahaTebaa, Said El Hajji "From Single to Multi-Clouds Computing Privacy and Fault Tolerance", Science Direct (ELSEVIER), InternationalConference on Future Information Engineering,(2014),pp112-118.

[11] YuukiKajiura,Shohei Ueno,Atsushi Kanai, ShigeakiTanimoto, Hiroyuki Sato "An Approach to Selecting Cloud Services for Data Storage in Heterogeneous-Multicloud Environment with High Availability and Confidentiality Autonomous Decentralized Systems" (ISADS) IEEE Twelfth International Symposium,2015,(pp 205 – 210).

[12] Tatiana Ermakova, Benjamin Fabian "Secret Sharing for Health Data in Multi-provider Clouds Business Informatics" (CBI), 2013 IEEE 15th Conference,2013,pp 93-100.

[13] Xu, L., Wu, X., & Zhang, X. "CL-PRE: A certificate less proxy re-encryption scheme for secure data sharing with public cloud". In Proceedings of the 7th ACM Symposium on Information, Computer and Communications Security,2012 pp. 87-88 .

[14] Seo, S. H., Nabeel, M., Ding, X., &Bertino, E." An Efficient Certificate less Encryption for Secure Data Sharing in Public Clouds. Knowledge and Data Engineering", IEEE Transactions on, 26 (9), 2013,pp2107-2119.

[15] Abdul Nasir Khan , M. L. Mat Kiaha, Sajjad A. Madanib, MazharAlic, Atta urRehmanKhana , ShahaboddinShamshirbanda "Incremental proxy re-encryption scheme for mobile cloud computing environment". The Journal of Supercomputing, 68 (2), 2014 Pp624-651.

[16] Yashaswisingh, Farah Kandah, WeiyiZhang "A Secured Cost-effective Multi-Cloud Storage in Cloud Computing" IEEE INFOCOM Workshop on Cloud Computing 2011,pp 619-624.

[17] Dr. K.Subramanian, F.Leo john "Data Security in Single and Multi-Cloud Storage- an Overview" *International Journal of innovative Research in Communication Engineering* 2016 pp 19046-19052.

[18] BorkoFurht, Armando Escalante "The Handbook of Cloud Computing",Springer Publications 2012.

[19] Jens-Matthias Bohli,NilsGruschka,MeikoJensen,Luigi Lo Lacono andNinja Marnau,"Security and Privacy-Enhancing Multicloud Architectures" IEEE Transactions On Dependable and Secure Computing 2013 pp-212-224.

[20] VenkataJosyula, Malcom Orr, Greg Page,"CloudComputing:Automating the Virtualized Data Center" Cisco Press 2012.

[21] Alycia Sebastin,Dr.L.Arockiam "A Study on Data Security Issues in Public Cloud", International Journal of Scientific and TechnologyResearchVolume 3 Issue 5 May *2014 pp 144-146.*

[22] B.Rex Cyril, Dr.S.Britto Ramesh Kumar "Cloud Computing Data Security Issues, Challenges, Architectures and Methods-A Survey" InternationalJournal of Engineering and Technology(2015).

AUTHORS PROFILE

Dr. Subramanian Krishnasamy is currently working as an Assistant Professor in H.H The Rajah's College. His area of interest includes Data Mining, Networking, Cloud Computing, Network Security, Big Data, Multi-Cloud and so on.

Mr. Leo John is a part-time research scholar in Computer Science Department, JJ.College of Arts and Science pudukkottai. His area of interest includes Cloud Computing, Unstructured Data Security in multi-cloud, cryptography and so on.

An Empirical Investigation of the Correlation between Package-Level Cohesion and Maintenance Effort

Waleed Albattah
Department of Information Technology
Qassim University
Qassim, Saudi Arabia

***Abstract*—The quality of the software design has a considerable impact on software maintainability. Improving software quality can reduce costs and efforts of software maintenance. Cohesion, as one of software quality characteristics, can be used as an early indicator for predicting software maintenance efforts. This paper improves Martin's cohesion metric, which is one of the well-known and well-accepted cohesion metrics. The strong correlation found between package cohesion, using our proposed metric, and maintenance efforts shows the improvement made on measuring cohesion, and how it would be for predicting maintenance efforts. The experimental study included data from four open source Java software systems. The results show that the package cohesion is good and low maintenance is required.**

Keywords—package; cohesion; metric; maintenance effort; maintainability; software; measurements

I. INTRODUCTION

Software maintainability refers to the ease of maintaining software products in order to prevent or correct defects and their causes, and to respond to new requirements and environmental changes [1]. The quality of the software design has a considerable impact on software maintainability [2]. Predicting software maintainability during the software design phase can reduce much of the maintenance costs and efforts, and improve software maintenance. While a number of research studies performed were based on measures taken after the coding phase, the cohesion metric we developed has an advantage of measuring cohesion in an earlier phase, the design phase. Another advantage of this metric is that it has been developed based on well-known and well-accepted package cohesion principles [3]. Further, if there is a relationship between our metric and software maintainability, then we will potentially establish a relationship between these principles and software maintainability.

This paper investigates the relationship between package cohesion, using the proposed metric CH, and software maintenance efforts. For this purpose, the package cohesion metric has been developed, based on a solid theory of the package design principles [3]. A number of experiments and statistical analyses have been designed and performed to investigate this relationship.

Looking carefully to the existing studies, some studies were conducted using a cohesion metric on the class level. Others were not validated or only validated theoretically without any empirical validation of the relationship with software maintenance. Some studies [4] used a subjective expert's surveys. Some related experimental studies [5-10] were performed to investigate some aspects of software maintenance, such as defect density or fault-proneness, but they don't consider other types of maintenance, such as adaptive maintenance. Some studies [11][12] did not rely on the reported maintenance history of the studied software systems. The drawback in such studies is that the maintenance data collected for the experimental studies does not represent the actual maintenance data. Some studies, such as [13][14], have relatively a small sample size of the experimental study, which makes the results hard to be generalized.

In contrast, we found that our study is unique in several different ways. It proposes a cohesion metric on a package level based on the well-known package cohesion principles, both theoretically and experimentally validated, uses actual maintenance data history of software, uses objective data instead of subjective ones, and considers all types of maintenance activities. To the best of our knowledge, there is no study that has investigated the relationship between package level cohesion and software maintainability, which makes this research original and vital in this matter.

The rest of this paper is organized as follows: The related studies are briefly introduced in Section II. Section III presents an overview of the studied package cohesion metrics. Section IV details the empirical study. Section V investigates and discusses the correlation between package cohesion and maintenance effort. Finally, Section VI concludes the paper with future works.

II. RELATED WORK

Many researchers and practitioners proposed software metrics in relation to software maintainability and its characteristics. While some of them were theoretically validated, only a few were empirically validated. Several research studies were conducted to investigate the relationship between class-level cohesion and software maintainability. One of the early investigation studies was by Li and Henry [13] to investigate the validity of object-oriented metrics in predicting

software maintenance efforts. The study tested if there is a strong relationship between object-oriented software metrics and maintenance efforts. LCOM, a cohesion metric developed by Chidamber and Kemerer [15], was among ten software metrics that were investigated. The results of the statistical analysis performed on two software systems showed that there is a strong relationship between the studied software metrics and maintenance efforts. Briand et al. [16] proposed cohesion and coupling measures based on object-oriented design principles to evaluate software maintainability. However, this approach was not validated. Briand et al. [17] defined a ratio-scale metric for cohesion to predict the error-proneness in the software design. The results of the experiments proved that software metrics can predict software error-proneness. Dagpinar and Jahnke [14] provided empirical evidence that software metrics can effectively be used to predict software maintainability. However, they found that Bieman and Kang's Loose Class Cohesion (LCC) [18], metric was not a significant predictor for class maintainability. Basili et al. [19] were concerned about fault detection and the fault prone-ness part of maintenance. They showed by their experiments' results that the Chidamber and Kemerer's metrics [15] are, individually, good indicators for faulty modules. This was supported by Gyimothy et al. [122] where a validation of the ability of the LCOM metric as a good indicator of software fault-proneness was indicated. The study was conducted on open source software, Mozilla. Koru et al. [20] showed that there is a correlation between the number of bugs and size. Al Dallal [119] empirically investigated the relationship between a number of internal class quality attributes (size, cohesion, and coupling) and class maintainability. Prediction models, based on statistical techniques, were constructed and validated to estimate the class maintainability. The results showed that internal attributes (size, cohesion, and coupling) have an impact on class maintainability. The higher the cohesion is, the higher the class maintainability is.

III. Package Cohesion Metrics

A. The proposed metric (CH)

In our previous work [22], which is motivated by Martin's package cohesion principles [3], we proposed two different cohesion metrics to measure two different cohesion concepts or types based on Martin's package cohesion principles in [3]. The first cohesion type, Common Reuse (CR), includes the factors that help in assessing CR cohesion. Similarly, the second cohesion type, Common Closure (CC), includes the factors that help in assessing CC cohesion. After each type of cohesion is measured by itself, the two values of CR and CC may be combined to one unified value of package cohesion, while still recognizing the two types.

The CR metric measures cohesion based only on the common reuse factors of the package. The elements of a package have different degrees of reachability. Reachability of a class in a package is the number of classes in the same package that can be reached directly or indirectly. The CR metric is defined as follows:

"Let c ∈ C, and suppose there is an incoming relation to c from a class in a different package. Then c is called an in-interface class. The cardinality of the intersection of the hub sets of all the in-interface classes in C divided by the number of classes in C is the CR of P ".

$$CR = |\cap_{\text{In-interface class hub sets}}| / |C| \quad (1)$$

where

Hubness(c) = {d ∈ C: if there is a path c → d}

C: set of classes in package *P*

c and d: classes in C

The CC metric considers the package dependencies on other packages as well as the internal dependencies between classes of the package. The classes of the package should depend on the same set of packages and, thus, they will have the same reasons for a change. The CC metric is defined as follows:

"The cardinality of the intersection of the reachable sets divided by the cardinality of the union of the sets represents the CC of P ".

$$CC = (|\cap_{\text{Reachable Package sets}}| / |\cup_{\text{Reachable Package sets}}|) \quad (2)$$

The combined cohesion *CH* is defined as follows:

$$CH = \frac{\sqrt{2} - D}{\sqrt{2}} \quad (3)$$

$$D = \sqrt{(1 - CR)^2 + (1 - CC)^2} \quad (4)$$

B. Martin's metric (H)

Martin proposed a rational cohesion metric for the package,

$$H = (R+1)/N \quad (5)$$

Where R: number of relationships between classes in the package

N: number of classes in the package

Although Martin's cohesion principles [3] are well known and well accepted, H metric doesn't conform to them. H measures the ratio of the relationships between classes of the package. This simple concept doesn't measure the common reuse or the common closure of the package, but rather, in its best situation, it may measure the classes' extent of being connected. The H metric depends on the number of relations rather than how these relations are designed. In this case, a well-designed package and a badly designed package could have the same cohesion value. In our previous work [22], further discussions are presented.

IV. Descriptive Statistics

This empirical study is based on four open-source Java software systems used to investigate the relation of package cohesion measure to software maintainability. This section provides descriptions about the studied software systems and the maintenance data collection. Two package cohesion metrics are included in this study: Martin's cohesion metric (H) and the proposed package cohesion metric (CH), which is developed based on Martin's package cohesion principles [3].

A. The software systems

Four open-source Java software systems were involved in the empirical study. All the four systems were selected based on the following criteria to allow results' generality; they had: (1) to be implemented using the Java programming language, (2) to have maintenance repositories available, namely Apache Subversion (SVN), (3) to have sufficient number of versions for each system that have been maintained, (4) to be organized using packages, (5) to have different sizes ranging from very large to small systems in terms of number of packages and number of classes, (6) to be from different domains, and (7) to have positive reviews and to be mature. We expect these criteria will allow the generalization of the results obtained from the study. The first system, Camel [23], is a rule-based and mediation engine to configure routing and mediation rules. The second system, Tomcat [24], is an open source webserver developed to implement Javaservlet and Java Server pages (JSP). Apache Tomcat is developed by the Apache Software Foundation. It has been developed and released under Apache License version 2. The third system, JHotDraw [25], is a Java GUI framework for technical and structured graphics. The fourth system, JEdit [26], is an open source Java text editor for programmers. It is licensed by GPL General Public License version 2.0. Table I provides details of the maintenance history; and Table II provides details about the studied systems.

TABLE I. MAINTENANCE HISTORY

	Base Release	End Release	History Studied
Camel	2.0.0	2.2.0	Aug/24/09 – Feb/6/10
Tomcat	7.0.6	7.0.22	Jan/14/11 – Oct/1/11
JHotDraw	7.5	7.6	July/29/10 – Jan/9/11
JEdit	4.5.0	5.1.0	Jan/31/12 – July/28/13

TABLE II. THE STUDIED SYSTEMS

	#LOC	#Methods	#Classes	#Packages	#Revised-Packages
Camel	143732	17369	5111	264	179
Tomcat	170461	15372	1725	113	62
JHotDraw	77194	7122	1026	65	65
JEdit	111861	7386	1238	35	23

B. Maintenance data

The source of the maintenance data for this study is the Version Control System (VCS), subversion (SVN), which is publicly available. The public can view the history of maintenance activities that have been made on the software system using SVN client. Each log entry in the repository log has a revision number, date and time, and short message that explains the maintenance activity. We considered all types of maintenance activities: perfective, adaptive, corrective, and preventive. We don't differentiate between different maintenance activities.

For this empirical study, as suggested by Al Dallal [21], we considered two package maintenance measures: the number of revisions (#Revisions) in which the package has been involved, and the number of revised lines of code (RLOC) during the studied maintenance history. The number of revised lines of code RLOC is calculated as suggested by Li and Henry [13], where a line added or deleted is considered one revised line, and a line modified is considered two revised lines, one deletion and one addition. We consider these two measures for two reasons. First, the number of revisions refers to the maintenance rate, while the number of RLOC is found to be correlated with maintenance cost [27][21] and maintenance effort measured in unit of time [28][21]. Packages with lower maintenance rates are better than those with higher rates because the code with more revisions becomes less organized, less understandable, and more fault-prone [29][21]. Second, these two measures are measurable using the freely available software maintenance history [21].

To collect maintenance data, we used the free software tool, TortoiseSVN [30], which is a subversion client developed to access the subversion (SVN) repositories. For each software system, the log of the SVN repository includes the following revision information: revision number, revision description, all the packages and classes affected by the revision, the previous and the current class versions, and the number of lines added, deleted, or modified. We had to create a list of all the packages and the classes within the package to relate each revision's information to the appropriate package. Then, revisions and revised lines of code were collected on package level. We considered different versions for each system, and collected the maintenance data reported during the entire maintenance period. Table III summarizes maintenance data for each system.

TABLE III. MAINTENANCE DATA

	#Revisions	Mean #Revisions	#RLOC	Mean #RLOC
Camel	1614	6.11	60688	229.87
Tomcat	636	5.63	22027	194.93
JHotDraw	354	5.45	21857	336.26
JEdit	323	9.23	9981	285.17

Two computer science PhD students were dedicated to collecting the maintenance data. The data was collected manually from the maintenance repositories. We have randomly checked the validity of the data collected. This process increased our confidence about the validity of the data collected.

For the purpose of a system's list of classes and list of packages, we have used the JHawk tool [31]. Then, each revision reported in the maintenance history was specified to the associated class along with the number of revised lines of code RLOC. Finally, maintenance data was collected on the package level.

C. Package cohesion data

Package cohesion data was gathered from two package cohesion metrics. The first metric is our proposed package cohesion metric, CH. The second metric is Martin's cohesion metric, H. These two metrics have been used to investigate the correlation between package cohesion and maintainability. For the purpose of data gathering, we have developed our Java tool to measure the CH package cohesion metric. The tool has been extended to calculate Martin's package cohesion metric, H. For each system, a list of all the packages, the number of classes in

each package, and the associated cohesion values were generated.

V. Exploring the Correlation between Cohesion and Maintenance Effort

The correlation analysis aims to determine whether each individual package cohesion metric (CH and H) is significantly related to the maintenance measures (#Revisions and RLOC) of the package. For this purpose, we have performed Spearman's rank correlation due to the non-parametric nature of the metrics' data. We have used the well-known SPSS software for the correlation analysis of the empirical study. We have created and analyzed a correlation matrix for each software system in the study. Each correlation matrix has all the studied variables (cohesion and maintenance), a correlation coefficient (r), and significance level. For each pair of variables, r value can range between -1 and +1, where 1 represents a perfect positive correlation between the pair variables; -1 denotes a perfect negative correlation; and 0 indicates that there is no relationship between the variables. The magnitude of the coefficient determines the degree of the correlation.

Besides the strength of the correlation, the relationship between any pair of variables should be assessed for its significance as well. The significance is assessed by the p-value, which corresponds to the probability that the found correlation might be due to purely random effects. The smaller the p-level, the more significant is the relationship between variables [32]. The significance of the correlation in this empirical study was tested at a 95% confidence level (i.e., p-level ≤ 0.05). While the correlation can establish the relationship, it cannot establish a cause-effect relationship between the pair variables [32].

A. Hypotheses

Our objective is to assess to what extent is the package cohesion metric related to the maintenance effort of the software packages. The hypotheses of the empirical study are:

H_{01}: There is no significant correlation between package cohesion, CH, and the number of Revisions, #Revisions.

H_{02}: There is no significant correlation between package cohesion, CH, and the number of revised lines of code, RLOC.

H_{03}: There is no significant correlation between Martin's package cohesion, H, and the number of Revisions, #Revisions.

H_{04}: There is no significant correlation between Martin's package cohesion, H, and the number of revised lines of code, RLOC.

In this experiment, rejecting the null hypothesis indicates that there is a statistically significant relationship between the pair of variables (significance level $\alpha = 0.05$).

B. Statistical Analysis

The number of software revisions (#Revisions) and the number of revised lines of code (RLOC) on the package during the maintenance history assess software package maintainability. A lower number of package revisions and a smaller number of revised lines of code during the package maintenance history indicates less effort needed to maintain the software and thus, indicate high maintainability.

Table IV provides descriptive statistics (mean and standard deviation) for the variables used in analyzing software maintainability across the four systems, Camel, Tomcat, JHotDraw, and JEdit. We included Martin's package cohesion metric (H) in the list of variables for the purpose of comparison.

TABLE IV. Means and Standard Deviations of the Variables Used in the Maintainability Analysis

Variable	Camel N=264		Tomcat N=113		JHotDraw N=65		JEdit N=35	
	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
H	.636	.361	.817	.524	.705	.502	1.059	1.075
CH	.530	.388	.358	.374	.288	.317	.374	.417
#Classes	13.700	29.637	16.17	23.063	16.31	18.332	35.37	47.888
#Revisions	6.114	14.91	5.575	12.238	5.45	3.192	9.23	17.066
RLOC	229.879	732.318	194.69	511.186	336.26	401.748	285.17	573.679

C. Results and Discussion

A Spearman Rho correlation is the appropriate measure of a bivariate relationship when normality and linearity conditions for the Pearson's product moment correlation do not hold. For this study, the Spearman Rho correlation provides a measure of association between the proposed measure of package cohesion CH, the Martin's package cohesion metric H, package size (#Classes), and the two measures of package maintainability, the number of package revisions (#Revisions) and the number of revised lines of code (RLOC), within each of the four data sets. Table V provides the list of these correlations for the four sets of data.

TABLE V. SPEARMAN'S RHO CORRELATIONS FOR MAINTAINABILITY ANALYSIS

Data Set		H	CH	#Classes	#Revisions
	CH	.281**			
Camel	#Classes	-.350**	-.655**		
N=264	#Revisions	-.101	-.562**	.720**	
	RLOC	-.129*	-.533**	.702**	.962**
	CH	.169			
Tomcat	#Classes	-.069	-.736**		
N=113	#Revisions	-.010	-.545**	.686**	
	RLOC	-.007	-,521**	.663**	.792**
	CH	.157			
JHotDraw	#Classes	-.041	-.706**		
N=65	#Revisions	-.07	-.594**	.674**	
	RLOC	.098	-.631**	.769**	.792**
	CH	.468**			
JEdit	#Classes	-.205	-.709**		
N=35	#Revisions	-.024	-.650**	.754**	
	RLOC	-.008	-.623**	.711**	.983**

** Correlation is significant at the .001 level

* Correlation is significant at the .05 level

Table V reveals that the new proposed measure of package cohesion, CH, consistently has a negative large correlation with the two measures of package maintainability, number of package revisions (#Revisions) and the number of revised lines of code (RLOC), across all the four data sets. The correlation values between package cohesion CH and number of revisions (#Revisions) across the four data sets range from -0.545 (for the Tomcat system data set) to -0.650 (for the JEdit system data set). Similarly, the correlation values between package cohesion CH and the number of revised lines of code (RLOC) across the four data sets ranges from -0.521 (For the Tomcat system data set) to -0.631 (for the JHotDraw system data set). The statistically significant correlations confirm that the expectation of a highly cohesive software package requires less effort to maintain. That is high values of the proposed measure of package cohesion are associated with a lower number of its revisions and a lower number of revised lines of code.

In this study, the correlations between Martin's package cohesion metric H and the two package maintainability measures, number of package revisions (#Revisions), and the number of revised lines of code (RLOC) are not as strong as the ones with the newly proposed measure of package cohesion CH. These correlations are consistently weak and statistically insignificant across all the four data sets, except for the correlation with the revised lines of code (RLOC) for the Camel system's data. The value of the correlation is -.129, which relatively small yet statistically significant at an .05 level. The significance of the weak correlation might be justified by the large sample size of the Camel system data set. The correlation values between Martin's package cohesion H and number of revisions (#Revisions) across the four data sets range from -0.010 (for the Tomcat system data set) to -0.101 (for the Camel system data set). Similarly, correlation values between Martin's package cohesion H and the number of revised lines of code (RLOC) across the four data sets ranges from -0.007 (for the Tomcat system data set) to -0.129 (for the Camel system data set).

Table VI summarizes the results of the examined null hypotheses. In this experiment, rejecting the null hypothesis indicates that there is a statistically significant relationship between the pair of variables (significance level $\alpha = 0.05$).

TABLE VI. THE RESULTS OF THE NULL HYPOTHESES

	Camel	Tomcat	JHotDraw	JEdit
H_{01}	Rejected	Rejected	Rejected	Rejected
H_{02}	Rejected	Rejected	Rejected	Rejected
H_{03}	Accepted	Accepted	Accepted	Accepted
H_{04}	Rejected	Accepted	Accepted	Accepted

VI. CONCLUSION

This study investigated the relationship between the software internal attribute, package cohesion, and the software external attribute, package maintainability. We found that package cohesion, using our proposed metric (CH), is highly correlated with package maintainability, measured by number of revisions (#Revisions) and number of revised lines of code (RLOC). As high cohesion, the package is the easiest to be maintained. Such relationship is explained by the Spearman's ranking correlations involving data sets of four Java open-source software systems. This high correlation will lead us in future to perform regression analyses to predict package maintainability using package cohesion. Predicting software maintainability during the software design phase can reduce much of maintenance costs and efforts.

One strength of this study is the number of the studied systems and the stability of the correlation of CH across all experiments performed that allows us to draw optimistic conclusions about the possibility of using it as an indicator.

The experiments support the relationship between package cohesion and software maintainability, although it may behave differently based on a system's domain. So the results in this study should be viewed as indicative rather than conclusive.

The study only involved systems developed in Java, and the results could be different with systems developed in other object-oriented languages (such as C++).

REFERENCES

[1] IEEE, IEEE standard glossary of software engineering terminology, IEEE Std 610.12-1990, Institute of Electrical and Electronics Engineering, 1990.

[2] Madhwaraj, K. G., and Chitra Babu. "An Empirical Investigation of the Influence of Object Oriented Design Quality Metrics on the Package Maintainability of Open Source Software." (2011).

[3] Martin, Robert Cecil. Agile software development: principles, patterns, and practices. Prentice Hall PTR, 2003.

[4] Muthanna, S.; Kontogiannis, K.; Ponnambalam, K. and Stacey, B., "A maintainability model for industrial software systems using design level metrics," *In Proceedings of 7th Working Conference on Reverse Engineering*, pages 248-256, 2000.

[5] Marcus, Andrian, Denys Poshyvanyk, and Rudolf Ferenc. "Using the conceptual cohesion of classes for fault prediction in object-oriented systems." *IEEE Transactions on Software Engineering*, 34.2 (2008): 287-300.

[6] [Al Dallal, Jehad. "Fault prediction and the discriminative powers of connectivity-based object-oriented class cohesion metrics." *Information and Software Technology* 54.4 (2012): 396-416.

[7] Al Dallal, Jehad, and Lionel C. Briand. "An object-oriented high-level design-based class cohesion metric." *Information and software technology* 52.12 (2010): 1346-1361.

[8] Briand, Lionel C., et al. "Predicting fault-prone classes with design measures in object-oriented systems." *The Ninth International Symposium on Software Reliability Engineering Proceedings*, 1998. IEEE, 1998.

[9] Gyimothy, Tibor, Rudolf Ferenc, and Istvan Siket. "Empirical validation of object-oriented metrics on open source software for fault prediction." *IEEE Transactions on Software Engineering*, 31.10 (2005): 897-910.

[10] Olague, Hector M., et al. "Empirical validation of three software metrics suites to predict fault-proneness of object-oriented classes developed using highly iterative or agile software development processes." *IEEE Transactions on Software Engineering*, 33.6 (2007): 402-419.

[11] Kabaili, Hind, Rudolf K. Keller, and Francois Lustman. "Cohesion as changeability indicator in object-oriented systems." *Fifth European Conference on Software Maintenance and Reengineering*, 2001. IEEE, 2001.

[12] Ajrnal Chaumun, M., et al. "A change impact model for changeability assessment in object-oriented software systems." *Proceedings of the Third European Conference on Software Maintenance and Reengineering*, 1999. IEEE, 1999.

[13] Li, Wei, and Sallie Henry. "Object-oriented metrics that predict maintainability." *Journal of systems and software*, 23.2 (1993): 111-122.

[14] M. Dagpinar and J. Jahnke, "Predicting Maintainability with OO Metrics – An Empirical Comparison", *10th Working Conference on Reverse Engineering Proc* (WCRE'03), 13-17 Nov, 2003, pp 155-164, 2003.

[15] Chidamber, Shyam R., and Chris F. Kemerer. Towards a metrics suite for object oriented design. Vol. 26. No. 11. ACM, 1991.

[16] Briand, Lionel C., Sandro Morasca, and Victor R. Basili. "Measuring and assessing maintainability at the end of high level design." *Conference on Software Maintenance Proceedings, 1993*. CSM-93, IEEE, 1993.

[17] Briand, Lionel, Sandro Morasca, and Victor R. Basili. "Defining and validating high-level design metrics." (1994).

[18] Bieman, James M., and Byung-Kyoo Kang. "Cohesion and reuse in an object-oriented system." ACM SIGSOFT Software Engineering Notes. Vol. 20. No. SI. ACM, 1995.

[19] Basili, Victor R., Lionel C. Briand, and Walcelio L. Melo. "A validation of object-oriented design metrics as quality indicators." *IEEE Transactions on Software Engineering*, 22.10 (1996): 751-761.

[20] Koru, A. Gunes, Dongsong Zhang, and Hongfang Liu. "Modeling the effect of size on defect proneness for open-source software." *Proceedings of the Third International Workshop on Predictor Models in Software Engineering*. IEEE Computer Society, 2007.

[21] Al Dallal, Jehad. "Object-oriented class maintainability prediction using internal quality attributes." *Information and Software Technology* 55.11 (2013): 2028-2048.

[22] W. Albattah and A. Melton, "Package cohesion classification", in: *5th IEEE International Conference on Software Engineering and Service Science* (ICSESS), 2014, IEEE, 2014, (pp. 1–8).

[23] http://camel.apache.org (accessed March 2014)

[24] http://tomcat.apache.org (accessed March 2014)

[25] http://www.jhotdraw.org (accessed March 2014)

[26] http://www.jedit.org (accessed March 2014)

[27] Granja-Alvarez, Juan Carlos, and Manuel José Barranco-García. "A method for estimating maintenance cost in a software project: a case study." *Journal of Software Maintenance* 9.3 (1997): 161-175.

[28] Hayes, Jane Huffman, Sandip C. Patel, and Liming Zhao. "A metrics-based software maintenance effort model." *15th European Conference on Software Maintenance and Reengineering*. IEEE Computer Society, 2004.

[29] K. Erdil, E. Finn, K. Keating, J. Meattle, S. Park, D. Yoon, Software maintenance as part of the software life cycle, Comp180: Software Engineering Project, Department of Computer Science, Tufits University, 2003.

[30] http://tortoisesvn.net (Accessed March 2014)

[31] http://www.virtualmachinery.com/jhawkprod.htm (accessed Dec 2013)

[32] Gupta, Varun, and Jitender Kumar Chhabra. "Package level cohesion measurement in object-oriented software." Journal of the Brazilian Computer Society 18.3 (2012): 251-266.

Congestion Control using Cross layer and Stochastic Approach in Distributed Networks

Selvarani R
Department of Computer science and Engineering
Alliance College of Engineering and Design
Bangalore, India

Vinodha K
Department of Information science and Engineering
The Oxford College of Engineering
Bangalore, India

***Abstract*—In recent past, the current Internet architecture has many challenges in supporting the magnificent network traffic. Among the various that affect the quality of communication in the massive architecture the challenge in maintaining congestion free flow of traffic is one of the major concerns. In this paper, we propose a novel technique to address this issue using cross layer paradigm based on stochastic approach with extended markovian model. The cross layer approach will bridge the physical layer, link layer, network layer and transport layer to control congestion. The resource provisioning operation will be carried out over link layer and the mechanism of exploring the congestion using stochastic approach will be implemented over the network layer. The Markov modeling is adopted to identify the best routes amidst of highly congested paths and it is carried out at the transport layer. An analytical research methodology will be adopted to prove that it is feasible to develop a technique that can identify the origination point of congestion and share the same with the entire network. It is found that this approach for congestion control is effective with respect to end to end delay, packet delivery ratio and processing time.**

Keywords—Distributed Network System; cross layer; congestion; Traffic Flow; Rate Control Metric

I. INTRODUCTION

Internet plays a vital role in dissemination of knowledge and servicing seamless and ubiquitous communication in the present era. With the advancements in technologies like cloud computing and optical network, offering high speed data delivery, data storage and retrieval is not an impossible task [1] yet, there is still a problem with the existing internet architecture. A closer look into the existing internet architecture revels that it is packed with various complexities Viz. incompatible in allowing connectivity with heterogeneous networks and its respective protocols, operating with various distributed networks [2][3]. The existing internet design principles can only permit networking with less complexity in its routing and communication process. These principles are not scalable for the requirement of the future internet architecture. The reason behind this is untrusted communication, more customer-oriented user environment, availability of many commercial network operators, data-centric utilities, and the worst part is intermittent connectivity [4]. Another challenging problem is its inclination towards Internet Protocol (IP) paradigm that makes it suitable for static internet users but not for mobile internet users. Therefore, whenever an application meets heterogeneity, it introduces a great deal of challenges for the network-based architecture and at the same time, it also leads to significant problems of resource allocation that can potentially affect the quality of performance. The connection technique of this architecture is characterized by one-to- many and many-to-many connections and also supports smart virtualization process. The significance of user-based participation is quite high with compatibility of multi-hop transmission scheme [5]. Unfortunately, none of the above mentioned schemes are present even to a lesser extent in the existing internet architecture.

The present paper deals with the problems related to congestion control in future internet architecture. Although understanding the user-behavior over traffic and predicting it is an NP(nondeterministic polynomial time) hard problem, there are studies existing in past that has already focused on congestion control mechanism but less evidence of studies have focused towards congestion control in future internet architecture. Essentially, this is built over three components viz. service, architecture and infrastructure. The next problem is interoperability. Given a scenario of multiple and heterogeneous network, it is a challenging task to process the control messages. This phenomenon is definitely a big impediment towards congestion. The next issue is for a given congestion over the dynamic network, it is quite challenging to maintain a balance between identifying the point of congestion and processing heterogeneous control messages. Hence, it can be said that it is quite a difficult task to identify and mitigate the level of congestion in this architecture.

This paper presents a joint algorithm that incorporates cross layered mechanism with stochastic approach and Markov modeling to mitigate the potential issues of congestion in massive distributed system (future internet architecture). Section II reviews the existing literature for congestion control. The motivation and problem identification is discussed in section III. Section IV deal with the proposed study and its significant contribution. The algorithms that are implemented to attain the goals are presented in section V. The results of the proposed study are analyzed in Section VI. The concluding remarks are discussed in section VII.

II. RELATED WORK

The existing research in this area is reveled here.

Gholipour et al. [6] have carried out an investigation on congestion problems in sensor network. The working principal of the sensor network operate with distributed algorithm. Here the authors discussed a technique based on cost metric. The

results were compared with respect to energy and packet. Efthymiopoulos et al. [7] have presented a study on congestion minimization pertaining to real-time streaming. The authors have introduced a technique that can provide traffic management in different domains of the network based on the bandwidth. The system is purely made for the internet-based peer-to-peer traffic. Jose et al. [8] have presented a congestion minimization technique that evaluates the rate of communicating signals in highly distributed manner. The outcome of the study was evaluated with respect to transmission rate and found that it offers better rate control mechanism for minimizing the congestion. Zaki et al. [9] have presented a solution towards mitigating congestion that is witnessed over highly unpredictable mobile networks. The authors tested their finding over the continuous date occurred on 3G network. The outcome of the study was evaluated with respect to throughput and delay to find that proposed system offers better resiliency for internet-based congestion. Ichrak et al. [10] have also investigated the problems of congestion in TCP-IP(Transmission Control Protocol/Internet Protocol)based connection. Sonmez et al. [11] have presented a technique that focuses congestion identification and reduction owing to multimedia transmission. The study focuses on the congestion control and its effect on the quality of the transmitted multimedia files using fuzzy logic mechanism. The outcome of the study was evaluated with respect to Peak Signal-to-Noise Ratio (PSNR).

Reddy and Krishna [12] has presented cross layer approach in order to mitigate the congestion issues in mesh network using TCP New Reno protocol. They focused on efficient channel capacity optimizing during the massive multimedia transmission and the results were assessed using packet transmission rate and delay. A scheme for controlling the congestion over TCP-IP based network was presented by the authors Carofiglio et al. [13]. They has used the principle of active queue management to control the congestion and found that the technique possessed an effective window size, round trip time, and queue size. Further studies towards distributed system were carried out by Antoniadis et al. [14]. Although they worked on a small network, the principle applied was considered as a guiding factor for large scale distributed network as it focuses on addressing an effective traffic management technique using game theory. Cai et al. [15] have presented a model for controlling congestion in TCP-based communication system. The proposed methodology controls congestion over the wireless network based on node-to-node interactions. Using the case study of adhoc-based network, they have proved that their method offered better congestion control. Under the constraint of the fading channel, Ye et al [16] used probability theory to show that the congestion control model for vehicular network offered improvement in the energy efficiency and data packet transmission over adhoc-based networks. Similar kind of work was carried out by Bouassida and Shawky [17] .They presented dynamic scheduling algorithm based on the priority of the messages. The focused on improving data reliability of real-time vehicular network.

Kas et al. [18] have presented a technique for performing scheduling over dynamic channels. The aim of this is to increase the throughput from application viewpoint. A specific level of weight is assigned to each node that is arbitrarily fine-tuned based on saturation level of the queue. The results were evaluated with respect to end-to-end delay and packet delivery ratio over Constant Bit Rate traffic. Li et al. [19] have investigated congestion control for delay-based network. The authors have compared their work with respect to voice and data traffic and showed that it can control congestion based on the available delay information. Misra et al. [20] has presented a unique technique based on automata theory for managing the congestion over wired network. The author have also applied stochastic-learning based mechanism and cellular automata for managing an effective queue size. The outcomes were assessed using sequence number, queue factor, etc. Uthra et al. [21] have proposed a rate control mechanism for governing the traffic so that efficient throughput can be managed to ensure transmission free from any sorts of collision. The outcome of the simulation-based study is recorded and compared with the existing predictive-based mechanism to control congestion and found that the presented system minimizes the traffic congestion and also enhances the traffic performance.

The following section presents the problem that is identified after reviewing the work that was carried out by researchers in the field of congestion control.

III. Motivation and Problem Identification

The following are the areas to be considered for efficient performance of future Internet. The prevailing research in this area fail to address the following:-

- The network quality parameters like delay, latency and channel capacity are not considered efficiently for congestion control mechanism.
- The current cross layer design allows manipulation of various layer parameters which leads to complication of congestion control and error management. In addition to that it is observed that the complexity of identifying the source of congestion is difficult because of the inefficient handling of randomness of traffic in heterogeneous network.

IV. Proposed System

The aim of the proposed system is to develop a novel algorithm that can identify the origin of congestion in distributed network system. Here, the emphasis on network resource allocation for dynamic data flow control is given. As future internet architecture will possess all the possible complexities of existing internet as well as other networking standards (owing to reconfigurable nature), it is essential to address the issues through empirical and analytical modeling. In addition to regular quality parameters our research addresses the issues related to air medium e.g interference and different levels of noise over wireless channel for modeling the traffic. This paper is a continuation of our work where we have offered a packet level congestion by introducing a parameter i.e., Rate Control Metric (RCM)[24] . This metric is designed to offer an efficient control over the highly distributed network. The system is designed with the principle of cross layer paradigm. The randomness in the heterogeneous network is studied

through stochastic based probability model. This system is viewed as a massive network through graph theory modeling for better analysis of traffic congestion. In order to study and mitigate the traffic congestion in heterogeneous network (Future internet architecture) the performance metric were analyzed through RCM.

In this research we have considered cross layer approach for effective communication between networks. The resource provisioning technique is enhanced through stochastic based approach where, the model based on markovian modeling provides an optimized search for favorable node for routing. This model supports in identifying the best possible node amidst of congested nodes for speedy transfer of data which enhances the throughput of the network.

A. Cross Layer network model

Identifying the origination point of congestion and determining the control messages for processing the routing process to mitigate congestion requires a robust mechanism. Cross-layered approach is used to overcome this problem by controlling physical layer, data link layer, network layer, and transport layer in the protocol stack of future internet architecture. The cross layer network model plays an important role in arbitrary provisioning of network resources for congestion control. The presented scheme uses a significant routing factor that supports dynamic communication through multiple hops. It also uses a scheme that controls and manages the rate of traffic flow to achieve the fairness in sharing of network resources. A cost efficient provisioning algorithm is designed that models a novel queuing technique for maintaining queue stability. One of the significant focuses of the presented technique is to include the scenarios of noise and interference. This approach helps in processing control messages of multiple layers to adjust the rate of traffic flow during peak hours and also select favorable nodes for communication between two different networks. Hence, the cross-layer scheme offers flexibility to process the control request with less delay and also ensures that it is applicable for distributed network with heterogeneity.

B. Stochastic Approach

The stochastic approach of the proposed system mainly involves an integrated implementation of resource provisioning, communication and controlling the traffic directions. This approach initializes the discrete networking states followed by selection of highly stabilized links, and apply provisioning. The system uses graph theory to design an algorithm that works over the distributed machines. The significant contribution of this approach is to develop a network model that uses noise and signal power to categorize the quality of the links. It also consider the constraints of first two layers (link, physical) where there is no assured link rate for assigned time instances in distributed networking system. There is also a possibility that the capacity of the route may vary over a period of time that will lead to a significant stochastic problem. In order to solve this issue, we have implemented Rate Control Metric (RCM) [24] that can extract the exact information about the traffic rate thereby giving more information about the capacity of the routes. In order to solve the problems related to computational complexity, we also initialize a hypothetical matrix that stores and extracts the best provisioned values, which acts as alternative for the congestion states of traffic. Hence, there is no significant control overhead due to this. Moreover, the provisional matrix is regularly updated which makes the proposed system independent from any degree of congestion found in a specific transmission area.

C. Markov Modelling

The main aim of Markov Modeling is to optimize the stochastic approach used for congestion control. The goal of this module will be to minimize the end-to-end delay in distributed networking system. We apply probability theory along with Markov model to find out alternate routes by exploring non-congested paths for routing. Markov chain is used for mapping the network model that uses queuing theory over the layered design. (The study doesn't emphasize much on queuing mechanism explicitly as there is already robust mechanism specific to routing protocols in distributed network). The system maintains two types of traffics in a matrix i.e. local traffic and global traffic. Local traffic can be accessed at any instance of time and global traffic information can be accessed only when the node has better residual energy. The energy model based on first order radio model or Radio Frequency (RF) circuitry principle [26] will be implemented in the proposed system. The next section discusses the implementation of cross layer algorithm, stochastic approach algorithm and markov modeling.

V. Algorithm Implementation

A. Algorithm for Cross-Layer Approach

In this work, we address the mitigation of congestion in the distributed network system through cross layer approach. The algorithm and implementation are as depicted below.

Algorithm:

Input: n (nodes), ρ (queue)

Output: updation of Link

Start

1. init n, ρ, // n is the number of nodes & ρ is the queue size

2. Estimate generated packets pkt on condition

$$pkt = \left[\frac{\rho}{J} \right]_0^I$$

3. Define link cost

$$C_{s,d}(t) = \arg_{\max}[\Delta\rho]$$

4. L[t]→{Ls,d[t]}

5. Define enhancement in cross-layer provisioning

$$L_c(t) = \int_0^L (\rho[x], (L_o - L_i))dx$$

6. Select→rand[t]∈L || rand[t]=Lc[t]

7. update L[t]

End

The algorithm is formalized by considering the number of nodes n and initial queue size ρ. Fig.1 illustrates the complete process flow of the proposed cross layer based provisioning.

The queue stability of the distributed networking system is defined by the following equation which is used to filter out all the links that have their queue size tending to infinity.

$$\lim_{t_{cum}\to\infty} |\rho(t)| < \infty$$

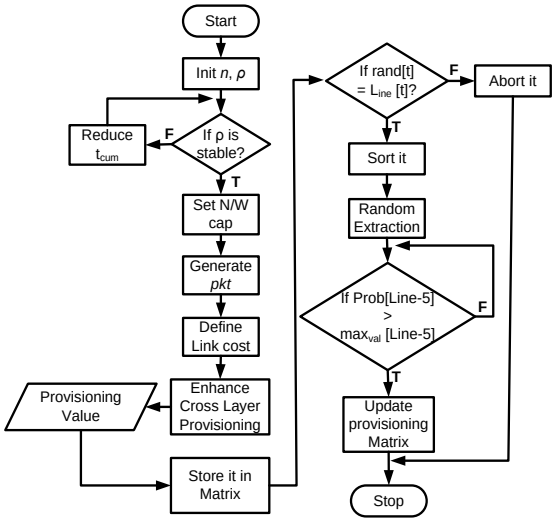


Fig. 1. Process flow for Cross Layer Provisioning

The link capacity of the network is expressed in terms of the *pkt*. Here, it is assumed that at time t the nodes in the network generate data packets equivalent to queue size ρ with controlled variables *I* and *J* (I and J are positive integers) are as shown in line2 of the algorithm. As the future internet architecture supports higher range of heterogeneity in device integration, there are possibilities of signal collision that leads to channel interference. In order to distinguish the quality of links (or routes) a new measurement link cost *C* is considered and is estimated based $\Delta\rho$, where the variable $\Delta\rho$ represents difference between the source queue ρ_s and destination queue ρ_d at the time *t.* The link provider metric $L[t]$ that is equivalent to $L_{s,d}[t]$, where *L* represents a *matrix* of provider that consist of non-colliding links between any source(s) and destination(d) is considered as the main parameter of the algorithm. It is assumed that initially the buffer is shared among each recipient node. The link provisioning matrix is updated considering the maximum value of the two arguments i.e. queue ρ[x] and difference between outgoing capacity of link Lo and incoming capacity of link Li.

The link provider will arbitrarily select an element from the matrix that satisfy the condition i.e. probability of selected element is equivalent to enhanced value (Line-6). It is updated as follows

$$L_c(t) = \int_0^L (\rho[x], (L_o - L_i))dx$$

Here, the link metric L[t] is estimated in terms of its queue size and the link capacity helps in the route selection process.

B. Algorithm for Stochastic Approach

In distributed networking system the traffic may undergo uncertainties like dynamic topology, random mobility etc. in high degree of randomness. The state of the network with uncertainty is analyzed through stochastic process in which the future node is identified with the theory of cross layer architecture. The nodes are initialized and their details are maintained on a data structure managed by graph theory. Owing to the distributed nature of the system, we assume that the control messages are free from errors or noise. After the implementation of cross-layer approach, we assume that there is no deviation or variance in the route capacity over the advancement of time.

Algorithm

Input: Es(energy for transmitting), δs,d (gain factor of the power), β (capacity of the channel), ψ (noise density)

Output: Provisioning state

Start

1. Evaluate SNR

$$SNR_{s,d} = \frac{E_s.\delta_{s,d}}{\beta.\psi}$$

2. Evaluate capacity of link

Lcap = β log2(1+SNRs,d)

3. Define duplicated groups

$$dp = \{dpS \mid s,d \in N, dpS \subseteq N_s \cap N_d\}$$

4. Function for duplicated groups

$$f(s,dp) = \{dpS \mid (s \in dpS)\text{^}(dpS \subseteq dp)\text{^}(|\, dpS \,| \geq 2)\}$$

5. If $s \subseteq S$ Than

6. for all $di \subseteq D$ do

7. rcm(t)←argmin(rcmmax, scaler_mult(t));

8. Apply Algorithm-1

9. Transmit data from s to d

10. Update scalar_multi(t)→state of provisioning

End

The algorithm is implemented by defining a network model, Signal-to-Noise Ratio (SNR) and Link Capacity (Line-1 and 2 of the algorithm). The duplicate control messages for analysis purpose are generated using Line-3 of the algorithm. In the above algorithm the source node *s* is identified as s_{id} and matrix of duplicate control messages containing information

about s as *s*.dp. The duplicated groups are formulated using the equation as shown in Line-4. The cross layer architecture of the future internet is designed in such a way that each source node s can access its routing table N_s. For reliable routing during peak traffic the algorithm allows node *s* to construct multiple hops with other nodes for providing alternate routes. The one dimensional matrix is generated by scalar multiplication of s and dp and the same is stored at every node. However, for all the duplicate control messages dpS, only the node that has highest value of id is chosen and is used in the computation process. The algorithm looks for all the source nodes s (S is total source nodes) and attempts to control the flow of packets. It then checks all the respective destination nodes and uses rate control metric (RCM) [24] to further enhance the provisioning for the data transmission. Finally, with the help of cross layer provisioning algorithm the date is transmitted towards the destination d.

The significance of the stochastic based approach algorithm is that it further enhances the resource provisioning offered by cross-layer based provisioning technique at the link layer and also supports better communication in the network layer by favoring multiple hops routing in distributed networking system. Finally, the data transmission is improved by applying rate control metric [24] which assigns an appropriate rate at the transport layer for effective end to end communication. Hence, the algorithm completely supports the cross-layer paradigm for future internet architecture to ensure interoperability among heterogeneous networks and achieve efficient data transmission.

C. Algorithm for Markov Modeling

The Markov modeling is used to further strengthen the algorithm discussed in the above sections and to apply stochastic modeling to further enhance the congestion control algorithm and offer a better solution to control traffic congestion. In Markov modeling each node is represented as Mc that is composed of the total number of layers corresponding to Lcap+1 (numerically). The amount of data packets pkt processed on each layer should be equivalents to Lcap such that 0<pkt<Lcap. We consider two different forms of layers Viz. passive layer PL and active layer AL. Passive layer represents the passive process when the nodes doesn't have any packet to forward (pkt=0) whereas in active layer, nodes always have packets for forwarding (pkt>0). As the future internet architecture possess different wireless nodes it is assumed that there are other feasible communication outages that will call for retransmission phenomenon. We denote φ as the amount of retransmission and Wn to be amount of unit trail of transmission. The algorithm for Markov modeling is given below.

Algorithm

Input: pkt (Packet), Lcap (Link capacity w.r.t queue), PL(Passive Layer), AL(Active Layer), φ (Maximum amount of retransmission)

Output: Identification of free/busy routes

Start

1. init pkt, Lcap, PL, AL, φ
2. Determine TP, PP, γP, IP.
3. Define area of collision A
4. $\{s,d\} \in A_{s,d}, \forall A_{s,d} = A_s \cap A_d$
5. Obtain |Fs,d|=As / As,d
6. Evaluate size matrix |As,d|, |Fs,d|, and |Fd,s|
7. Estimate number of Nodes

size(|As,d|, |Fs,d|, and |Fd,s|).network density

8. Estimate the probability of minimum transmission

$p_{A_s} = 1 - \vartheta$

9. Evaluate b1, b2, b3 & Mc=Algorithm-2{b1, b2, b3}
10. Find busy routes and free routes.

End

The problem of congestion in future internet architecture leads to network jamming that disrupt the process of identifying the best nodes for forwarding the data packets. This problem can be addressed by designing an algorithm that applies Markov modeling for evaluating the free and busy routes at the peak traffic situation. The algorithm takes the required inputs and computes maximum probability of passive state transition TP, preliminary state component PP, passive state probability component per states γP, and inter-arrival probability IP. Fig.2 shows the process flow for Markov modeling.

The Markov model is designed by considering three probability matrices viz. b1, b2 and b3. The matrix b1 and b2 represents the probability of a node identifying the busy channel in the first and second Markov process. The matrix b3 represents the feasibility that the packet forwarding process fails due to data packet collision or interference or noise. Line-3 shows the collision area A for the source node s. The area A is defined as a transmission zone where there is interference of the neighboring nodes resulting in traffic congestion in that particular transmission area. Hence, area A represents the possible congestion area. As shown in Line-4, it can be interpreted that both the sender node s and destination node d will lie within As,d. There can also be another possible transmission zone Fs,d as per Line-5 which may be undetected in the area As.d. It means that there may be an area e.g. F, which goes undetected and the state of congestion is not determined owing to dynamic mobility of nodes in mobile networks. In this case a source node or any intermediate node in area $F_{s,d}$ cannot forward the message to destination node d.

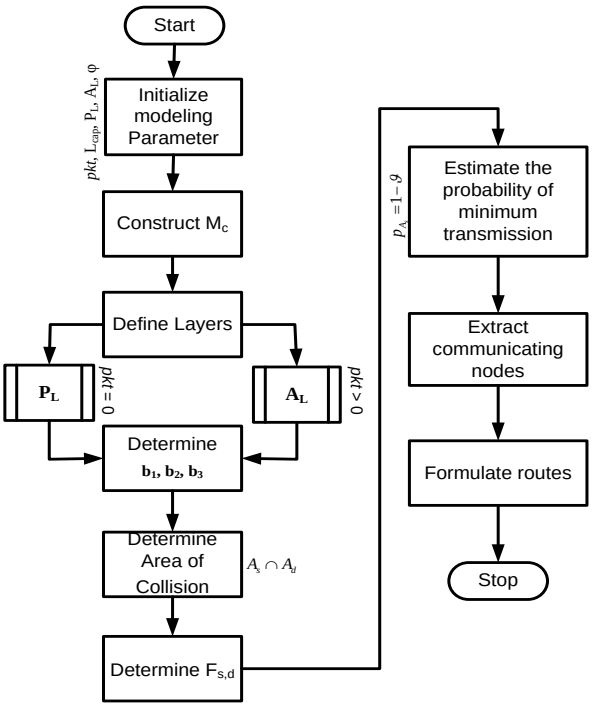


Fig. 2. Process Flow for Markov Modeling

The next phase of the algorithm is to compute the size of the transmission zones as per line-6. The algorithm computes the number of nodes in transmission zones by scalar multiplication of network size and network density as per line-7. The probability of minimum number of nodes required for forwarding data packets is computed as per Line-8. We use a simple variable ϑ that is equivalent to summation of the probability of all nodes carrying out data packet forwarding divided by total probability of the nodes forwarding data packets from the congested area. This phenomenon will mean that proposed Markov modeling attempts to find the existence of atleast one node which is in fair position to perform data transmission. The Markov modeling proceeds further to find similar kind of nodes and updates the matrix of data communication path that was previously managed by the algorithm of stochastic approach. The updated matrices helps to find the links between favorable nodes as the best possible alternate routes for packet forwarding during the peak traffic condition.

VI. Results and Discussion

This section discusses about the results generated from the network simulation through NS2 simulator. The simulation parameters are as shown in Table 1.

TABLE. I. Simulation Parameters

Parameter	Value	Parameter	Value
Network area(Simulation) area	1000 x 1200 m2	Control packet size	32 bits
Simulation Time	200 seconds	Data packet size	2000 bytes
Routing Protocol	NetFlow	Antenna Model	Omni-directional
Pathloss exponent	0.5	Maximum Speed of node	50 m/s
MAC Type	802.11	Minimum Speed node	1m/s
Traffic Model	CBR/VBR	Transmission range	10m
Mobility Model	Random	Transmission Energy consumption	0.5 J
Channel Model	Urban	Receiving Energy consumption	0.25J
Channel capacity	300 Mbps	Ideal mode Energy consumption	0.035 J
Channel sensing time	0.2 sec	Sleep mode Power consumption	0.02J
		Initial battery Energy of each node	10J

The proposed work focuses in finding an effective solution for congestion control in distributed networking system. The performance parameters like packet delivery ratio, end-to-end delay and processing time are considered to analyze the effectiveness of the proposed system. It is benchmarked with similar studies of Otoshi et al. [25] and Sahuquillo et al. [27]. Otoshi et al. [25] who have presented a stochastic modeling with predictive analysis for identifying discrete states of traffic in distributed networking system. This technique has used a predictive control scheme to minimize the possibilities of predictive error considering network constraints e.g number of hops, length of the hops etc. The mean length of the hops was considered as cost function, which was subjected to optimization using CPLEX solver. The outcome of the work was quite convincing as it has offered better scalability for future internet architecture. Similarly, we consider the work carried out by Sahuquillo et al. [27] as it offers solution to the congestion control for a practical case study of distributed networking system eg. High Performance Computing. The authors have used a mechanism that integrates injection throttle and segregation of congested traffic. We perform a minor modification to techniques introduced in [25] [27] in order to make a suitable testbed for carrying out the comparative analysis. The parameters considered for analysis are end-to-end delay, packet delivery ratio and processing time.

A. Comparative Analysis of End-to-End Delay

The end to end analysis is carried out by transmitting the test data of 2000 bytes. The result is as shown in Fig.3.

The graph shows that proposed system is able to minimize the end-to-end delay to a larger extent as compared to existing studies of Otoshi et al. [25] and Sahuquillo et al. [27]. The reason behind this is the technique that is adopted for processing search request and control messages by the proposed system.

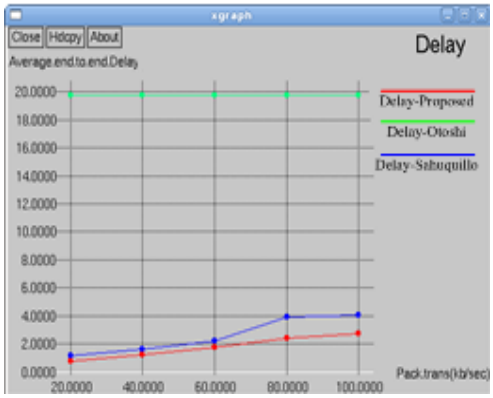


Fig. 3. Comparative Analysis of Delay (sec)

In the proposed system, owing to Markov modeling, it becomes essential for a node to obtain the significant address information of another communication node which could possible reside in transmission zone of $F_{s,d}$ or $A_{s,d}$. As both $F_{s,d}$ and $A_{s,d}$ are different transmission zones, extraction of the node address will be a quite difficult. We simplify this problem by developing a cross layer paradigm that can carry out the task of processing control messages in transport layer thereby minimizing the complexity.

Here, the task of one layer is to aggregate the respective addresses of the nodes and keep on exchanging it with other layers. This operation of interoperability is managed by the network layer. It is the responsibility of the network layer for carrying out the processing of control message as it maintains the communication standards of each transmission zone. This process helps in identifying the point of congestion and makes it aware to the entire network. This process has two advantages viz. i) all nodes can quickly decide about alternate routes and decrease the impact of congestion during peak traffic and ii) degree of congestion at the origination point is reduced by implementing active queue management that directs the packets from highly congested area to less congested point. Hence, end-to-end delay of the proposed work is reduced in the presence of mobility of the nodes which varies at every simulation track points. The problem explored in Otoshi et al. [25] is a predictive scheme. Here, the stochastic processing is adapted to predict and identify the possible prediction error. Hence, the delay factor using this technique cannot be implemented for distributed system of dynamic nature like that of future internet architecture. Similarly, the work done by Sahuquillo et al. [27] have focused on identifying congestion by using control messages which is quite time consuming in its nature. Using Markov modeling, proposed system offers optimized solution for identifying the point of congested and also offers best quality routes for packet forwarding thereby reducing the delay.

B. Comparative Analysis of Packet Delivery Ratio

Packet delivery ratio is computed by analyzing the amount of data packets received by the destination node to total amount of data transmitted by the source node. The result shown in Fig.4 exhibits that the proposed system offers better packet delivery ratio compared to Otoshi et al. [25] and Sahuquillo et al. [27]. This is because the proposed system provides a better processing of data generated by multiple networking domains in future internet architecture through cross layer paradigm. We start by analyzing the work done by Sahuquillo et al. [27]. The authors have implemented a technique where the incoming packets are organized at the input ports of the switches. The system emphasizes more on organization and less on queuing. This operation when implemented in our scenario reduces the packet delivery ratio. Moreover, the process of identification of the congestion and notify it to other nodes for updates are not discussed in that paper [27]. It is also not sure whether the updates were done over the highly congested area. This issue creates a negative impact on other neighboring nodes by consuming more time to take decision for routing. Hence, packet delivery ratio will be affected when this technique is used in future internet architecture.

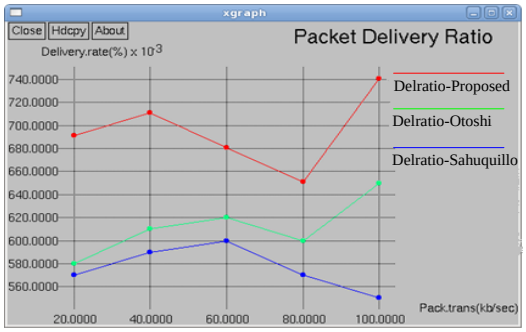


Fig. 4. Analysis of Packet Delivery Ratio

The technique proposed by Otoshi et al. [25] has used the concept of traffic engineering. This technique was implemented through stochastic modeling which is more predictive in nature. The predictive analytic model is assessed for its accuracy of traffic modeling using randomness by adopting traffic engineering with cost as a function on the stochastic model. This is much better than the technique discussed by Sahuquillo et al. [27] as it can accomplish better packet delivery ratio. The main drawback of this technique is that it uses control server to optimize the cost function which leads to less efficient distributed routing. Although, the authors have used relaxation mechanism to sort out this problem, but the probability factors assumed is less when compared to real-time traffic constraints. Hence, its packet delivery ratio is not better than the proposed system. The proposed system overcomes this problem by the algorithm-2 (stochastic) and algorithm-3 (Markov Modeling). These algorithms assist in identifying the best possible routes from non-congested area as well as congested area. The updating mechanism is quite instantaneous with a pause time of 0.0025 seconds in

simulation study that leads to better packet delivery ratio for a longer period of time.

C. Analysis of Processing Time

It is known that an effective congestion control mechanism must have a reduced processing time as far as possible. Lower the processing time means the network can ensure better instantaneous data delivery process. We analyze the processing time with increasing traffic load (packets per seconds). A closer look into the Fig 5 shows that processing time gets reduced linearly with increasing traffic load, which is one of the unique patterns of the proposed study. Usually with increased network traffic, the processing time should be increasing but due to cross layer approach the time complexity is reduced.

The cross layer approach bridges physical layer, link layer, network layer and transport layer. The provisioning operation is carried out over link layer, the mechanism of exploring the congestion using stochastic is implemented over network layer and Markov modeling for further optimizing the best routes (even from highly congested area) is carried out at the transport layer.

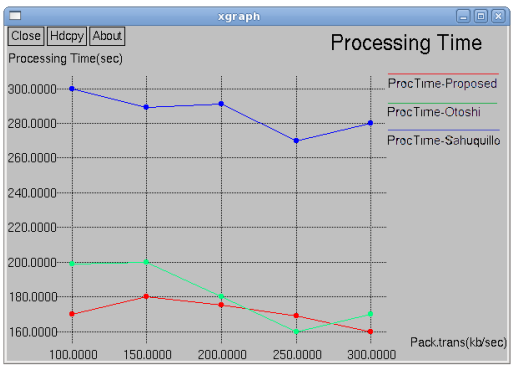


Fig. 5. Analysis of Processing Time

Hence, the system maintains different functionalities over different layers of protocol stack resulting in reduced processing time in the proposed system. For a given simulation environment, Otoshi et al. [25] and Sahuquillo et al. [27] work doesn't meets the demands of the distributed traffic scenario with dense congestion leading to higher processing time when compared to proposed system.

VII. Conclusion

Owing to the complexity in the design principle of distributed networking systems e.g. future internet architecture, the existing algorithms and techniques do not provide solution for mitigating congestion. The proposed system, therefore, presents a technique that uses conglomeration of cross layered approach, stochastic approach, and Markov modeling for addressing the problems of congestion in highly distributed networking system. We have adopted an analytical research methodology to prove that it is feasible to develop a technique that can identify the origination point of congestion and share the same with the entire network. The interesting point of implementation is that proposed technique attempts to use the existing network resources for harnessing the channel capacity in accordance with the state identified by the proposed system. The outcome of the study were compared with existing system respect to end-to-end delay, packet delivery ratio, and processing time and found that proposed system offers better solution for congestion control.

References

[1] C. White, Data Communications and Computer Networks: A Business User's Approach, Cengage Learning, Computers, 2015

[2] P. Verissimo, L. Rodrigues, Distributed Systems for System Architects, Springer Science & Business Media, Computers, 2012

[3] N. Rajan, The Digitized Imagination: Encounters with the Virtual World, Taylor & Francis, Social Science, 2012

[4] J. Holler, V. Tsiatsis, C. Mulligan, S. Avesand, S. Karnouskos, D. Boyle, "From Machine-to-Machine to the Internet of Things: Introduction to a New Age of Intelligence", Academic Press, Technology & Engineering, 2014

[5] H-Y Wei, J. Rykowski, S. Dixit, WiFi, WiMAX and LTE Multi-hop Mesh Networks: Basic Communication Protocols and Application Areas, John Wiley & Sons, Technology & Engineering, 2013

[6] M. Gholipour, A. T. Haghighat, M. R. Meybodi, "Hop-by-hop traffic-aware routing to congestion control in wireless sensor networks", *Springer- Eurasip Journal on Wireless Communications and Networking*, vol.15, 2015

[7] N. Efthymiopoulos, A. Christakidis, M. Efthymiopoulou, "Congestion Control For P2P Live Streaming", *International Journal of Peer to Peer Network*sm, Vol.6, No.2, August 2015

[8] L. Jose, L. Yan, M. Alizadehy, G. Varghese, "High Speed Networks Need Proactive Congestion Control", ACM-Proceedings of 14th ACM Workshop on Hot Topics in Network Article, No.4, 2015

[9] Y. Zaki, T. Potsch, J. Chen, "Adaptive Congestion Control for Unpredictable Cellular Networks", ACM-SIGCOMM, 2015

[10] T. Ichrak, S. Nawal and A. Mustapha, "Systematic Mapping Study on the Congestion Control Problem in TCP/IP", *Contemporary Engineering Sciences*, Vol. 7, no. 27, pp.1509-1515, 2014

[11] C. Sonmez, O. D. Incel, S. Isik, M. Y. Donmez, "Fuzzy-based congestion control for wireless multimedia sensor networks", *Springer-Eurasip Journal on Wireless Communications and Networkin*, vol. 63, 2014

[12] C. P. Reddy, P. V. Krishna, "Cross Layer Based Congestion Control in Wireless Mesh Networks", *Cybernetics And Information Technologies*, Vol.14, No 2, 2014

[13] G. Carofiglio, M. Gallox, L. Muscariellox and M. Papaliniz, "Multipath Congestion Control in Content-Centric Networks", *IEEE-Conference on Computer Communication Workshop*, pp.363-368, 2013

[14] P. Antoniadis, S. Fdida, C. Griffin, Y. Jin, "Distributed medium access control with conditionally altruistic users", *Springer- Eurasip Journal onWireless Communications and Networking*, vol.202, 2013

[15] Y. Cai, S. Jiang, Q. Guan, and F R. Yu, "Decoupling congestion control from TCP (semi-TCP) for multi-hop wireless networks", *Springer-Eurasip Journal onWireless Communications and Networking*, vol. 149, 2013

[16] F. Ye, R. Yim, J. Zhang, S. Roy, "Congestion Control to Achieve Optimal Broadcast Efficiency in VANETs", *IEEE-International Conference on Communication*, pp.1-5, 2010

[17] M. S. Bouassida and M. Shawky, "A Cooperative Congestion Control Approach within VANETs: Formal Verification and Performance Evaluation", *Hindawi Publishing Corporation, Eurasip Journal on Wireless Communications and Networking*, Article ID 712525, 2010

[18] M. Kas, I. Korpeoglu, and E. Karasan, "Utilization-Based Dynamic Scheduling Algorithm for WirelessMesh Networks", *Hindawi Publishing Corporation, Eurasip Journal on Wireless Communications and Networking*, Article ID 312828, 2010

[19] Y. Li, A. Papachristodoulou, M. Chiang, A. R. Calderbank, "Congestion control and its stability in networks with delay sensitive traffic", *Elsevier-Computer Networks*, vol.55, pp.20-32, 2011

[20] S. Misra, B. J. Oommen, S. Yanamandra, "Random Early Detection for Congestion Avoidance in Wired Networks: A Discretized Pursuit

Learning-Automata-Like Solution", *IEEE Transactions On Systems, Man, And Cybernetics*, vol. 40, no. 1, February 2010

[21] R. A. Uthra, S. V. K. Raja, A. Jeyasekar, A. J. Lattanze, "A probabilistic approach for predictive congestion control in wireless sensor networks", *Journal of Zhejiang University-Science*, vol.15, Iss.3, pp.:187-199, 2014

[22] K.Vinodha, R. Selvarani,"Congestion Control in Distributed Networks-a comparative study", *Springer –Advances in Computer Science and Information Technology*, pp.115-123, 2012

[23] K.Vinodha, R. Selvarani, "Congestion Control in Distributed Networking System-A Review", *International Journal of Computer Applications*, Vol.83, No 6, December 2013

[24] S. Rangaswamy and V. Krishnareddy, "An efficient traffic regulation mechanism for distributed networks", *Springer- Eurasip Journal on Wireless Communications and Networking*, vol.154, 2015

[25] T. Otoshiy, Y. Ohsitay, M. Muratay, "Traffic Engineering Based on Stochastic Model Predictive Control for Uncertain Traffic Change" *IEEE International Symposium on Integrated Network Management*, pp.1165-1170, 2015

[26] A. Grebennikov, N. Kumar, B. S. Yarman, *Broadband RF and Microwave Amplifiers*, Taylor & Francis, Computers, 2015

[27] J. E.Sahuquillo, E. G. Gran, P. J. Garcia, "Efficient and Cost-Effective Hybrid Congestion Control for HPC Interconnection Networks", *IEEE Transactions On Parallel And Distributed Systems*, vol. 26, no. 1, January 2015

Electronic Health as a Component of G2C Services

Rasim Alguliyev
Department of information society problems
Institute of Information Technology of ANAS
Baku, Azerbaijan

Farhad Yusifov
Department of information society problems
Institute of Information Technology of ANAS
Baku, Azerbaijan

***Abstract*—This paper explores electronic health as a segment of electronic government. International practice in electronic health field and electronic health strategies adopted in Europe are analysed. Current practices in delivery of electronic health services in G2C are investigated and perspectives are specified.**

***Keywords*—electronic health; electronic government; Government to Citizen (G2C); electronic services; medical information; Electronic Health Records (EHR)**

I. Introduction

At present, the active implementation of information-communication Technologies (ICT) is observed in various fields of human activity, including in public administration, economy, education and health.

The development in the sphere of ICT and web-technologies has substantially altered the organization, proposal and delivery of government services [1]. Currently, the establishment and use of new government services is mostly dependent on the development of electronic government (e-government) within the framework of national and international programs, as well as the demand of citizens for online use of electronic services.

More specifically, e-government forming in developed countries is based on electronic interactions of three models Government to Government (G2G), Government to Business (G2B) and Government to Citizen (G2C) [2, 3]. G2C is the expression of mutual relationships between citizens and government, and includes several electronic services (e-services) such as the forming of tax relations, issue of birth certificates, registration and voting by voters, conduct of referendums, provision of medical information. Electronic health (e-health) services delivered to citizens are among services that are approached with special sensitivity. Currently, citizen expectations are being expanded regarding e-services delivered in healthcare sphere such as accessibility of good practice and useful information, improvement of the quality of health services, new treatment methods, delivery of long-term medical support and medical insurance.

In 2000's e-health was used as a general term for explaining the use of electronic tools and electronic data related to information technologies in healthcare sector [4]. It has been reckoned, in a broad sense, that the use of e-health will facilitate the solution of several problems encountered by healthcare system, improvement of the efficiency of health services, and effective organization of management system.

Electronic health – is a broad term and can be defined as the use of electronic tools for the purpose of delivery of information, resources and services related to health protection. Several terms are being employed in e-health sector, among which electronic health record, mobile health, telehealth, telemedicine, electronic education, social networks where health issues are discussed, analysis of medical data and big data can be mentioned [4, 5].

The paper considers e-health as one of the segments of e-government. International practice in electronic health field and electronic health strategies adopted in Europe are analysed. Existing practice and solutions on the delivery of e-health services in G2C, advantages and problems of e-health are explored.

II. Electronic Health System

At present, there is no sole world practice for establishment of e-health system. Even in developed countries several models are proposed for establishment of government e-health system. The model selection depends on financing mechanism and the state of healthcare administration in the country in the first instance. The reforms conducted in healthcare system show that the role of commercially interested enterprises and insurance companies has recently been increasing in healthcare sector [6].

Socio-economic and financial impact and consequences of formation of e-health system have been explored in several thematic research works in Europe. The analysis of prospective spending and expected benefits in e-health system indicate that its socio-economic benefit for the society outweighs the spending in each separate case. The mutual link between e-health data and other clinical-medical systems is considered as the main advantage capable of bringing benefits.

In several research works, it is indicated that the delivery of health services with the implementation of ICT not only provides more advanced health services with lower costs, but also creates broad opportunities which stimulate economic growth. It is, specifically, noted that the application of ICT in healthcare sector will facilitate better quality, delivery of safer health services, the elimination of blunders related to medical drugs and satisfaction of patient needs, and the use of innovative models in delivery of medical aid [7].

It must also be mentioned that, despite several advantages, the application if ICT in medicine is observed to be relatively slower in comparison with the other sectors. Researchers usually explain this matter with the fact that an administration of health services is a complex process. As the main reasons, the requirements of participation of parties of interest, as well as central and local governance entities, doctors and other

experts in the field of medicine, the availability of medical information, information exchange between parties, the security maintenance and accurate processing of individual information are indicated.

Another important factor is the maintenance of semantic and technical compliance of different systems at local, national and inter-boundary level for the purpose of delivery of high quality health services encompassing the citizens within and beyond the country boundaries.

It is clear from indicated factors that ICT implementation in healthcare sphere requires the conduction of appropriate measures in different spheres (normative, organizational, administrative and technical) at local, regional, national and international levels.

While investigating international practice in e-health sector, it can be seen that one of the goals is the assessment of degree of the implementation of advanced technologies in delivery of e-health services to citizens which accords with European Union strategy and indicators by taking existing problems into account.

III. Electronic Health Strategies of European Countries

In 2000, European Union (EU) member countries have adopted "eEurope 2002" initiative in order to use the facilities of Internet and ICT, and the structuring of European state policy in ICT sector have been commenced [8]. One of the primary goals of this initiative was to stimulate the use of Internet as well as to propose the notions of "online government" and "online medicine". In other words, the maintenance of electronic availability of government services and the facilitation of transparency, openness and availability of medical-sanitary information as much as possible were considered by employing new technologies.

In "Electronic Europe 2005" Activity Plan, strategically goals and activity directions have been specified based on the initiatives in e-government and e-health sectors [9]. Its primary aim is directed towards the provision of access to online government services and content, maintenance of user satisfaction, convenience and multiplatform availability, and the delivery of services conforming to the needs of citizens.

In order to achieve these goals, European Commission has proposed several measures such as the maintenance of availability of access to broadband Internet connection, mutual relations between national systems at EU level, delivery of Pan-European e-government services to citizens, the maintenance of secured information structure.

In 2004, European Commission has approved the first action plan on e-health, and the notion of e-health was described as the application of ICT to all functional fields affecting the health sector [10]. In general, ICT-based tools and services facilitating the improvement of preventive measures, diagnostics, treatment, monitoring and administration are included here.

In last 10 years, several plans, strategies, and directives of measures have been adopted and carried out by European Commission on improvement of e-health sector [8-12]. "E-health Action Plan 2012-2020" adopted in 2012 is directed towards the elimination of the obstacles for complete and effective improvement of e-health in compliance with goals of "Europe 2020" program dwelling upon the large potential of e-health [12].

One of the main goals of Action Plan is the creation of route map of e-health based on adopted framework programs, and the maintenance of four-stage (legal, organizational, semantic and technical) relations according to mutual exchange program [12].

The application and improvement of ICT in e-health sector is carried out by *DG INFSO* competence group (*ICT for Health Unit of DG INFSO*) at the level of European Commission [13]. This group supports rich data base which stores the information on all issues of EU level policy and scientific research.

Currently, the development tendency existing in e-health sector in European states can be divided into several stages: member states shape the strategy for e-health sector; the standards are developed for Electronic Health Records; moreover, large volume of Patient Summaries are already being stored in 4 countries; e-prescription service is being applied in 3 countries and other countries are at the stage of realization; telemedicine is experimentally implemented at the regional level in Northern European countries; legal issues are being prepared in some countries. Large-scale Pan-European pilot projects are developed based on Electronic Health Records standard [13, 14].

In general, it can be noted that European Commission plans to facilitate the establishment of effective and operative state services in member countries, especially, the expansion of health system by developing this strategy with ICT implementation.

IV. Electronic Health Services in G2C

G2C (government to citizens) – mainly covers e-taxes, e-employment, e-voting, e-health and other citizen centric services by reflecting the mutual relations between citizens and the government [1-3]. Amidst the services mostly required by citizens in everyday life in G2C sector, information, education, e-employment, and e-health services can be mentioned. E-health – is considered as an important sector which is being developed at the crossroads of medical informatics, medicine and business sectors. E-health is attributed to such health services and information which encourages the formation of new information relations between patients and medical experts by being enhanced with the help of Internet and technologies, and by increasing efficiency and raising the quality. At the same time, e-health services are associated with medical experts or medical staff as G2C services.

Citizens or potential patients search for satisfactory and reliable e-health services proposed by medical experts via appropriate web-pages. E-health provides a mutual link between citizens, patients and medical facilities for information transfer. Several matters such as methods of specification of treatment or therapeutic schedule, the

accuracy of diagnoses, discovery of experts or entities with desired expertise are attributed to mostly encountered problems [3].

In practice, the following can be considered as the e-health applications:

- Electronic health records;
- Electronic health card;
- E-prescription service;
- Medical information network;
- Telemedicine services;
- Portative systems;
- Portals specialized in health.

Alongside, several other ICT tools can be mentioned for disease prevention, diagnostics, health monitoring and the management of the quality of life.

EU-funded ICT based research and innovation projects addressing societal challenges in the areas management of chronic diseases, surgical treatment to the recovery phase, such as an epilepsy project, a stroke recovery project, patient safety in robotic surgeons', various projects to help surgeons in making critical decisions and other projects from which patients can benefit [15].

One of these projects is EPILEPSIA [16]. The main goal of the project is a brain-computer interface to help patients stay in control. EPILEPSIAE project works to improve the safety and the quality of life of epilepsy sufferers. The project researched the technology of brain-computer interaction and developed an intelligent system capable of collecting and analysing patient's data and predicting epileptic seizures for patients. Also, this system was integrated in a prototype of a transportable alarm device which helps the patient to be in control of their health status.

Another important project is CONTRAST [17]. This project is based on the human (inter) face of ICT to aid recovery after a stroke. This project addresses the gap between clinical rehabilitation and patient support at home.

In practice, the system works as the patient sets up a headset with electrodes connected via the Internet to the assigned doctor's headset. On the basis of information collected from different sources (EEG, heart rate,) the doctor decides together with the patient on the most appropriate training, addressing attention and memory, in particular. Note that results from this project have started being translated into commercial products but some modification is needed for it to be turned into a fully available product in health services as well as in G2C services.

Another interesting project is PASSPORT which was EU funded [18]. The PASSPORT project is based on using a virtual liver for real patient safety. In this project developed a "virtual liver" that helps surgeons take critical decisions on operating process. Note that, it is a virtual reproduction of the patient liver that enables the surgeon to obtain much of the needed information to decide on the treatment and programme the patient's recovery. One of the interesting moments is that the virtual software used in the project is based on open source technology available online making it easier for doctors to collaborate and share their analysis. In 2015, this service is used in more than 10 different countries with more than 200 clinical cases modelled during the last months. The main benefit is the possibility to optimize preoperative planning, significantly improved in more than 20% of clinical cases.

One of the important citizen-centric services is a STORK project [19]. This project makes it possible for EU citizens who are resident in a Member State other than their own or work in one country and live in another one to access online public services, wherever, they are located. This project proposes a solution to make it easier for EU citizens to access the relevant public service online wherever they are located. Using electronic-identity authentication system, citizens can access their national electronic identities in any Member State that was participating in STORK. Public institutions can connect their services to the European e-ID interoperability platform in key areas like e-banking, e-health, public services for business and eLearning and academic qualifications.

Considering the international practice, the implementation of special cards is being initiated for the use of health services for the purpose of delivery of e-services in G2C sector in several countries [3, 20, 21]. It is no coincidence that e-health records are considered as the main element of e-health concept in several European countries. Structured electronic medical documents are collected in these records based on the information transferred from distributed data bases. European epSOS system aimed to design, build and evaluate a service infrastructure that demonstrates cross-border interoperability between electronic health record systems in Europe. The epSOS project was completed in 2014 was indicated as one of the popular projects of information transfer on patient and e-prescriptions [22]. In other words, epSOS supported the convergence of the e-health progress in the EU by cooperating and providing the e-health network with the epSOS data set of the Patient Summary and e-prescriptions. Thus epSOS helped the process towards interoperable healthcare in Europe.

Note that, as a next step, the different countries and governments may take up other initiatives to approach the overall goal of establishing an interoperable cross-border healthcare system across Europe. Already, many initiatives (national and cross-border) base their work on epSOS results highlighting that the efforts in epSOS support public healthcare systems and cross-border exchange of information between them, in Europe. epSOS project will ensure that its deliverables are providing enough practical guidance and recommendations on how to make use of its results in a more long-term implementation [22].

At present, a unified information system is being implemented in health sector in Europe. Within the framework of e-government, the establishment of information system segment is considered based on this program. Experts estimate the investment requirement of e-health sector as 21,6-43,2 million dollars in next 10 years. Electronic health records, national information structure in health sector, regional health

information organizations (RHIOs), electronic exchange of medical data are attributed as the main priorities of currently conducted works [20].

In European practices, electronic health records are real-time, citizen-centric, patient-centred records that provide immediate and secure information to authorized users. Electronic health records typically contain a record of the patient's medical data, results of diagnoses and treatment, medications, information about allergies and immunizations, as well as radiology images and different laboratory results. That is why an electronic health record system plays a vital role in universal health coverage by supporting the diagnosis and treatment of patients through real-time, comprehensive and timely patient information at the point of care.

In World Health Organization (WHO) Report 2016 it is described through project funding, studies, research and policy initiatives that, the EU is active in developing and supporting cross-border interoperability of e-health [14].

As a result, use of international standards to support national electronic health record systems promotes interoperability with other health-oriented ICT systems and with cross-border health services. In order to engage inter-sectoral partners and patients, integrating with public services, especially G2C services, it is important to better understand the need for sharing health information in the process of electronic health record system development. Also note that appropriate national legislation governing electronic health record systems and their use by the full health and social care team should be defined and local, regional systems should be integrated with national systems.

Note that, most e-Government projects in this area are aimed at facilitating exchanges of information and helping medical staff, doctors, concentrate on care and treatment and nothing else. For example, faster registration of patients at treatment centres.

One of the main stages in the restructuring of the relationship between patients, health and social care staff, professionals and public sector, especially G2C services are obviously the implementation of digital technologies and the creation of e-health cards, also referred as electronic health records.

To summarize the current state of the most important implementations of e-health cards use in the healthcare sector in European countries, France was the first country to launch large scale use of smart cards in the healthcare sector with the Sesam-Vitale system in 1998 [23,24]. The e-health cards initially included only some information about health insurance, but later on complementary health insurance administrative data information was added. Starting from 2007, "Carte Vitale 2" card is implemented in France. These cards are prepared in compliance with IAS (Identification, Authentication and Signature) standard. Vitally important information is stored in this card and it is used as a unique tool for accessibility of health services.

In Germany the health card, *electronische GesundheitsKarte (eGK)*, is often described as one of the largest IT projects in the world, and aims to redesign the way health services are provided. Germany started considering the use of e-health cards in 1996 with the *Krankenversichertenkarte*, and started development in 2000. In order to benefit from health services, electronic records are implemented in Germany [24]. By creating appropriate infrastructure in healthcare system, the implementation of these cards, the issue of e-prescriptions, also the use of patient's electronic records is facilitated. The card can be used with administrative, as well as medical software. By supporting card operations, administrative software provides the solution to such problems as the control of entitlement to benefit from services and possession of insurance and the issue of electronic prescription. The availability of medical tools for the patient during the use of card is provided by mutual consent. This facilitates the access to information regarding electronic medical records, chronic diseases, allergic reactions and relevant medical data.

Sweden carried out a pilot project related to the large-scale implementation of e-health records [25]. The information regarding patient's individual information, pathologies, chronic diseases, allergic reactions, analysis results, certificates on prescribed medications and other related information is stored in the card. In Sweden, a special organization was created for the project management within the framework of national e-health strategies. It must be mentioned that the organization is directly responsible for the issue of e-prescriptions. The cost of implementation of information technologies in the health sector in Sweden constitutes 2-3% of healthcare spending.

Healthcare systems in Estonia are considered one of the most developed in Europe [20]. The government of Estonia has created a central database of health and social care. The central database collects information about each inhabitant of the country. The database records disease history from birth till death. Citizens' ID card enables confidential authorised access to all medical information regarding the patient via the www.eesti.ee portal.

Countries throughout Europe are providing their citizens with e-health cards. Some use the cards in their national healthcare programs. Others have smart card based national ID programs as described in Table 1. [23, 24].

From the point of view of practices across European countries, despite the many benefits and opportunities of e-health cards in the healthcare sector, smart cards have not achieved large-scale use due to several limitations. One of the main limitations is the cost of replacing the existing infrastructure, with e-health card health information services. Furthermore, e-health card implementation requires the agreement between all those involved: public authorities, healthcare providers, health insurance companies and the citizen, and thus demands greater cooperation.

TABLE. I. EXAMPLES OF E-HEALTH CARD IMPLEMENTATION IN EUROPEAN COUNTRIES

Country	Card	Number Deployed	Launch Year	Smart card uses
Austria	e-card	11 million (patient) 24,000 (professional)	2005	e-Prescribing Insurance check, e-Referral, eGovernment
Belgium	Social system identity	11 million	1998	e-Prescribing Insurance check, e-Referral eGovernment
France	Sesam Vitale Sesam Vitale-2	60 million (total)	1998 2007	e-Prescribing Insurance check, HER
Germany	Gesundheitskarte (health card)	80 million 375,000 professional	2006	e-Prescribing Insurance check, Medication Log, EHR, e-Referral
Hungary	MOK, Hungarian Chamber of Doctors	40,000	2006	Insurance check, HER eGovernment
Italy	Carta Nazionale dei Servizi (national service card)	3 million	2004	e-Prescribing Insurance check, EHR, e-Referral
Slovenia	Health insurance card	2 million (patient) 70,000 (professional)	1999	Insurance check
Spain	Carte Santé	5.5 million	1995	Health care Insurance check
Estonia	ID-kaart	1.2 million	2002	e-Prescribing national health insurance card eGovernment

Currently, the European Health Insurance Card (EHIC) entitles the holder to emergency healthcare cover in European countries. The card was phased in from 1 June 2004 and throughout 2005, becoming the sole healthcare entitlement document on 1 January 2006. Nowadays, a free card that gives citizen access to medically necessary, state-provided healthcare during a temporary stay in any of the 28 EU countries, Iceland, Lichtenstein, Norway and Switzerland, under the same conditions and at the same cost (free in some countries) as people insured in that country [26].

Implementing ICT into health system adds value to the services and to the skills of the medical staff using them. Although significant progress is being made towards citizen-centred service models in Europe, critical gaps in the design and delivery of health services remain. Note that, most importantly, effective national health reform needs to adopt the perspective of the patient as citizen in order to understand how e-health tools and services can be used to facilitate better care.

It is clear that the usefulness of e-health in terms of improvement of efficiency of services for both citizens and government, the adequacy of treatment, the reduction in management spending are not possible without the integration of clinical-medical, economic and administrative indicators. From this point of view, timely, accurate, complete and clear information is needed for the analysis of costs and effectiveness, assessment of existing strategy and the conduct of national and international comparisons.

This research facilitates the discovery and sharing of best practice, assessment of quantitative and qualitative indicators of e-health, control of the efficiency of budget resources, determination of successful medical facilities and the improvement of indicators characterizing citizens' health. Moreover, the current information in citizen-oriented systems must be accessible for medical facilities as well as patients, so that both parties can use it in accordance with their needs and requirements.

V. ADVANTAGES AND PROBLEMS OF E-HEALTH

E-health has significant impacts on the lives of citizens, the working conditions of health personnel (doctors, healthcare professionals and administrative staff) and the activities of health offices and whole e-government system.

In general, the outlook for e-health covers several perspectives: technological, research, economic, political, international cooperation and stakeholders.

Some advantages of e-health technologies are the following:

- Electronic health data is useful for doctors, practitioners and scientists for treatment and research purposes.
- With the advent of innovation and new technologies, multimedia files and data in the form of pictures, videos, and text can be shared in real time on a range of platforms;
- E-health services may play an important role in having more balanced proportion between medical staff and patients all round the world;
- Multi-location real time video-conference can be used to conduct training sessions, treatment of diseases, collaborations, and more.
- E-health resources can be accessed by large number of people to study and to get knowledge about health related issues at their own convenience.
- Possibility using satellite based medical diagnosis and care of elderly in their homes and other.

Main problems of implementation of e-health are the following:

- Lack of concept and national strategy in health sphere specially in developing countries;

- Uncoordinated efforts between government, public administration and medical facilities;
- Lack of e-health services in realizing strategy of e-government, as well as in G2C segment;
- Inadequate financing and ineffective organizing;
- Lack of legal framework;
- Confidentiality, privacy, protecting personal information and other issues;
- Lack of universal software platforms in health system in different medical facilities.

VI. CONCLUSION

Currently, a single world practice does not exist regarding the establishment of e-health systems. Even in developed countries several models are being proposed and trialled for creation of government e-health systems. One of the biggest problems for e-health is the matter of administration of medical facilities and management. Integration of daily data can be very supportive in this regard. The investigation of international practice and the implementation of best practices can help to eliminate such problems.

In this article, e-health is considered as a segment of e-government while international practice and adoption of e-health strategies are considered. Perspectives of the delivery of e-health services in G2C are explored and explained. The study of best practices in this field in future research works will facilitate the expansion of citizen-centric e-services.

REFERENCES

[1] M. Domenichiello, "State of the art in adoption of e-health services in Italy in the context of European union e-government strategies," Procedia Economics and Finance, No 23, 2015, pp.1110-1118. www.sciencedirect.com

[2] R.M. Alguliyev, F.F. Yusifov, "Some actual scientific and theoretical problems of formation of the electronic government and the perspectives of their solutions," Scientific-practical journal of Problems of Information Society, No 2, 2014, pp. 3-13.

[3] W. Zhang, Y. Zhu, and et al., "Personalized Recommendation of E-Health Services Based on Mutual Information," International journal of innovative computing, information & control, vol. 11(3), 2015, pp. 903-919.

[4] J. Mitchell, "Increasing the cost-effectiveness of telemedicine by embracing eHealth," Journal of Telemed and Telecare, vol. 6 no. suppl. 1, 2000, pp. 16-19.

[5] E-health in practice, www.euro.who.int

[6] World practice of establishment of electronic health in governments, www.cnews.ru.

[7] E. Ronchi, J. Adler-Milstein, G. Cohen, L. Winn, and A. Jha, Better Measurements for Realizing the Full Potential of Health Information Technologies, Chapter 1.7 in the 2013 Global Information Technology Report, World Economic Forum, www3.weforum.org

[8] Council and the European Commission, eEurope (2002): An Information Society for All, 2000, Brussels. www.umic.pt

[9] European Commission, eEurope 2005: An Information Society for All, COM (2002) 263 final, 2002, Brussels. www.etsi.org

[10] European Commission, eHealth - making healthcare better for European citizens: An action plan for a European eHealth Area, COM (2004) 356 final, 2004, Brussels. http://ec.europa.eu

[11] European Commission, i2010 An European information society for growth and employment, COM (2005) 229 final, 2005, Brussels. http://europa.eu

[12] European Commission, eHealth Action Plan 2012-2020 - Innovative healthcare for the 21st century, COM (2012) 736 final, 2012. Brussels. http://ec.europa.eu

[13] Analysis of development of electronic healthcare in Europe, http://zerde.gov.kz

[14] From innovation to implementation, eHealth in the WHO European Region, Report, 2016, www.euro.who.int/en/ehealth

[15] EU funded societal challenges projects, https://ec.europa.eu

[16] Project EPILEPSIAE, A brain-computer interface to help you stay in control, https://ec.europa.eu

[17] Project CONTRAST, The human (inter)face of ICT to recover after a stroke, https://ec.europa.eu

[18] Project PASSPORT, A virtual liver for real patient safety, https://ec.europa.eu

[19] Project STORK, Take your e-identity with you, everywhere in the EU, https://ec.europa.eu

[20] Electronic healthcare as a factor of development of quality and availability of health services to population, 2014, www.dompressy.by

[21] Foreign practice of solution of electronic government building and delivery of government services, http://aisup.economy.gov.ru

[22] epSOS results & outlook, www.epsos.eu

[23] Smart Cards and Healthcare Providers, www.smartcardalliance.org

[24] A.P. Keliris, V. D. Kolias and K.S. Nikita, Smart Cards in Healthcare Information Systems: Benefits and Limitations, IEEE 13th International Conference on Bioinformatics and Bioengineering (BIBE), 2013, www.computer.org

[25] A.A. Kukhtichev, "Electronic health records as a basis of ehealth services of information system "digomed"," NSU News, Series: Information technologies, vol.14, № 1, 2016, pp. 61–75.

[26] European Health Insurance Card, http://ec.europa.eu

On the Dynamic Maintenance of Data Replicas based on Access Patterns in A Multi-Cloud Environment

Mohammad Shorfuzzaman
Department of Computer Science
College of Computers and Information Technology
Taif University, Taif, KSA

***Abstract*—Cloud computing provides services and infrastructures to enable end-users to access, modify and share massive geographically distributed data. There are increasing interests in developing data-intensive (big data) applications in this computing environment that need to access huge datasets. Accessing such data in an efficient way is deterred with factors such as dynamic changes in resource availability, provision of diverse service quality by different cloud providers. Data replication has already been proven to be an effective technique to overcome these challenges. Replication offers reduced response time in data access, higher data availability and improved system load balancing. Once the replicas are created in a multi-cloud environment, it is of utmost importance to continuously support maintenance of these replicas dynamically. This is to ensure that replicas are located in optimal data center locations to minimize replication cost and to meet specific user and system requirements. First, this paper proposes a novel approach to distributed placement of static replicas in appropriate data center locations. Secondly, motivated by the fact that a multi-cloud environment is highly dynamic, the paper presents a dynamic replica maintenance technique that re-allocates replicas to new data center locations upon significant performance degradation. The evaluation results demonstrate the effectiveness of the presented dynamic maintenance technique with static placement decisions in a multi-cloud environment.**

Keywords*—Multi-cloud environment; data replication; distributed algorithm; response time; dynamic maintenance; QoS constraint

I. INTRODUCTION

Cloud computing is appearing as an emerging technology that provides scalable computing system and offers endless opportunities for the computing community. It provides large-scale computing and storage resources comprising data centers [1], [2]. In a cloud computing scenario, computing and storage resources can be delivered as a service irrespective of the location and time as other necessary utilities in life [3]. In general, there are three categories of cloud computing architecture for the delivery of services, namely, Software as a Service (SAAS), Platforms as a Service (PAAS) and Infrastructure as a service (IAAS) are common [1]. In a SaaS architecture, a cloud service provider hosts and manages software applications intended for the end-users instead of letting them using locally-run applications. The IaaS architecture offers hardware and software resources for data storage and processing as well as networks and other necessary infrastructure for deployment of operating systems and applications. In a PaaS architecture, users are furnished with necessary tools and programming languages and the service provide hosts an application delivery platform to continuously support development and delivery of end-user applications [4].

Due to the use of scalable data centers in cloud computing, end users are relieved of the burden associated with application provisioning and management. Popular cloud services providers such as Amazon S3 [5], Google [6], App iCloud[1] , Microsoft Azure[2], and DropBox[3] are serving thousands of millions users through a huge number of servers distributed over many datacenters around the world. Hence, cloud computing infrastructures are being used for effective processing of large data sets without the huge upfront investments required to purchase traditional data centers. Accordingly, there are increasing interests in developing data-intensive (big data) applications in this computing environment that need to access huge datasets. For instance, Facebook, Twitter, and big data analytics applications, such as the Human Genome Project [7], are using cloud computing infrastructures for processing and analyzing their petabyte-scale data sets, using a computing framework such as MapReduce and Hadoop[4].

Data availability and accessing data efficiently is an important demand for these data intensive applications [8]. Furthermore, cloud infrastructure has heterogeneous resources and the resources have diverse performances. Also, there may be demands of different service quality requirements from different users. Besides, overall system performance is also a critical factor. To effectively address these challenges, the need for data replication is apparent. In data replication, data are replicated at different replicas to provide data access to the users in a nearby location. Replication techniques increase data availability and reduce cost of data access and response time [9], [3], [10]. It also distributes the workload to the replica servers by routing user requests to different sites. Replication techniques decrease congestion-related performance degradation probability by distributing the load of network to network of multiple paths. To obtain maximum gain from replication, strategic placement of the file replicas is critical [9], [11], [12].

[1] https://www.icloud.com/

[2] https://azure.microsoft.com/

[3] https://www.dropbox.com/

[4] http://www.slideshare.net/kevinweil/hadoop-at-twitter- hadoop-summit-201

Data replication technique is widely used in the cloud computing system to provide high data availability [13], [14], [15], [16]. While research has been done on replica placement, little has focused on the dynamic maintenance of the replicas over the time. Moreover, until now, replication in multi-cloud environments is not considered in the existing research. Therefore, determining data center locations of replicas in a multi-cloud environment mimicking general network topology is an open problem.

This paper first presents a new distributed algorithm that determines static locations for placing replicas in multi-cloud environments representing a general network topology to improve data availability and satisfy user demands for time critical data. The algorithm also maximizes the degree of satisfied users while minimizing aggregated response time upon data access. The proposed distributed replica placement algorithm provides benefits with respect to reliability and scalability issues. A multi-cloud environment is highly dynamic where user requests and network latency fluctuate persistently. Data centers that hold replicas currently may not be the best locations to obtain replicas later. Hence, an algorithm is proposed for the dynamic maintenance of replicas based on the user data access patterns and the cumulative aggregated response time. Overall, the paper makes the following contributions:

- Unlike existing data replication strategies in cloud computing system, the proposed replication algorithm is targeted for multi-cloud environments.
- A novel algorithm is presented for distributed replica placement in multi-cloud environments mimicking general graph topology.
- A dynamic maintenance algorithm is presented to relocate replicas in optimal locations to sustain overall response time to a minimal.

The rest of this paper is organized as follows. Section II presents some selected existing approaches from the literature. Section III describes the clustering approach to static replica placement. The proposed dynamic replica maintenance technique is presented in Section IV. Performance evaluation is described in Section V. Section VI concludes the paper and suggests directions for future work.

II. Related Work

There is a handful of data replication strategies found in the cloud computing environments. One of the earliest work proposed by Ghemawat et al. [6] where a static distributed data replication algorithm in clouds (Google File System – GFS) places data chunks based on a number of factors such as: 1) to choose replica locations in the chunk servers with below-average disk space utilization; 2) to set a maximum count on the replicas that are created recently on each chunk server; 3) to scatter replicas of a particular chunk across racks. The creation of data chunk replicas is triggered when the number of replicas drops below a specific level. The data in DataNode is protected against failure through replication technique. For instance, in Hadoop Distributed File System (HDFS) [14], to safeguard against failure a data block is replicated twice in different racks containing two DataNodes. The authors in [17] propose a dynamic data replication algorithm which is distributed in nature and targeted for cloud environment. The algorithm works a cloud system based on HDFS technology and decides to place replicas considering the current workload on the nodes and conforming to a lower bound on the number of total replicas that are created. Wang et al. [18] present a prediction based centralized cloud data replication technique that considers weighting factor for replication decision.

The authors in [19] made a number of contributions for cloud data replication as such: 1) came up with a model to link system availability and the number of replicas; 2) identify data items for replication based on popularity; 3) computed an appropriate number of replicas and locations based on an effective system byte rate. An adaptive technique [20] for cloud data replication is proposed which works based on file prediction. The technique takes availability and efficient access into account while predicting files in data centers. Besides replication, Gai and Daio [15] investigate replica consistency issue in cloud computing system. The authors adopt a lazy update scheme to split data access and update that can improve throughput and cut down response time. In addition, the authors in [21] work on a data management technique focusing on reliability issues. The technique works by checking the replicas proactively in an attempt to decrease the replica count which will reduce the storage usage in turn. A data storage mechanism [22] is proposed to enhance data availability and privacy which works by combining multiple clouds and a set of protocols such as Byzantine quorum system protocols, cryptographic secret sharing, and erasure coding. The evaluation results demonstrate that the proposed system brings forth a cost effective way to improve data availability and privacy for critical applications. In a relatively recent effort done by X. Wu [23], a cost effective data set replica placement strategy is proposed for cloud environments. The technique starts with designing a cost model for data management which includes storage cost and transfer cost. Then a replica placement technique is presented that decides the location of replicas using based on a location graph problem. The technique uses access frequency and average response time to decide which data set should be created. Experimental results demonstrate the benefit of the technique in term of lower management cost with fewer replicas.

Even though a quite a bit of work has been done to deal with replication issues in cloud computing environments, less attention was paid to address the issue of QoS requirement. The authors in [24] present two data replication algorithms that address the issue of QoS requirements in cloud systems. First, replication is done based on meeting high QoS requirement on a first come first service basis. Nevertheless, the performance of this algorithm is not satisfactory in reducing the replication cost and the number of unsatisfied requests. Hence, the second algorithm converts the replication problem into a minimum-cost maximum-flow (MCMF) problem which can generate an optimal solution to the problem in polynomial time. This problem has also been studied widely in data grid systems [25], [26], [27].

Even though the issue of dynamic maintenance of replicas in cloud environments is not well studied in the literature, some researchers addressed this issue. Liao et al. [28] propose a dynamic replica deletion strategy for distributed storage systems under cloud computing environments. The strategy aims at reducing storage that is occupied by the replicas and their maintenance cost. The authors come up with a mathematical model to illustrate the relationship between QoS requirement for the requested data and the number of replicas that need to be created. Performance results suggest that the proposed technique can save disk space and reduce maintenance cost while satisfying the QoS requirements. The authors in [29] present a file replication and consistency maintenance technique in Hadoop cluster. As claimed by the authors, Hadoop's strategy to maintain three replicas of each file leads to poor storage utilization and hence they propose integrated data replication and consistency maintenance (IRM) technique. They only create replicas for popular files and this method has already been proven to be an effective technique in the literature. The implementation includes the use of MapReduce programming model and HDFS in the clusters of commodity computers. The experimental results show that the technique can reduce data access time and increase data locality at the same time. A relatively earlier effort was done by Chun et al. [30] that deal with efficient replica maintenance in distributed storage systems. The proposed strategy works by aggregating disk spaces of many nodes over the Internet. The novelty of the strategy comes from the viewpoint of fault-tolerance in case of network and disk failure.

III. Clustering Approach to Static Replica Placement

This section will first describe the static replica placement (SRP) problem in a multi-cloud scenario that minimizes the number of replicas created and aims at meeting the user QoS requirements. The proposed solution starts with a system model that shows the targeted multi-cloud architecture where each cloud provider contains different number of data centers for storing replicas. Then, a static replica placement approach is presented based on clustering data mining technique.

A. System Model

Figure 1 shows a high level architecture of a multi-cloud setup comprising of a number of cloud providers interconnected multi-cloud proxies. User communication with the data centers in the cloud providers are facilitated through the multi-cloud proxies. Before describing the proposed clustering approach to static replica placement in this multi-cloud environment, a logical structure of the multi-cloud system model is presented as shown in Figure 2. The multi-cloud system which is considered here consists of $DC = \{DC_1, DC_2, \dots, DC_n\}$ different data centers and $U = \{U_1, U_2, \dots, U_m\}$ distributed end-users that share both resources and data in a multi-cloud system. This is modeled using an undirected graph $G = (DC, E)$. Here, DC is the set of data centers, and $E \subseteq DC \times DC$ is the set of links among the data centers. It is assumed that there is an upper bound bandwidth constraint $(bc(e_i))$ on each link $e_i \in$ Ecapacity. Also, each data center $DC_i \in DC$ is characterized by the following 4-tuple:

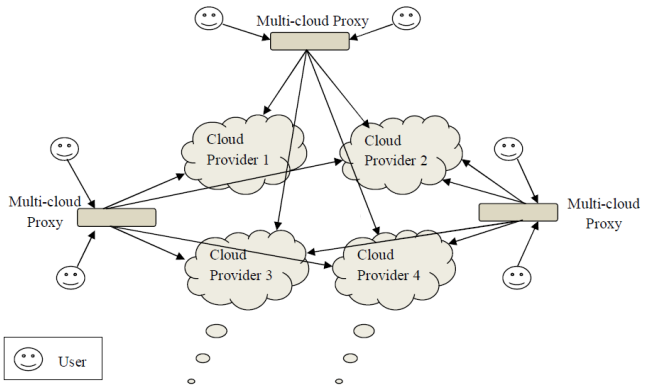


Fig. 1. A multi-cloud architecture comprising a number of cloud providers

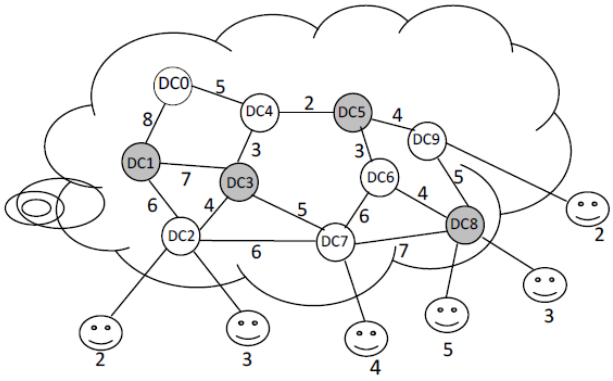


Fig. 2. A logical structure of a multi-cloud system model consisting of data centers and users

$$DC_i: \langle ID, S, C, L, d(u, v) \rangle \quad (1)$$

where ID is the data center *id*, S is the storage capacity of the data center, L is a workload capacity constraint, C is the storage cost to store a replica and $d(u, v)$ is a communication cost over link $(u, v) \in E$. It is assumed that for a same data file different data centers may have different storage costs. A data center containing a replica can be considered as a replica server.

The workload capacity constraint is defined as an upper bound on the user requests' number that is processed by the replica server during a specified time period. The replication strategy has to ensure that the user requests are satisfied while limiting the workload of each replica server to its capacity. Associated with different replica servers the workload constraint can be different. A replica server is overloaded if the total workload of the replica server becomes greater than the server's capacity constraint. In addition, as shown in Figure 2, end users are connected to the data centers. Note that each replica server (data center) provides services to multiple users and a user $U_i \in U$ is always associated with a specific server $DC_i \in DC$ in the multi-cloud structure. The number on the link denotes the cost of communication. It is considered that the logical graph is connected. Therefore, a replica server is able to communicate with any other server in the multi-cloud via some path.

A user $U_i \in U$ sends his/her requests to its associated server for retrieving data specifying QoS requirements. The

QoS is defined as an upper bound q (U_i) on the cost retrieval. Note that QoS requirements may be different for different users. The requirement specifies that all the requests by U_i will be handled by some servers by a communication delay q (U_i). The requirement is satisfied when a request is handled by the closest replica server which satisfies q (U_i). The requirement is violated otherwise.

Now, the QoS-aware static replica placement problem in hand is precisely defined. Let DC' represent a set of data centers that contain the replica of a requested file. Let $min(d_i)$ be the minimum communication time for the data center i to retrieve a requested file from all the probable retrieval paths from i to the set of replicas DC'. In case a data center holds the target replica, the value of $min(d_i)$ becomes 0. To this end, our goal is to calculate a replica set of minimum size, $DC' \subseteq DC$ such that the retrieval of the requested file at each data center in DC meets the QoS requirement, i.e., $min(d_i) < q$ (U_i) for each i in DC. Here, the communication cost between the users to the local data centers is not counted since the cost does not change the decision of replication.

B. Data Model

A central database (Central DB), located at the origin server (represented as $DC0$ in the graph) stores all the data required by the cloud applications. To reduce the response time, each data center maintains a local database, called datacenter database (Datacenter DB). The goal is to replicate the most frequently used data items from the central database. For maintaining consistency of data, the origin server sends updates to each replica server. It is assumed that updates must first be performed on the authoritative copy stored at the origin server, where they are then propagated to the replica servers. That is, updates only come from the origin server and the replica servers act as repositories for data retrieval. Popularity of a file is calculated using access rate and it is assumed that popularity persists into the near future.

In order to retrieving data a user $U_i \in U$ sends his/her requests to its associated data center specifying some QoS requirements. The request is processed locally if the data center contains the requested data. Otherwise, the request is forwarded to the nearest data center that contains the replica of the requested data. Thus, associated with every user $U_i \in U$ will be a non-negative weight $count(U_i)$ representing the traffic (data access counts which represent the data popularity) originating from that node over a certain period of time. These data access counts will contribute to the response time for servicing the data requests from the users.

When data is accessed, the information about requesting data center is stored. Additionally, the statistics showing the number of accesses and updates are obtained for each data item. The access rate or popularity is calculated in terms of the number of accesses over a period of time. The popularity of a data item varies over time in relation to the types of stored data. Generally, a newly created data has the highest popularity. Then, the access rate decreases over time. For example, a newly posted YouTube video is watched by most of the visitors. However, over the time its popularity and the number of visitors start to decrease.

C. Clustering Approach to Replica Placement

Let each data center in the multi-cloud be i which by now is acquainted with its directly adjacent data center. In addition, this data center i also knows about the communication delay to reach any neighboring data center k in the multi-cloud structure through the $i \rightarrow k$ path. Furthermore, a data center always maintains information about these direct adjacent data centers that store replicas of the requested data and the in-degree of each of them. The in-degree of a data center having replicas determine the percentage of other data centers that are connected to users and their requests are satisfied by this particular data center.

Each data center, i, connected to a user knows its QoS requirement (i.e., q (U_i)) to retrieve a data item. The data center incessantly checks its own status and the status of its neighboring data centers whether a replica of the requested data item exists in its own storage or at the storage of any of its neighboring data centers that can be communicated through a delay $\leq q$ (U_i). In course of time, if data center i or any of its neighboring data centers reachable within the stipulated delay becomes devoid of replica, any of the following two will occur. First, data center i will become a cluster core and it will copy the requested replica in its own storage if all its neighboring data centers are away with a delay $> q$ (U_i). Alternatively, data center I will select a core for the new cluster from any of the neighboring data centers that can be reached within q (U_i) delay and it has a highest in-degree. This core data center will now store replica to be accessed by other data centers that are connected to the users. This strategy tries to increase the count of data centers that can read this latest copy of replica.

IV. Dynamic Maintenance of Static Replica Placement

Clustering approach to replica placement ensures that replicas are created in near-optimal locations based on the user requirements for the current session. However, the data centers currently holding the replicas may not be the best locations for replicas to be there in future due to the changes in user requests and network conditions. Hence, dynamic maintenance of replicas which means relocation of replicas in optimal locations is necessary to sustain overall response time to a minimal. At the same time, the goal is to maximize the percentage of users whose QoS requirements are met.

A. Problem Formulation

The dynamic replica maintenance problem is now formulated in a multi-cloud environment. Formally, the problem of replica maintenance problem is expressed as an optimization problem as follows:

$$minimize_{\ response\_time} \ \& \ maximize_{\ user\ QoS}$$

The problem is to choose M replica locations among N potential data centers ($N > M$) by minimizing the overall response time and maximizing the rate of satisfied user QoS requirements considering a given traffic pattern (i.e., access frequencies' recurring pattern of users for different data items). More specifically, the goal is to identify a set of new data center locations with the minimal response time where

services for each user request can be provided by a data center within its quality requirement.

Now, the replica maintenance problem in hand is generalized. It is assumed that in cloud applications, updates to the data are infrequent and that the consistency can be more relaxed than in, for example, high-performance commercial databases. Given this, the proposed approach achieves greater scalability while making modest compromises in terms of update propagation and replica synchronization. In particular, the updates are delivered from the origin server (data center) to all other data centers having replicas via application-level multicast, in which each server (data center) receives the updates from its parent and is responsible for further distributing the updates to its children. Without loss of generality, it is assumed that this origin server constitutes the root of the tree that is going to be embedded. Given the embedded tree over the general multi-cloud, the replica maintenance problem can be modeled as a dynamic programming problem and its solution can be obtained in a distributed fashion. To solve the problem in hand using a dynamic programming approach, it is needed to solve different parts of the problem (also called sub-problems), then combine the solutions of the sub-problems to attain an overall solution. This approach tries to solve each sub-problem only once, thus reducing the number of computations. Hence, the new dynamic programming problem is a generalization of the replica maintenance problem we intend to solve, and thus the solution to this new problem also solves the original replica maintenance problem.

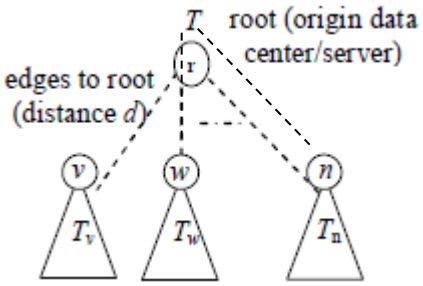


Fig. 3. An embedded tree with possible sub-trees representing the sub-problems

Let $T = (V, E)$ be the embedded tree over the general multi-cloud rooted at the origin server. T can be viewed as a combination of a number of sub-trees such as Tv, Tw, and so on where each sub-tree is connected to the origin server with a list of edges of length d. Now, the replica maintenance problem in T can be solved, by first solving an extended problem in a slightly different tree built from each sub-tree such as T_v and an additional edge list connected to it as shown in Figure 3. The edge-list is of length d and one side of it is connected to the root of T_v and the other side to the origin data center. At any data center from v to the root, there is a replica server data center containing the copy of the data file which can be accessed by the sub-tree T_v. Now, the total response time for sub-tree T_v can be calculated by considering the replica server at any distance towards the root. This response time considers the data access cost from the edge-list. Thus, this became a generalization of the original replica maintenance problem: when the edge-list length is zero and $T_v = T$, the replica maintenance problem is reduced to the original one.

Now, we formally formulate our goal for finding an updated set of replicas that has minimal response time so that QoS requirement of each user v is satisfied by $RU\{r\}$, i.e.,

$$\min_{s \in RU\{r\}} d(v, s) \leq q(v)$$

where, d(v, s) denotes the distance between v and s.

B. Dynamic Replica Maintenance Strategy

In dynamic replica maintenance strategy (DRMS), a shortest path tree is built at the onset. As noted previously, the updates are forwarded from the origin server data center to all the replica server data centers using application-level multicast. It is assumed that this origin server constitutes the root of the tree that is going to be embedded. To do this, the all-pairs shortest paths are calculated and a shortest path tree is built which has root, the origin server (see Figure 4).

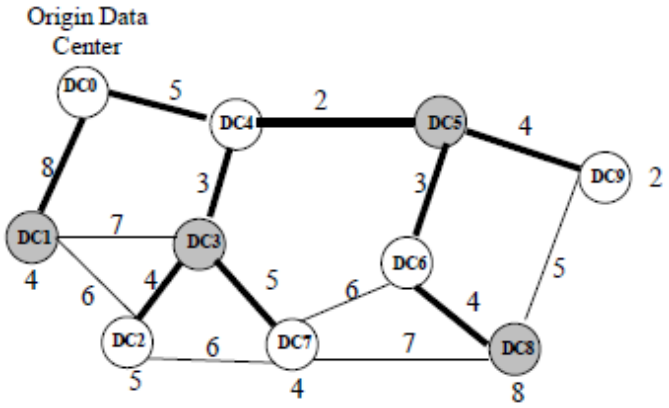


Fig. 4. An embedded all pairs shortest path tree in the multi-cloud system

Given the embedded tree over the multi-cloud model each data center node can determine the cost for generating a local replica and the cost of data transfer from a replica server data center anywhere towards the root. Then a parent data center node assesses the cost of creating a local replica versus the cost of data transfer using the results provided by its children. This process continues until it reaches the root data center of the embedded tree. Then, based on the assessed data, actual replica placement starts at the root data center and finishes at the leaves.

Let $RMC(v, rd)$ represents the replica maintenance cost contributed by the sub-tree with the root data center v when the replica is placed at a distance of rd towards the origin data center. The value of rd can vary from 0 (when the replica is in v itself) to the distance to the origin server. When rd is equal to 0 the replica maintenance cost considers read cost of all the descendants of data center node v, the storage cost, and the update cost for the replica at node v. On the contrary, when rd is greater than 0 (i.e., the replica is placed to any of the ancestors of node v except the root data center), RMC (v, rd) denotes the cost of replica maintenance for the sub-tree rooted at data center v that contains all its descendants' read cost.

In Figure 4, it is considered that the number of updates originated by the origin data center is 3 for a specific time period and the storage cost for the data file in the replica server is 10. As before, it is also assumed that accesses and updates need one unit of bandwidth for per file transfer per network hop. Now, if at "node 3" a replica is placed, the replica maintenance cost for the sub-tree with the root data center 3 is *RMC* (3, 0) = 40(read cost) + 24(update cost) +10 (storage cost) = 74. However, if at "node 4" a replica is placed, then the corresponding cost will be *RMC* (3, 1)= 67(read cost) + 0(update cost) +0(storage cost) = 67.

Now, in order to determine the locations of new replica based on changing popularity and replica update frequencies, the dynamic maintenance algorithm is called at regular intervals. If old replicas are still in a newly found replica set they are retained. From the old set, other replicas are removed and new replicas are generated as required. Let *OR* represents the set of replicas that are created by the distributed static replica placement strategy and *NR* represents the newly calculated updated set of replicas during maintenance process. Thus, the dynamic maintenance algorithm will create replicas contained in the set $NR - (OR \cap NR)$. The selected interval is calculated by the rate of request so that a short interval results for high arrival rates. This produces higher overhead but adapts more quickly in changing access patterns.

Calculation of replica maintenance costs by the terminal, non-terminal, and root data center nodes and the determination of new replica locations based on these calculated costs are done in a bottom-up fashion starting from the terminal data center nodes. Each data center, *v*, calculates the optimal replica maintenance cost for its sub-tree considering the replica location at any distance from *v* to the root data center. Then data center *v* determines the location of the replica whether it should be fetched from any data center located up on the path towards the root, or it should be created in its own storage based on the calculated replica maintenance cost. In case *v* is a non-terminal data center node, it is also possible that the replica should be created in children data center nodes of *v* if placing replicas in them results in lower replica maintenance cost. Data center *v* records these costs and data center locations for potential replica creation. It is important to note that a data center node commences the calculation of costs once all of its children nodes have finished calculating the same.

Now, the actual determination of new replica locations occurs in a top-down fashion starting from the root data center. More specifically, a data center will determine if it needs to create a replica on its own storage or not. The root data center starts the process by determining the minimum of the two replica maintenance costs calculated before and its associated replica location either in itself or in the children data center nodes down the hierarchy. Root data center node then forwards this location information to all of its children. Upon receiving this location information, each of the children data center nodes verifies it with its location value calculated before. If they match, a replica is created in its own storage and this information is further forwarded down the hierarchy. Otherwise, it forwards an updated location information that was received from the root data center to its children and this process continues till the bottom of the embedded tree is reached.

V. Performance Evaluation

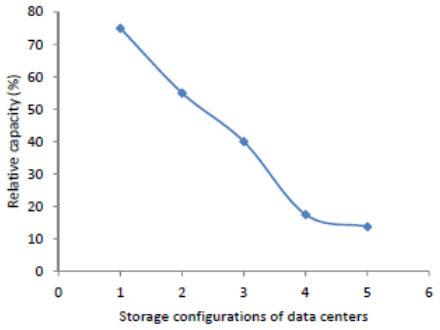
This section describes the performance analysis of the dynamic maintenance algorithm and compares it with the clustering based static placement as factors such as data center storage capacities, data access patterns, and user QoS constraints are varied.

A. Simulation Method

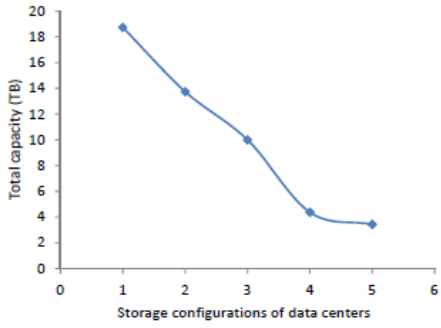
For the assessment of the algorithms, a Java based simulator program is used. The detailed configurations for the simulation are given below.

a) Testbed Environment

In the experiment, a multi-cloud composed of 156 data centers is used. Waxman model [31] is used to generate the multi-cloud topology. The available link bandwidth is computed using a uniform distribution with the range [0.622, 2.5] (Gbps). Data center storage capacities are also determined using the same distribution. The number of data files used in the system is 2500 with each data file size equal to 10 GB. This makes the total size of all data files used in the simulation approximately 25 TB. To measure the effectiveness of the algorithm a wide range of data center storage resource configurations are used in terms of the relative storage capacity, *r*, of the replica data center servers. Here, *r* is defined as a ratio of the total storage size of replica data center servers to the total size of all data files in the system. If *r* is 100%, it can be assumed that every data file could have a replica in the system. For the experiments *r* has been varied from 75% to 13% as shown in Figure 5.



(a)



(b)

Fig. 5. Relative storage capacity and total capacity for different data center storage settings in (a) and (b)

Each replica data center server can serve a number of data access requests from the users. The replica servers will run short of storage during the simulation. To place new replicas, a replacement strategy is necessary to ensure that new data files do not replace popular files. To this end, a modified Least Recently Used (LRU) replacement strategy was used based on the popularity of data to ensure that no replicas created in the in-progress replication period are removed.

B. Data Access Patterns

A number of data files are accessed by each job during the simulation. The simulation was conducted with 50 different jobs that were submitted with fixed probabilities. Some jobs were more popular than others. The data access requests from the users follow Poisson arrivals. Each user issues one access request on average per 2500 milliseconds and a data access pattern determines the sequence of the access requests. Two access patterns namely Gaussian random walk and the heavily tailed Zipf distribution were used. The Zipf distribution is given by: $P_i = K$, where P_i is the frequency of the ith ranked item, K is the popularity of the most frequently accessed data item and s determines the shape of the distribution. It is assumed that data access patterns can show temporal locality to some extent which means that recently accessed data are expected to be accessed again. Such an access pattern containing varying amount of temporal locality can be generated using Zipf distribution. Thus, in a system that is designed to react to file popularity, the Zipf distribution offers a natural testing ground. The index used to measure the amount of locality in the pattern is denoted by s. The observed parameter values are in the range of $0.65 < s < 1.24$. A higher value of s indicates an increased degree of locality. In this paper, we use $s = 0.85$ and 1.0 and refer to as Zipf-0.85 and Zipf-1.0 distribution respectively. Furthermore, Gaussian distribution is the most widely used family of distributions in statistics and many statistical tests are based on the assumption of normality. As such, it is a good base measure which can be used for easy informal comparison to known applications [27, 28].

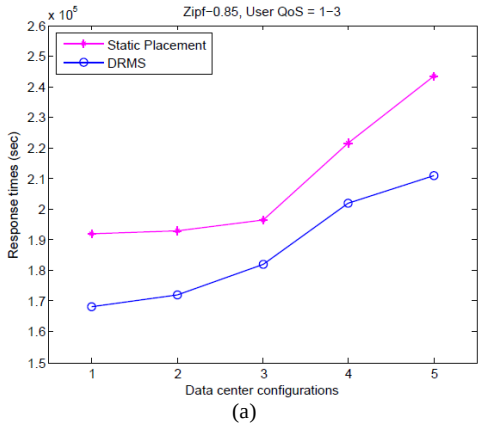
a) Performance Metrics

Each user data center site keeps record of the time required to receive a data item once it is requested. This elapsed time constitutes the foundation of assessing and comparing various replication strategies. Our dynamic replica maintenance strategy was assessed using the performance metrics which primarily include total response time in respect with job execution. Response time refers to the time that elapses from the moment when a data is requested until it is received and the specified job completes its execution. Total response time aggregates the response times of all the executed jobs for a simulation period. The goal is to achieve minimum total response time for our dynamic replica maintenance algorithm. The performance metrics also considers user satisfaction rate. User satisfaction rate is the proportion of users whose QoS constraints are met. The absolute values are actually of little interest but the relative performances demonstrate the superiority of dynamic maintenance algorithm over the static counterpart.

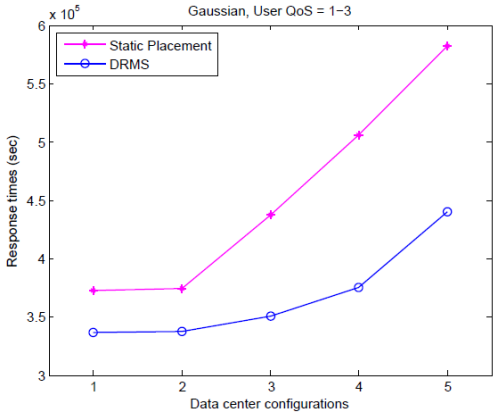
C. Results and Discussion

This section presents the experimental results of the static replica placement strategy and the dynamic replica maintenance strategy (DRMS) and compares them thoroughly based on the performance metrics. For a specific user, its QoS requirement is taken as a distance from the user to the closest replica data center server (such as number of hops) using a uniform distribution (i.e. the distance requests are uniformly distributed over the range). For example, a user QoS requirement of [1-3] implies that the closest data center with requested replica from the user should be any value between 1 and 3 and in such case the stipulated user QoS constraint deems to be satisfied.

We start by considering the algorithms' performances in terms of total response time, the major concern from the viewpoint of the data consumer. Figure 6 shows the approximate values of response times (y-axis) as a function of varying data center storage capacities (x-axis) for static replica placement and DRMS. In the experiment, a moderate workload capacity is considered for the replica server data centers. It is taken from a uniform distribution of [100-200] in terms of GB. The user QoS constraint on replica server data center distance of [1-3] is specified from a uniform distribution to allow relatively relaxed range. Our dynamic maintenance strategy DRMS that considers the relocation of replicas generally performs better than the static replica placement model in terms of response times for both Zipf and Gaussian data access patterns. The reason is that DRMS creates a modest number of well-placed replicas compared to the static counterpart which substantially reduces data access latency. Consequently, this decreases the overall response time. In addition, low running time of DRMS contributes to the reduction of its overall response time. With the decrease in storage capacity of data center replica servers the response time increases due to the creation of lower number of replicas.

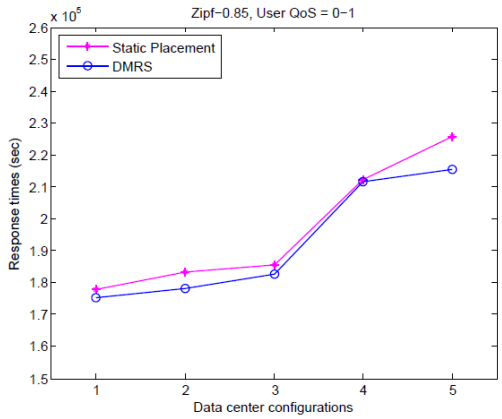


(a)

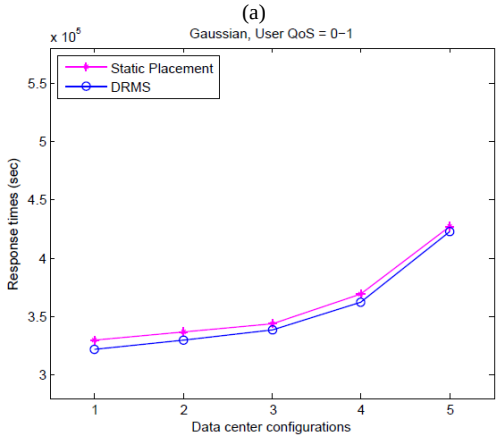


(b)

Fig. 6. Response times of static replica placement and DRMS considering relaxed QoS requirement [1-3] for Zipf-0.85 and Gaussian access patterns in (a) and (b)



(a)

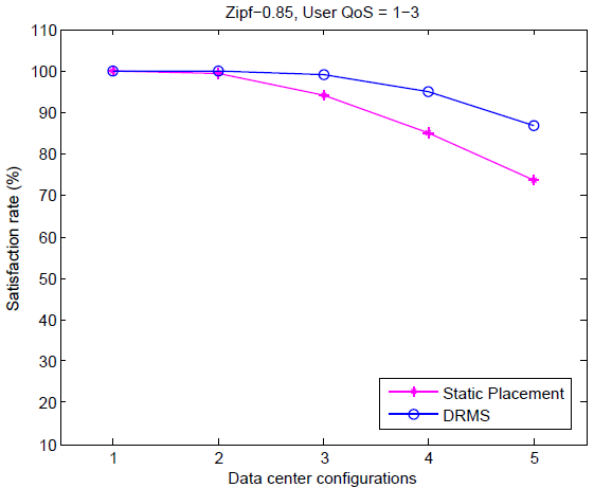


(b)

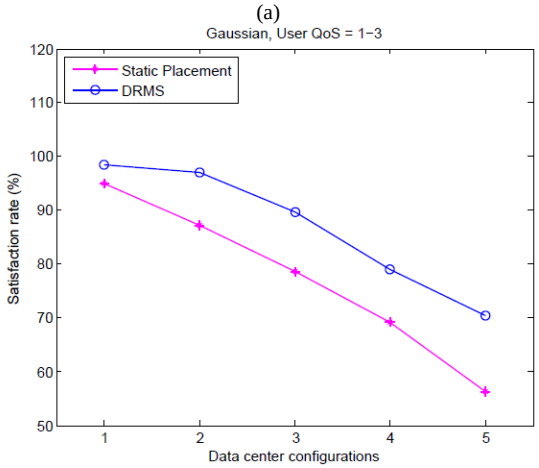
Fig. 7. Response times of static replica placement and DRMS considering more constrained QoS requirement [0-1] for Zipf-0.85 and Gaussian access patterns in (a) and (b)

For relatively more constrained QoS requirement ([0-1]) the performance improvement of dynamic maintenance model drops significantly for both Zipf and Gaussian access patterns as shown in Figure 7. In general, the performance benefit of DRMS over static placement becomes more obvious when user QoS requirements of wider ranges are considered.

Figure 8 demonstrates approximate values of user satisfaction rates for both the strategies using all storage configurations of data centers. DRMS performs better in most cases. Notably, the performance benefit of DRMS over static replica strategy is prominent when the storage capacity of data center servers becomes limited (for example in case of 17.5% and 13.75% relative capacities). However, user QoS satisfaction rates drop for both strategies in this case irrespective of the data access patterns and user QoS constraints used as shown in Figure 8.



(a)



(b)

Fig. 8. User satisfaction rates with relaxed QoS requirement [1-3] for Zipf-0.85 and Gaussian patterns in (a) and (b)

VI. Conclusions and Future Work

This paper investigates the dynamic replica maintenance problem in a multi-cloud scenario. To this end, a novel approach to distributed placement of static replicas in appropriate data center locations is proposed. Motivated by the fact that a multi-cloud environment is highly dynamic, the paper presents a dynamic replica maintenance technique that re-allocates replicas to new data center locations upon significant performance degradation. Performance analysis of the proposed techniques is done in terms of total response time and user satisfaction rates. The simulation results showed that the proposed dynamic maintenance technique, DRMS, can considerably reduce response times compared to the static

counterpart. In addition, user satisfaction rates are shown to be relatively higher due to dynamic replica maintenance. These benefits are attained using a wide range of storage configurations of data center servers and data access patterns with a degree of temporal locality and randomness. 501554-3-Distributed Systems.

In the future, we plan to implement the proposed dynamic replica maintenance algorithm in a real multi-cloud platform. Moreover, the algorithm will also be extended to deal with the peak bandwidth usage due to network link constraints and traffic patterns.

REFERENCES

[1] M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. H. Katz, A. Konwinski, G. Lee, D. A. Patterson, A. Rabkin, I. Stoica, and M. Zaharia, "Above the clouds: A berkeley view of cloud computing," Technical Report UCB/EECS-2009-28, Dept. of EECS, California Univ., Berkeley, Feb. 2009.

[2] J. M.D. Dikaiakos, D. Katsaros, P. Mehra, G. Pallis, and A. Vakali, "Cloud computing: Distributed Internet computing for IT and scientific research," IEEE Internet Computing, vol. 13, no. 5, pp. 10-13, Sept. 2009.

[3] H. Lamehamedi, B. Szymanski, Z. Shentu, and E. Deelman, "Data replication strategies in grid environments," in Proc. of the Fifth Intl. Conf. on Algorithms and Architectures for Parallel Processing, pp. 378–383, 2002.

[4] I.S. R. Buyya, C.S. Yeo, S. Venugopal, J. Broberg, and I. Brandic, "Cloud Computing and Emerging IT Platforms: Vision, Hype, and Reality for Delivering Computing as the fifth Utility," Future Generation Computer Systems, vol. 25, no. 6, pp. 599-616, June 2009.

[5] Amazon, Amazon simple storage service (Amazon S3). Available: http://aws.amazon.com/s3

[6] S. Ghemawat, H. Gobioff, and S. T. Leung, "The Google file system," SIGOPS Oper. Syst. Rev., vol. 37, no. 5, pp. 29-43, 2003.

[7] Human Genome Project, http://www.ornl.gov/hgmis/home.shtml.

[8] A. Chervenak, "Giggle: A framework for constructing scalable replica location services," IEEE Supercomputing, pp. 1–17, 2002.

[9] J. H. Abawajy and M. Deris, "Data replication approach with data consistency guarantee for data grid," IEEE Transactions on Computers, vol. 63, no. 12, pp. 2975 – 2987, 2014.

[10] K. Ranganathan, A. Iamnitchi, and I. Foster, "Improving data availability through dynamic model driven replication in large peer-to-peer communities," in Proc. of the 2nd IEEE/ACM Intl. Symposium on Cluster Computing and the Grid (CCGRID'02), pp. 376–381, 2002.

[11] J. Abawajy, "Placement of file replicas in data grid environments," in Proc. of the Intl. Conf. on Computational Science, vol. 3038, pp. 66–73, 2004.

[12] K. Kalpakis, K. Dasgupta, and O. Wolfson, "Optimal placement of replicas in trees with read, write, and storage costs," IEEE Transactions on Parallel and Distributed Systems, vol. 12, no. 1, pp. 628–637, 2001.

[13] F. Wang, J. Qiu, J. Yang, B. Dong, X. Li, and Y. Li, "Hadoop high availability through metadata replication," Proc. First Int'l Workshop Cloud Data Manage, pp. 37-44, 2009.

[14] K. Shvachko, H. Kuang, S. Radia, and R. Chansler, "The Hadoop distributed file system," Proc. IEEE 26th Symp. Mass Storage Systems and Technologies (MSST), pp. 1-10, June 2010.

[15] A. Gao and L. Diao, "Lazy update propagation for data replication in cloud computing," Proc. Fifth Int'l Conf. Pervasive Computing and Applications (ICPCA), pp. 250-254, Dec. 2010.

[16] W. Li, Y. Yang, J. Chen, and D. Yuan, "A cost-effective mechanism for cloud data reliability management based on proactive replica checking," Proc. IEEE/ACM 12th Int'l Symp. Cluster, Cloud and Grid Computing (CCGrid), pp. 564-571, May 2012.

[17] Q. Wei, B. Veeravalli, G. Bozhao, Z. Lingfang, and F. Dan, "CDRM: A cost-effective dynamic replication management scheme for cloud storage cluster," in 2010 IEEE International on Cluster Computing. pp. 188 – 196, 2010.

[18] S. Wang, K. Q. Yan, and S. C. Wang, "Achieving efficient agreement within a dual-failure cloud-computing environment," Expert Syst. Appl., vol. 38, no. 1, pp. 906-915, 2011.

[19] D. W. Sun, G. R. Chang, and S. Gao, "Modeling a dynamic data replication strategy to increase system availability in cloud computing environments," Journal of Computer Science and Technology, vol. 27, no.2, pp. 256-272, Mar. 2012.

[20] M. Hussein and M. Mousa, "A light-weight data replication for cloud data centers environment," International Journal of Innovative Research in Computer and Communication Engineering, vol. 2, no.1, pp. 2392-2400, Jan 2014.

[21] W. Li, Y. Yang, J. Chen, and D. Yuan, "A cost-effective mechanism for cloud data reliability management based on proactive replica checking," Proc. IEEE/ACM 12th Int'l Symp. Cluster, Cloud and Grid Computing (CCGrid), pp. 564-571, May 2012.

[22] A. Bessani, M. Correia, B. Quaresma, F. Andr'e, and P. Sousa, "DepSky: Dependable and secure storage in a cloud-of-clouds," in Proc. of the 6th Conference on Computer Systems, pp. 31-46, 2011.

[23] X. Wu, "Data sets replicas placements strategy from cost-effective view in the cloud," Scientific Programming, vol 2016, pp. 1-16, 2016.

[24] J. Lin, C. Chen, and J. M. Chang, "QoS-aware data replication for data-intensive applications in cloud computing systems," IEEE Transactions On Cloud Computing, vol. 1, no. 1, pp. 101-115, 2013.

[25] X. Fu, R. Wang, Y. Wang, and S. Deng, "A replica placement algorithm in mobile grid environments," Proc. Int'l Conf. on Embedded Software and Systems (ICESS '09), pp. 601-606, May 2009.

[26] A.M. Soosai, A. Abdullah, M. Othman, R. Latip, M.N. Sulaiman, and H. Ibrahim, "Dynamic replica replacement strategy in data grid," Proc. Eighth Int'l Conf. on Computing Technology and Information Management (ICCM), pp. 578-584, Apr. 2012.

[27] C. Cheng, J. Wu, and P. Liu, "Qos-aware, access-efficient, and storageefficient replica placement in grid environments," Journal of Supercomputing, vol. 49, no. 1, pp. 42–63, 2009.

[28] B. Liao, J. Yu, H. Sun, and M. Nian, "A QoS-aware dynamic data replica deletion strategy for distributed storage systems under cloud computing environments," in Int'l Conf.on Cloud and Green Computing (CGC), pp. 219-225, 2012.

[29] A. R. Varma and A. K. Shrivastava, "File replication and consistency maintenance in the Hadoop cluster using IRM technique," Int'l Journal of Advanced Research in Computer Engineering & Technology (IJARCET), vol. 3, no. 7, pp. 2424-2428, 2014.

[30] B. Chun, F. Dabek, A. Haeberlen, E. Sit, H. Weatherspoon, M. Kaashoek, J. Kubiatowicz, and R. Morris, "Efficient replica maintenance for distributed storage systems," in Proc. of the 3rd conf. on Networked Systems Design & Implementation, vol. 3, 2006.

[31] H. Wang, P. Liu, and J. Wu, "A QoS-aware heuristic algorithm for replica placement," Journal of Grid Computing, pp. 96–103, 2006.

Contribution to the Development of a Dynamic Circulation Map using the Multi-Agent Approach

Asmaa ROUDANE*, Mohamed YOUSSFI, Khalifa MANSOURI
Lab. SSDIA, ENSET, University Hassan II Casablanca, Morocco
Bd. HassanII, Mohammedia, 28820 Morocco

***Abstract*—Road traffic is considered one of the most difficult domains to manage and one of the fast- growing networks. This environment is geographically distributed and its actors are in continuous interaction in order to be able to succeed in a real-time exchange of the variable data of the road traffic. The modeling of this complex system requires powerful tools to be able to realize precise representations of reality hence the combination of mathematical modeling tools and intelligent systems. This paper presents a model of road traffic using graphs for the representation of the road infrastructure and the multi-agent technology for the management of the data representing the information circulating in the road network. This new approach aims at optimizing traffic paths, minimizing the time of a given journey and developing a dynamic traffic map that will enable the user to reach his destination by following the best possible path.**

Keywords*—Road traffic; road traffic management systems; intelligent systems; complex systems; multi-agent systems; graphs; real-time

I. INTRODUCTION

New technologies such as artificial intelligence, collective intelligence, multi-agents technology, expert systems and real-time systems are used to facilitate the management of complex environments and to improve their functioning **[1]**. The study, modeling, and processing of complex systems by methods of collective intelligence apply to various fields of application such as population flow modeling **[2]**, epidemiology, energy flow, social networks, and adaptive load distribution. Collective Intelligence systems are naturally complex and **[3]** have the ability to adapt to uncertain and unknown environments, **[4]** can organize themselves autonomously, and **[5]** present 'emergent' behavior. The complex environment in this contribution is road traffic.

The objectives of most road traffic studies are to improve the fluidity of traffic in the network by preventing and avoiding abnormal traffic events (saturation, congestion **[6]**, etc.), and minimizing the travel time of each user by enabling him to reach his destination under the best possible conditions **[7]**. The achievement of these objectives depends first and foremost on the infrastructure of the road network, which must be expanded to accommodate new users, who are increasingly numerous and who share the same network. Since the existing road infrastructure is insufficient and its enlargement is not always so obvious, another solution is to improve road traffic management methods by sharing the same network.

Improving road traffic management requires the use of new technologies, in particular, collective intelligence. As already mentioned, road traffic is a very complex, geographically distributed environment and its components are in continuous interaction, for all these reasons, and for others the multi-agent approach is the most suitable for better road traffic management.

An agent-based approach makes it possible to design non-centralized and self-organizing management models that improve the responsiveness of decision-making systems, which increases their autonomy. The term "agent" denotes hardware or software that supports an autonomous decision-making context. Agents are not simply objects or actors but also autonomous entities trying to achieve goals by acting on their environment. The agents must have the following properties: autonomy, proactivity, adaptability, sociability and mobility **[8]**. Multi-agent technology facilitates the management of geographically distributed environments, reduces their complexity and ensures continuous interaction between their different actors **[9].**

The combination of the multi-agent approach with the modeling tools gives powerful models that help to effectively solve many problems. One of the most widely used modeling tools is the graphs.

This contribution firstly presents the first proposed solution which was the result of a comparative study of existing solutions for the management of road traffic. This solution has limitations which will be detailed in the problematic part; these limitations led us to think of another solution by combining the modeling of the road network using the graphs and the management of the data circulating in this network using multi-agent technology. This modeling will make it possible to optimize traffic flows, to minimize the travel time of a journey, and finally to produce a dynamic traffic map that reflects the real-time state of the road network.

II. RELATED WORKS

The complexity of road traffic makes its management difficult and costly. Effective management depends on the availability of traffic data that are variable all the time. In general, the problems that a road traffic management system must solve can be summarized as follows:

- Identification of the critical areas of the system being studied.

- Detection of abnormal traffic events (saturation, congestion, accident ...).
- Rapid dissemination of information about abnormal events in order to keep a fluid state of traffic.

Several solutions have been proposed to meet the needs of road traffic managers. These solutions fall into two categories: tools to support real-time or deferred operations or decision-making tools. The tools for operating aid, or rather for regulation, are generally autonomous systems capable of performing an action in a given situation or under predefined conditions **[10]**. These systems make it possible to monitor the good progress of the operation and to monitor the quality of service offered to passengers. Decision support systems are solutions based on knowledge and collective intelligence; they collect information, analyze it, interpret it, symbolize it and finally present it to decision-makers in order to help them take the best decision for a given situation **[11]**.

Most of the systems studied are based on multi-agent technology as there are others based essentially on the object-oriented approach. The comparative study considers two factors to present the applications studied. These two factors are:

The Coordination in each system by using three attributes:

- Control: centralized or distributed.
- Structure: static or dynamic.
- Attitude: cooperative or competitive, or both.

The **Maturity** or level of experience, considering the following four levels of experience:

a) Architectural proposal: a description of the idea and the principles with their characteristics.

b) Simulation: the tool has been validated in a simulation environment with real or simulated data.

c) Real environment: The tool has been validated in a real environment with limited or complete data.

d) Deployment: This is the highest level where the application has been implemented in the real world and has been used.

The comparative study is presented in Table 1:

TABLE I. COMPARATIVE STUDY OF EXISTING ROAD TRAFFIC MANAGEMENT SOLUTIONS

System	Use	Structure	Coordination	Maturity	Type of data
TRYS (Work of Cuena et al) [11]	DSS	MAS dynamic	Distributed, cooperative	Simulation	Artificial, limited
InTRYS [12]	DSS	MAS static	Centralized, cooperative	Deployment	Real, limited
TRYS A2 [12]	DSS	MAS static	Distributed, cooperative	Simulation	Real, complete
Work of Adler et al [13]	DSS et OSS	MAS dynamic	Distributed, cooperative	Simulation	Artificial, complete
Work of Ossowski et al [14]	DSS	MAS	Distributed, cooperative	Simulation	Artificial, limited
Work of Li et al [15]	DSS	MAS static	Distributed, cooperative	Simulation	Artificial, limited
Work of Finder et Strap [16]	DSS	Distributed Control	Distributed	Proposition architectural	-
Work of Van Katwijk [17]	OSS	MAS	Distributed, cooperative	Simulation	Artificial, limited
TRANSYT [18]	OSS differed	OOP		Deployment	Real, limited
CLAIRE [19]	OSS real	OOP (C++)		Deployment	Real, limited
SCOOT [20]	OSS real	OOP		Deployment	Real, limited

***DSS**: Decision Support System **OSS**: Operations Support System **OOP**: Object Oriented Programming **MAS**: Multi-Agents System*

The solutions studied present some inputs in the management of road traffic but they are limited. These limitations are either in presenting time-delayed functionality, which is not efficient in the management of a variable environment, or in covering only one problem of road traffic. This comparative study has been an inspiration to design a first model of road traffic management based on multi-agent technology and the sharing of the studied area in sub-zones monitored by agents in continuous communication.

First model:

The idea of the model is to apply the two approaches: Distributed Artificial Intelligence (AI) and Multi-Agent approach. The AI approach is an approach for the control of large scale systems using decomposition and distribution. Large systems are decomposed into a series of small, interconnected subsystems, each responsible for controlling its domain and coordinating activities with adjacent subsystems. The AI approach creates a more robust environment in control side by allowing faster responses; sharing critical resources and increasing flexibility in adapting to changes in the system **[21]**.

Each domain is monitored by a reactive agent whose mission is to collect road traffic data using equipment such as detectors or sensors and then transfer this data to a coordinating agent so that it can be sent to its turn to the area agent to be finally communicated to the central agent, as described in Figure 1. All these communications must be done in real time since the data collected are variable and unpredictable.

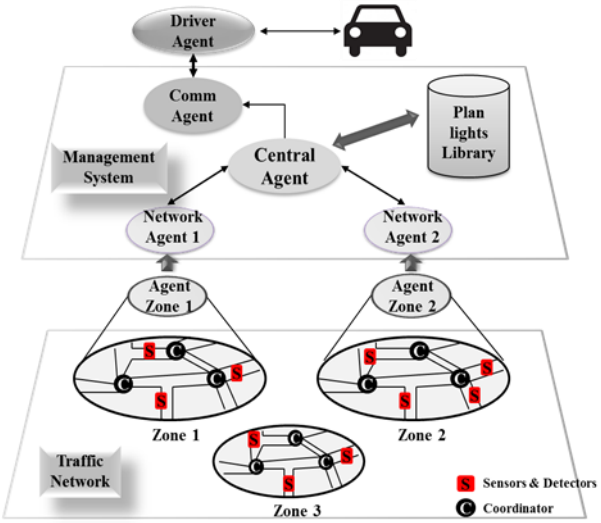


Fig. 1. First proposed model for road traffic management

The implementation of this model has shown that it is difficult to develop and succeed all communications between agents in real-time. In general, the modeling of complex systems, such as road traffic, by reactive agents uses a sub-symbolic representation of the world through stimulus/action structures, that is, following a hierarchy of many simple automaton agents who need information on their local environment in order to decide on their actions **[22]**. This hierarchical architecture that has been adopted for agents is too complex to construct. To understand this difficulty, a communication diagram has been elaborated, in Figure 2, between agents.

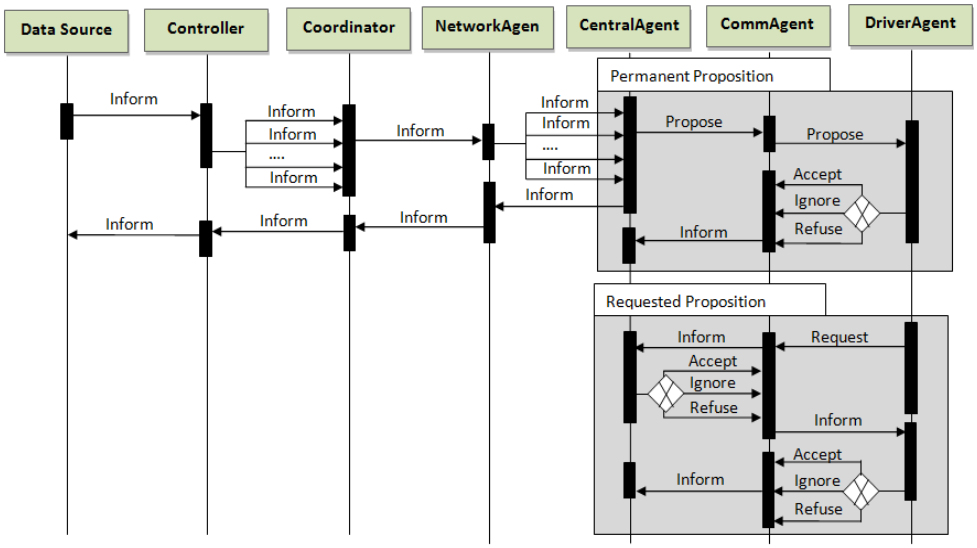


Fig. 2. Diagram of communications between the actors of the first proposed model for the management of road traffic

The analysis of the communications between the agents shows that there are several messages to be exchanged in real time, these messages contain variable data which must be communicated to the central agent before they change, which is not Always the case given the unpredictability of traffic data.

This analysis and other research have oriented the researches towards another idea in order to better manage road traffic. The new solution proposed is to separate the modeling of the infrastructure and the management of the data using one of the most powerful modeling tools: Graphs.

III. GRAPHS AND MULTI-AGENT SYSTEMS

The field of application in this study is road traffic; this system depends on two components:

- Road infrastructure.
- Data circulating in the road network.

The proposed solution for road traffic management consists of modeling the road infrastructure using graphs and developing a traffic data tree that will be dynamically fed by agents using the multi-agent approach. The use of graphs is commonly current as a representation tool: the schematic map of streets, a genealogical tree, and the representation of a computer network; are examples of modeling using graphs. Graphs are standard objects of mathematical modeling; they have an expressive power to cover very wild fields of problems and applications **[23]**.

A graph can be represented by several means, either by the description of the arcs, representing it by a list of summits, each being characterized by its name, a possible label, and the list of arcs that have this summit for origin. The representation can be sagittal that means in a drawing form, it can also be a matrix (adjacency matrix), or by a dictionary which consists of a table enumerating the following and preceding of each summit.

To represent a network or a complex system by a graph, it is first necessary to establish an analogy between the two. A network is a graph whose individuals and relationships have quantitative or qualitative characteristics. As for the complex system, it is a large number of interacting entities forming a set whose behavior is unpredictable for the observer because resulting from non-trivial interactions **[24]**.

Then the entities of a complex system are associated to the summits of the graph, and the interactions between these entities are associated in turn with the graph arcs. However, graphs are discrete structures whose structure, topology is static, while complex systems are inherently dynamic and evolve over time, hence the separation of the infrastructure modeling, which is in general static, and data that are variable all the time and unpredictable.

Modeling of road infrastructure:

The complex system in the actual study is road traffic; Figure 3 presents a first modeling of the road network using the graphs:

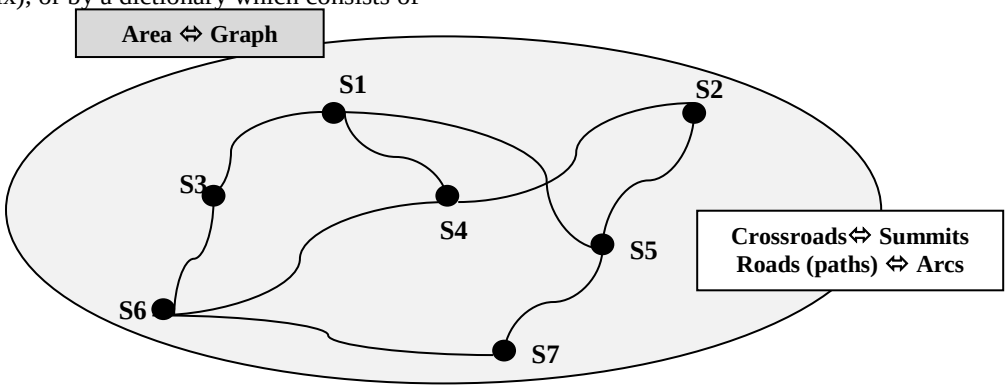


Fig. 3. Modeling the road infrastructure using graphs

Each area of the road network is considered as a graph whose summits are intersections or intersections, and the arcs are the roads or paths connecting each two crossroads. A zone is identified by a name, it contains several crossroads. A crossroads is identified by a name and its coordinates. A road connecting two crossroads is identified by a name, an origin and a destination, and the most important by a weight. In general; the weight of an arc is its value, in this case, the weight is the dynamic part of the data circulated in the network.

In this proposition the value of a road depends on several parameters, these parameters are either static or variable. Variable parameters are only the dynamic traffic data that is collected by the agents monitoring the road in question. Each crossroads has an agent who collects data concerning the roads connected to the crossroads in question, these data are issued by the mobile agents associated with the vehicles crossing the intersection. Then the agents of the crossroads are required to regularly transfer the collected information to the system agent whose mission is the updating of the dynamic traffic map which offers users the current state of the road network, as schematized in figure 4.

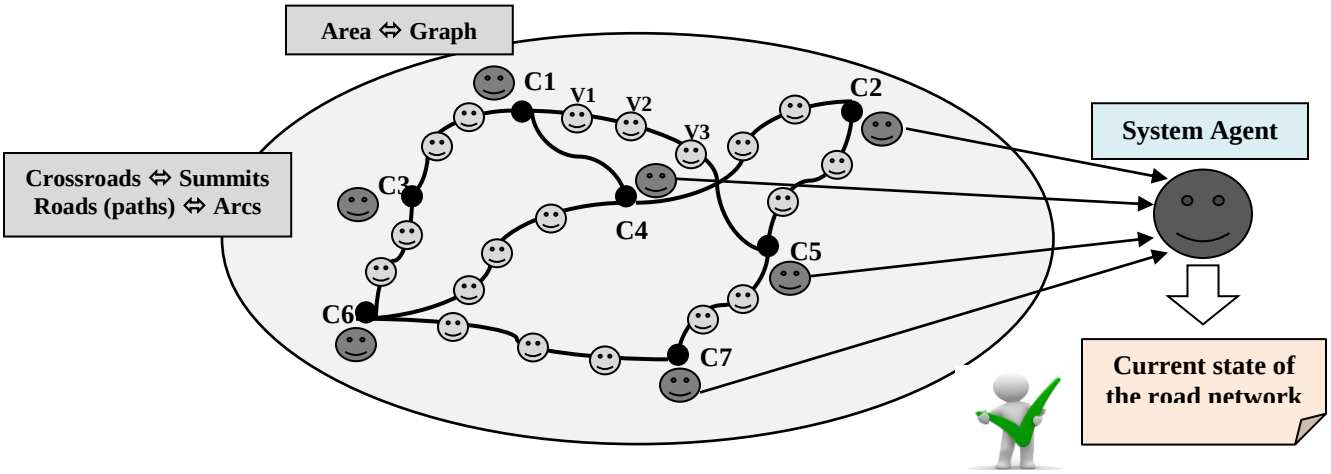


Fig. 4. Modeling of road traffic using graphs and the multi-agent approach

Each vehicle agent passing through a road is invited to send its current information to the agents of the nearby intersections. The latter transmits the received information to the system agent which analyzes the received messages in order to feed the traffic data tree to be able to update the dynamic traffic map to keep a real view of the road network. The interactions between the agents described above are schematized in the diagram presented in Figure 5.

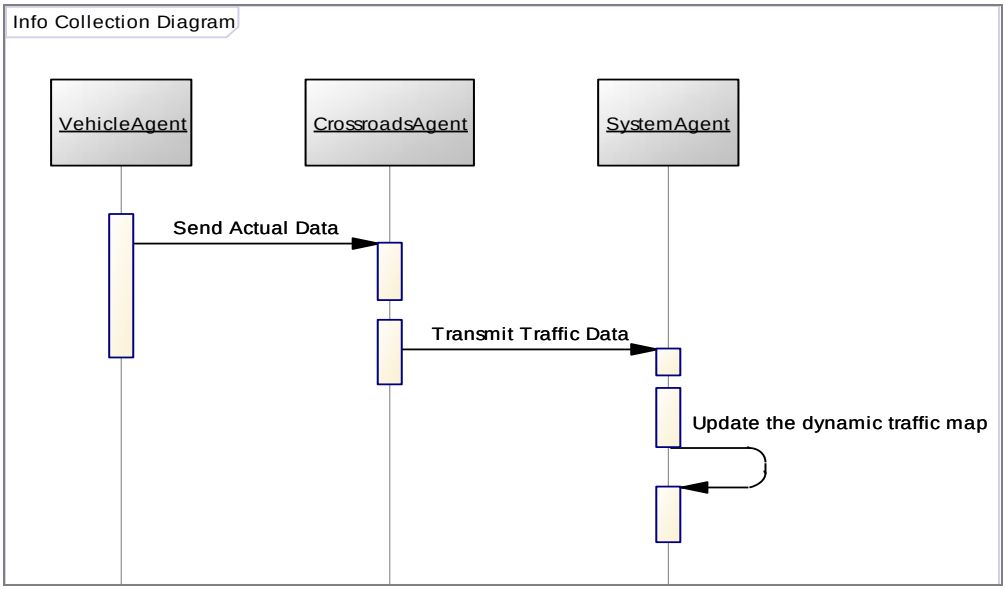


Fig. 5. Sequence diagram detailing road traffic data collection

For a user to reach his destination by browsing the best possible path he must connect to the system and indicate his destination, the system analyzes the information sent by the user and searches for all possible paths to the indicated destination. Thereafter, the system agent uses the optimization algorithms to choose the best possible path that is sent to the user afterward. All these exchanges are detailed in Figure 6.

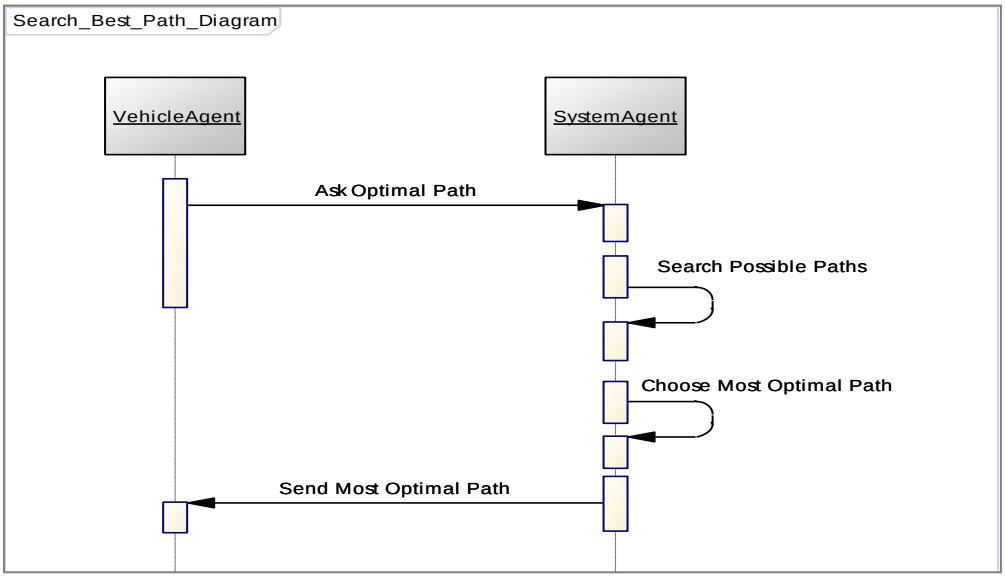


Fig. 6. Sequence diagram detailing the optimal path search

The tree of the data circulating in the road network:

As mentioned in the previous section, each road in the studied area is monitored by an agent. The mission of this agent is summarized in the collection of the variable data of the path in question and the transfer of this information to the system agent whose mission is to update the data tree described in Figure 7. The variable data (speed, actual distance ...) make it possible to evaluate each road in order to offer the user the best possible path to his destination.

The best possible path depends on several factors: it may be the shortest road (short distance), it can be the most secure road, or it can be a path that is distinguished by several characteristics. In the description of the data tree, a few parameters are presented which characterize each road; these parameters make it possible to improve the calculation of the value of each path as may be added to other information that may offer the user a better view of the best route to his destination. The parameters to be included in the assessment of the route must be well analyzed in order to give us an accurate classification of the paths to be covered.

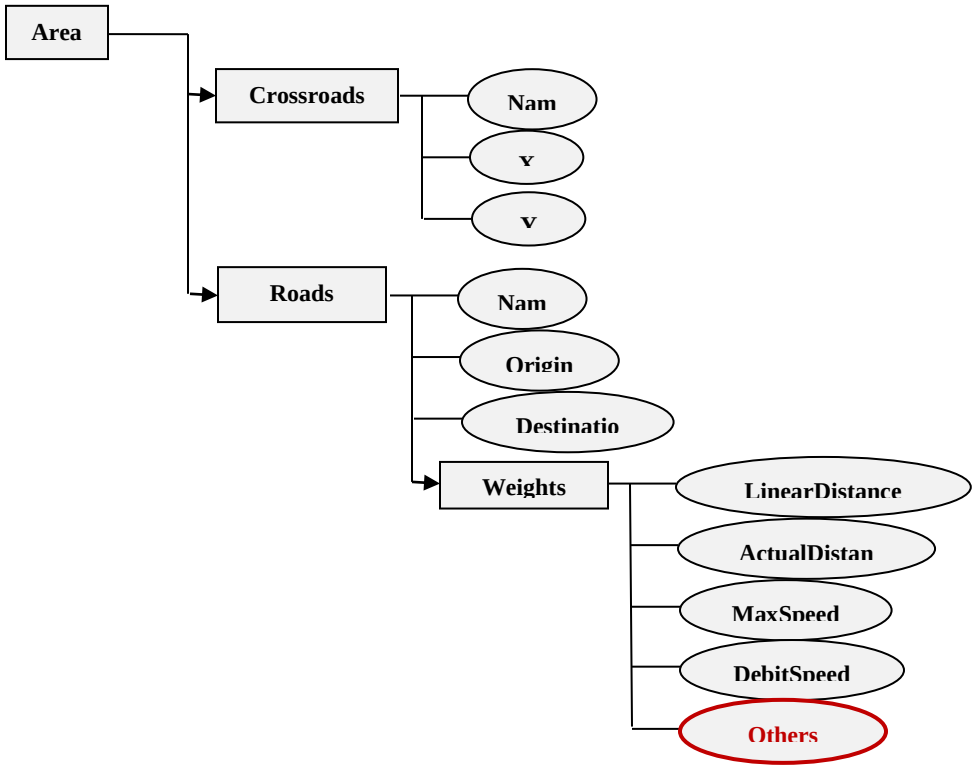


Fig. 7. The road traffic data tree

The solution presented in this article has several objectives: one of them is the minimization of the journey time of a path. The travel time is an equation with two variables: the speed of flow and the actual distance, in other words, it is the relation between the distance actually traveled and the average speed with which this distance has been browsed. In this solution, the travel time is a function with several parameters that change in real time, and these parameters can be changed according to the preferences of the user.

The system agent in the proposed solution is an agent that has several missions:

- Update information in the road traffic data tree.
- Find the most optimal path to a specified destination using the optimization algorithms that will allow choosing the best route according to the state of the network and according to the user preferences.

Thus, by connecting to the proposed system, the user can have a real-time view of the state of the road network thanks to the dynamic traffic map offered by the system; in addition, he can indicate his preferences to be able to personalize the calculation of his travel time for the best possible path to his destination.

IV. CONCLUSION

This paper presents a new modeling of road traffic based on graphs and multi-agent technology. This modeling consists in separating the representation of the road infrastructure using the specifications of the graphs and the representation of the data circulating in the road network using a data tree that will be managed in real time by agents based on the technology Multi-agents.

The new approach in this paper aims to optimize traffic routes, minimize travel time for a given destination, and develop a dynamic traffic map that will provide a real-time view of the current state of the road network to users; enabling them to reach their destinations by traveling the best path possible.

REFERENCES

[1] M. Mitchell, "Complex systems: Network thinking", Artif. Intell, vol. 170, no. 18, pp. 1194-1212, Dec. 2006.

[2] James M. Keller, Derong Liu, David B. Fogel, “Collective Intelligence and Other Extensions of Evolutionary Computation”, Wiley-IEEE Press 1, pp. 400, 2016.

[3] P. Ball, Material witness: Designing with complexity, Nature Materials 3 (2004) 78.

[4] A.-L. Barabási, Linked: The New Science of Networks, Perseus, New York, 2002.

[5] A.-L. Barabási, R. Albert, Emergence of scaling in random networks, Science 286 (2002) 509–512.

[6] Shen Wang, Soufiene Djahel, Zonghua Zhang, Jennifer McManis, “Next Road Rerouting: A Multiagent System for Mitigating Unexpected Urban Traffic Congestion”, IEEE Transactions on Intelligent Transportation Systems, Vol. 17(10), pp. 2888 - 2899, 2016.

[7] B.-Y. Chen, W. H. K. Lam, Q. Li, A. Sumalee, K. Yan, "Shortest path finding problem in stochastic time-dependent road networks with stochastic first-in-first-out property", IEEE Trans. Intell. Transp. Syst., vol. 14, no. 4, pp. 1907-1917, Dec. 2013.

[8] M. Wooldridge, M. Jennings, « Intelligent agents: theory and practice”. The Knowledge Engineering review 10(2), pp.115-152, 2002.

[9] B. Burmeister, A. Haddadi, and al, “Application of multi-agent in traffic and transportation”. IEEE Proc. Software Engineering 144(1), pp. 51-60, 1997.

[10] Mc Graw-Hill, Science and technology, 2003.

[11] J. Cuena, J. Hernandez, and al, “Knowledge based models for adaptative traffic management systems”, Transportation Research 3(5), pp. 311-337, 1995.

[12] J. Hernandez, S. Ossowski, A. Garcia-Serrano, “Multiagent architectures for intelligent traffic management systems”, Transportation Research Part C: Emerging Technologies, vol. 10, pp. 473-506, 2002.

[13] J. L. Adler, and V. J. Blue, “Cooperative multi-agent transportation management and route guidance systems”, Transportation Research Part C, vol. 10, pp. 433-454, 2002.

[14] S. Ossowski, J. Z. Hernandez, and al, “Decision support for traffic management based on organizational and communicative multiagent abstraction”. Transportation Research Part C 13, pp. 272-298, 2005.

[15] M. Li, J. Hallam, and al, “A cooperative intelligent system for urban traffic problems”. Proceedings of the 1996 IEEE International Symposium on intelligent Control, Dearborn, MI, pp.162-167, 1996.

[16] N. V. Finder, J. Strap, “A distributed approach to optimized control of street traffic signals”. Journal of Transportation Engineering 118, pp. 99-110, 1992.

[17] R. van Katwijk, P. van Koningsbruggen, “Coordination of traffic management instruments using agent technology”. Transportation Research Part C: Emerging Technologies, vol. 10, Issues 5-6, pp. 455-471, 2002.

[18] D. I. Robertson, “Research on the TRANSYT and SCOOT methods of signal coordination”. ITE Journal, 1986.

[19] G. Scemama, “Développement d’un système à base de connaissance historique pour la gestion du trafic”. Recherche Transport Sécurisé, R.T.S. 5(22), 1992.

[20] P. B. Hunt, D. L. Robertson, R. D. Bretherton, M. C. Royle, “The SCOOT online traffic signal optimization technique”. Traffic Engineering and Control 23, pp. 190-199, 1982.

[21] Drogoul, A. Callinot, “Applying an Agent-Oriented Methodology to the Design of Artificial Organizations: A case study in Robotic Soccer”. Journal of Autonomous Agents and multi-agent systems 1(1), pp. 113-129, 1998.

[22] R. L. Brooks, “On coloring the nodes of a network”, Mathematical Proceedings of the Cambridge philosophical society, vol. 37, Issue 2, pp. 194-197, 1941.

[23] M. Gondran, M. Minoux, “Graphes et algorithms”. 1995.

[24] G. Abu-Lebdeh, A. K. Singh, “Modeling Arterial Travel Time with Limited Traffic Variables using Conditional Independence Graphs and State-Space Neural Networks”. Procedia-Social and Behavioral Sciences, vol. 16, pp. 207-217, 2011.

Downlink and Uplink Message Size Impact on Round Trip Time Metric in Multi-Hop Wireless Mesh Networks

Youssra Chatei
Dept. Electronics, Informatics and Telecommunications
ENSAO, Mohammed I University
Oujda, Morocco

El Miloud Ar-reyouchi
Dept. Telecommunication and Computer Science
Abdelmalek Essaadi University
Tetouan, Morocco

Maria Hammouti
Dept. Electronics, Informatics and Telecommunications
ENSAO, Mohammed I University
Oujda, Morocco

Kamal Ghoumid
Dept. Electronics, Informatics and Telecommunications
ENSAO, Mohammed I University
Oujda, Morocco

***Abstract*—In this paper, the authors propose a novel real-time study metrics of Round Trip Time (RTT) for Multi-Hop Wireless Mesh Networks. They focus on real operational wireless networks with fixed nodes, such as industrial wireless networks. The main aim of the metric is to show the effect of the Downlink and Uplink message size (DMS and UMS) on RTT with and without Forward Error Correction (FEC), user data size without any headers, on RTT path between a source (Supervisory Control and Data Acquisition (SCADA) center) and a destination RTU (Remote Terminal Units). The metric assigns weights to links based on the RTT path of a packet size over the link path. They studied the performance of the metric by implementing it in three wireless scenarios consisting of 3, 4 and 5 nodes; each node represents a wireless radio IP router. They find that in a multi-hops environment, the real-time metric clearly shows the impact of DMS, UMS and FEC on RTT by making judicious use of the number of the hops.**

Keywords*—RTT; Remote wireless communications; Wireless radio router; Wireless multi-hop networks; Average message size; FEC; SCADA

I. INTRODUCTION

In this work, the authors consider the RTT as the time required for a packet to travel from the source (SCADA center) to the destination (remote RTU) and back again.

The SCADA center allows technical operation management for PV power plants on site. It permits us to visualize all measured values locally, in real time and in the event of a fault, reacts quickly and efficiently.

A remote terminal unit (RTU) is a multipurpose device used for remote monitoring and control of various devices and systems for automation.

In wireless network, the RTT is considered as an important measure in determining the accomplishment of a connection. RTT is effective and profitable in measuring the congestion window size and retransmission timeout of a connection [1] [2] [3] as well as the available bandwidth on an along path [4]. This information can help fixed, determine factors that limit data flow rates and cause congestion [5]. When a network link along the path is experienced, RTT can also aid efficient queue management and buffer provisioning. Moreover, RTT can be used to ameliorate node distribution in peer-to-peer and coverage networks [6].

With the wide range of RTU and PLC (Programmable Logic Controllers) currently on the market, SCADA system engineers [7] and decision makers face several challenges [8].

In basic SCADA architectures [9], information from a set of nodes (node-set) in wireless sensor networks (WSNs) [10], sensors or manual inputs are sent to PLCs or RTUs, which then send that information to computers with SCADA software. SCADA software analyzes and displays the data in order to help operators and other workers to reduce waste and improve efficiency in the manufacturing process.

The advent of Wireless Mesh Networking technology has introduced some very effective solutions for applications that previously required impractical and very expensive infrastructure. A Multi-Hop Wireless Mesh Networks [11] ,[12] can deliver large amounts of data and causes significant End-to-End delays across long distances using lot of nodes [13].

Wireless mesh networks also provide the added benefit of link redundancy for continued reliability across the infrastructure. Mesh networking helps to improve on the efficiency that Progressive Communications already delivers with custom SCADA and Digital Monitoring systems.

With Wireless Mesh Networks and SCADA monitoring, Progressive Communications can deliver solutions for applications such as:

1) Point-to-Multipoint wide area video surveillance
2) The peer-to-peer (P2P) renewable energy
3) Multipoint-to-Multipoint wide area video surveillance (law enforcement monitoring surveillance from their vehicles)

4) Public Wi-Fi access systems for municipal parks and commons

5) Data gathering for wireless Electric / Water /Gas meters

6) Remote monitoring of analog systems (SCADA)

7) Remote monitoring of sub-stations (such as city-wide water pumps or another municipal infrastructure)

8) Multi-purpose wireless backbones (multiple systems utilizing the same available network).

FEC is a coding technology widely used in communication systems[14]. The principle is that the sender inserts some redundant data into its messages. This redundancy allows the receiver to detect and correct errors.The traditional FEC technique can increase the RTT efficiency in single-hop wireless networks such as WLAN or WiMAX networks. Using the FEC approach, the source encodes additional information together with the data before broadcast them to the receivers. If the amount of lost data is sufficiently small, a receiver can recover the lost data using some decoding schemes.

The industrial narrowband land mobile radio devices, as treated in this paper, have been the subject of European standard ETSI EN 300 113 [15] ,[16]. The system functions on frequencies between 30 MHz and 1 GHz, with channel separations of up to 50 kHz, and is designed for private, fixed, or mobile, radio packet switching networks.

The rest of the paper is organized as follows. The system model is introduced in Section II. The experimental setup is given in Section III. The description of the performance network and the metric results are presented in section IV. Finally, the authors conclude the paper and announce the future work in Section V.

II. System Model

Let us consider the wireless network having n nodes as shown in Fig.1.

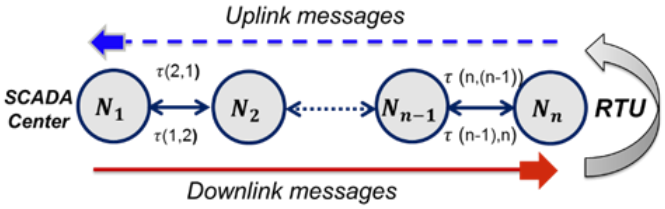


Fig. 1. Downlink and Uplink broadcast paths for each node

where, n is the number of nodes presented in the respective RTT path, $\tau(1,2)$ is the time between the node 1 & 2 and depends upon the distance among them. The $\tau(i, j)$ is the time among the node i and j respectively.

Each downlink message is sent by the SCADA center network server to only one end device (RTU) and is relayed by a radio router. While end devices (RTUS) send Uplink messages to the network server relayed by one or many radio routers. Downlink and Uplink messages use the radio packet explicit mode in which any headers are included.

Consider the condition of having *n* nodes in RTT path wich will be the maximum nodes in this Wireless Network. A worst-case application is constructed by the following path,

$$RTT\ path = Downlink\ RTT\ path + Uplink\ RTT\ path \quad (1)$$

So, the generalized equation for RTT time for the RTT path containing n nodes will be written as follows:

$$RTTpath = (N_1 \square\ N_2 ..... \square\ N_{n-1} \square\ N_n) \quad (2)$$

Where, the symboles $\square$ *or* $\square$ means the direction of the motion path of time that can takes a message to go from center to RTU and return to the center, through all specified intermediate nodes. $RTTpath(N_i \square\ N_j \square\ N_l)$ is the RTT among the node, i , j and l respectively, $(i, j, l \in \{1.2....n\})$.

The minimum number of routers required to form RTT path is three [1] and maximum routers in RTT paths should not be more than (n-1). The δ routers that build the RTT path for Wireless network are limited as following:

$$3 \leq \delta \leq n\text{-}1 \quad (3)$$

In router mode, the node works as a standard IP Router with two interfaces (Radio and Ethernet) and two COM port devices without any compromise. There is a sophisticated anti-collision protocol on radio channel, where every single packet is acknowledged. Moreover, each unit can simultaneously work as a store-and-forward repeater.

Packet size is also influenced, to a modest degree, by the transmission error rate. If a link is noisy, it is possible to add a FEC that allows the receiver in many cases to figure out which bits are corrupted, and fix them. This has the effect of improving the bit error rate at a cost of reducing throughput. In order to involve many more bits than are needed for error detection, additionally, FEC can be applied. Typically, if a communications technology proves to have an unacceptably high bit-error rate (such as wireless, case of the experimental measurements),

If propagation delay of the message is considered negligible , the transmission time or transfer time of a message is expressed as:

$$T_t = L/pD \quad (4)$$

Where, L is the message length (bits) is divided into p packets transmitted on the different carriers at the same rate of D (bit /s).

If only one packet is transmitted in a time L becomes;

$$L = m.8 \quad (5)$$

Where m is the user data (bytes) transmitted without any headers (IP, TCP, UDP, … (consider the longest possible reply from RT,U),

To minimize radio channel impairments, it is essential to use the FEC technique which is a very effective method. Therefore , the next step is to introduce an error-correcting

code to the different next senarions. This generally reduces the "virtual" bit-error rate (that is, the error rate as corrected) to acceptable levels. The improvement comes at the expense of the data rate throughput:

$$D = b \times FEC \quad (6)$$

Where, b is the modulation rate and FEC is the Forward Error Correction.

It apparent from this relation that the lower the FEC ratio, the better the error correction capability and correspondingly the lower the data rate. This implies an increase in the packet transmission time. Indeed, the packet transmission time through radio channel in miliseconds can be obtained from the packet size in bit (message= one packet) and the data rate in bit/s as:

Testing Connection using TCP (Transmission Control Protocol) [17] and it is noted that initial header options were with 20 bytes , but after the first ACK the header options were with 12 bytes, so, it is necessary to add12 bytes' in every TCP packet. (TCP packets include a 12 bytes' header).

Packet transmission time t_p can be calculated as follows:

$$t_p = (m + 12).8/(b\,.FEC) \quad (7)$$

Where, t_p (ms) is the time needed for the packet transmission, with:

FEC = 1.00 if FEC = Off

FEC =0.75 if FEC = On

It is noted that a Modbus RTU over TCP is used. Simply put, this is a Modbus RTU message transmitted with a TCP/IP wrapper and sent over a network instead of serial lines. The Server does not have a SlaveID since it uses an IP Address instead.

Fig.2 shows different RTT path scenarios of downlink transmission message for SCADA center and for uplink transmission message at remote RTUs. The node N_1 can be considered as the base station (BS).

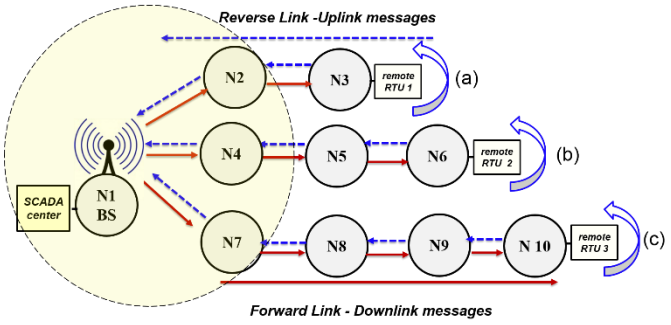


Fig. 2. Different RTT path scenarios of downlink and Uplink transmission message

In the topologies of the three scenarios, let us consider a graph including three, four and five nodes respectively. The RTT paths of network are calculated by using following equations:

1) Topology of the first scenario: Fig.2 (a)

$$T_{RTT_1} = \tau(1,2) + \tau(2,3) + \tau(3,2) + \tau(2,1) \quad (8)$$

2) Topology of the second scenario: Fig. 2 (b)

$$T_{RTT_2} = \tau(1,4) + \tau(4,5) + \tau(5,6) + \tau(6,5) + \tau(5,4) + \tau(4,1) \quad (9)$$

3) Topology of the third scenario: Fig. 2 (c)

$$T_{RTT_2} = \tau(1,7) + \tau(7,8) + \tau(8,9) + \tau(9,10) + \tau(10,9) + \tau(9,8) + \tau(8,7) + \tau(7,1) \quad (10)$$

The center broadcasts request and the RTUs 1, 2 and 3 generate the response and send it out to their respective nodes.

In this paper, through these metrics application, the frequency band (160 MHz) is the best choice when covering a hilly region and repeaters are not an option. The only frequency of the set of options, which can possibly make it to a distant valley, 20 km, Line of sight (LoS) as propagation mode, from the nearest point-of-presence, it can reach a ship 100 km from the shore base using radio router as the transmission medium with Simple Network Management Protocol (SNMP). Consequently, this frequency band is suitable for low speeds using robust modulation techniques only, and even then, somewhat lower long-term communication reliability has to be acceptable for this metric application.

III. EXPERIMENTAL SETUP

The system consists of two parts, related hardware and management software. The system hardware is divided into on-site data monitoring unit, wireless transmission unit and monitoring center. The software adopts decentralized collection and centralized control management model, providing users basic information management, real-time monitoring, fault alarm, fault records, maintenance reminder, maintenance records, failure statistics, bulletin boards, industry news and so on.

The equipment's, including computer software used, are:

- The SunSet E20c provides full transmission testing over 2.048 Mbit/s and V-series Datacom interfaces. It verifies Datacom circuits by monitoring the received data, control leads, and physical layer results. It also tests frame relay circuits, PING testing, stress testing, and statistics.
- Ten radio routers (use the frequency band 160 MHz) which are characterized by the SNMP (Simple Network Management Protocol) that will support the base MIB (Management Information Base).
- Omnidirectional antenna KA160.3 which is designed for base radio stations working in bands of 158-174 MHz The antenna, used in our application, has an omnidirectional radiation pattern with the gain of 3 dB and is adapted for the top-mounting. The values received at the level of each site vary between 38 and 70 dBµV.
- The output power of each radio router varies between 0.1 and 2 watts.

The Processing time is the time for the RTUs / SCADA devices to process queries; it has the same value and Interface speed, in either a center or remote, which is 20 msec and Ethernet-TCP/IP respectively.

The RTT measurement is affected by various parameters of the wireless network. The following Table I lists conditions selected.

TABLE. I. CONDITIONS SELECTED

Factors Affecting RTT measurement	Condition Chosen
Total node Number of sites	10
Average hops per path to remote	4-6-8
Operating Mode	Radio Router
Modulation rate [kbps]	16DEQAM (166.67)
ACK: Acknowledgement	On
FEC: Forward Error Correction	with (FEC= 3/4) / without
Processing time	20 ms

IV. RESULTS

This section describes the experiments results (see Fig. 2). The presented measurements show , firstly, how RTT (between source and destination) varies depending on the DMS (User data size without any headers) for several hops and UMS (500 Bytes, 1000 Bytes and 1500 Bytes) of wireless links, then their performances , of a variety of multi-hop wireless mesh networks to fixed value of UMS metric, are compared in various conditions.

The authors begin by comparing the performance of RTT to DMS as well as basic shortest-path routing using only three nodes per RTT path, 3 nodes and 4 Hops, (see Fig. 2):

$$RTT(N_1 \quad N_2 \quad N_3) \tag{11}$$

Then for four nodes per RTT path (4 nodes and 6 Hops):

$$RTT(N_1 \quad N_4 \quad N_5 \quad N_6) \tag{12}$$

Finally, the RTT path with five nodes (5 nodes and 8 Hops) is:

$$RTT(N_1 \quad N_7 \quad N_8 \quad N_9 \quad N_{10}) \tag{13}$$

Next, measurements results are obtained by selecting UMS= 500, 1000 and 1500 Bytes and the performance of RTT in this tree path routing is compared.

Various measurements are effectuated, in comparison to different values of packet length. Each router units' module may support up to 1500 bytes (User data size without any headers (IP, TCP, UDP …) of RF payload.

A. RTT for multi-hop wireless path

Fig.3 and Fig.4 show the RTT versus number of hops depending of different DMS/UMS and with or without FEC.

The results of the ping RTT, for multihop connections with all wireless hops (at the four, six, and eight hops) and with /without FEC, are showed in Fig.3 and Fig.4.

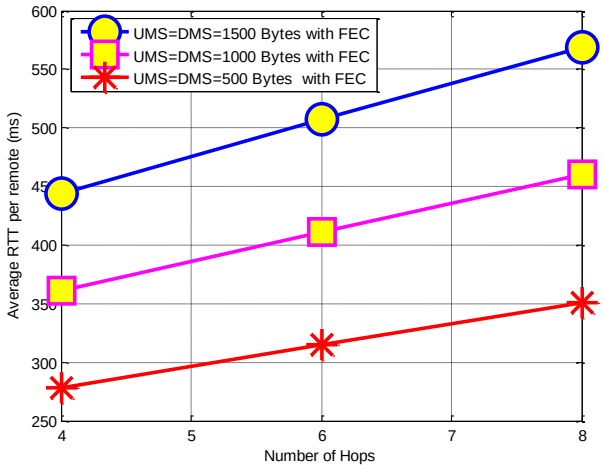


Fig. 3. The ping RTT for multihop connections with all wireless hops (at the four, six, and eight hops) and with FEC

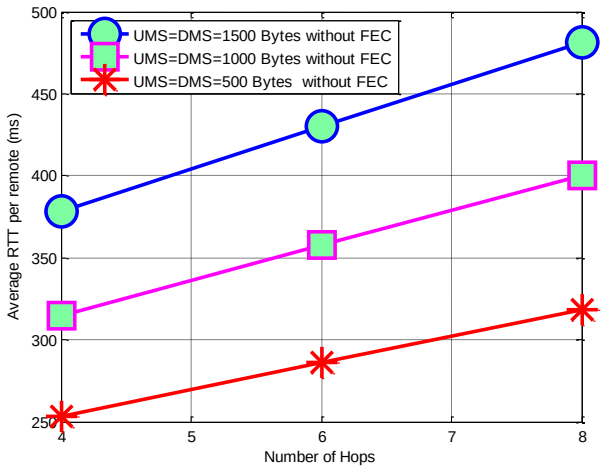


Fig. 4. The ping RTT for multihop connections with all wireless hops (at the four, six, and eight hops) and without FEC

As it can be seen in Fig.3 and Fig.4, both the number of hops and UMS/DMS can affect the RTT. In the two cases with / without FEC, the RTT increases when the number of hops or the UMS/DMS increases. Another important point obtained from these results is that the RTT measurements, by fixing the value of UMS=DMS at 1500,1000 and 500 bytes, increase of the same value from 4 hops to 6 hops and from 6 to 8 hops as illustrated in Table II.

TABLE. II. INCREASE OF RTT VALUES

	With FEC			Without FEC		
UMS=DMS	**500**	**1000**	**1500**	**500**	**1000**	**1500**
4 6 8	37	49	62	32	42	52

In order to avoid reception overcharge with the SCADA communication protocol, a packet is a block of data with length

that can vary between 0 to 1500 bytes. In this case, the shape of RTT curve is showed as a function of message size and the results are illustrated in Figs.5-10, for both situations with and without FEC, as described in the following three scenarios.

B. Average RTT per remote: First scenario

The first scenario is considered as shown in Fig. 2(a). It is worth mentioning that the direct transmission is considered between the SCADA center and the destination RTU1 using three nodes (radio router). Each node in Wireless Network is defined for a selected RTT path by configuring them, with SCADA center and RTU1, using the protocol SCADA software. The messages are relayed and RTT path selected here, as in (11), is:

$$RTT(N_1 \quad N_2 \quad N_3)$$

Fig.5 and Fig.6 show the results of a representative first scenario practice with and without FEC respectively.

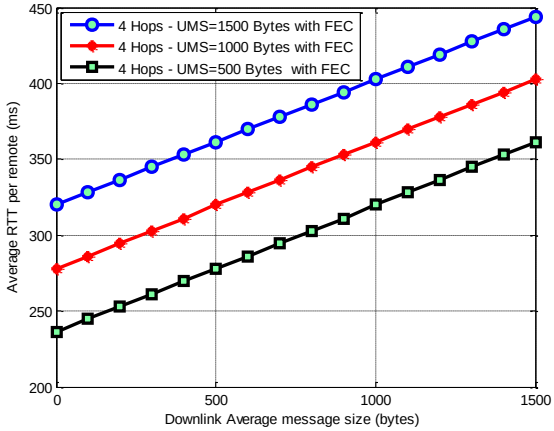


Fig. 5. Average RTT vs DMS: 4 Hops with FEC and different values of UMS

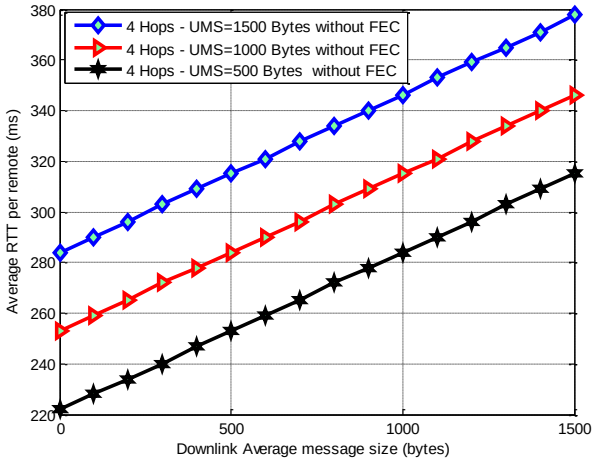


Fig. 6. Average RTT vs DMS: 4 Hops without FEC and different values of UMS

This experimental trial is repeated ten times. Furthermore, it was conducted to have a performance baseline for wireless networks.

The results in Fig.5 and Fig.6 shows that the RTT is increased when the DMS increases. The increase of UMS can also promote a slight increase in RTT. On the other hand, one can see from Fig.6 that the introduction of FEC can significantly increase the RTT.

For a fixed value of UMS= 1000 Bytes and ranging DMS from 0 up to 1500 Bytes and without FEC, the RTT for 4 hops increases to 93 msec, whereas for FEC=3/4 RTT increases to 125 msec.

C. Average RTT per remote: Second scenario

In Fig.7 and Fig.8, the topology showed in Fig.2 (b) is well respected. SCADA center unit, configured as a source, sends packets to destination RTU2 and comes back to (BS) then calculates the RTT. The direct transmission between the SCADA center and the destination RTU 2 is respected using four nodes (Router). The RTT path, as in (12), is:

$$RTT(N_1 \quad N_4 \quad N_5 \quad N_6)$$

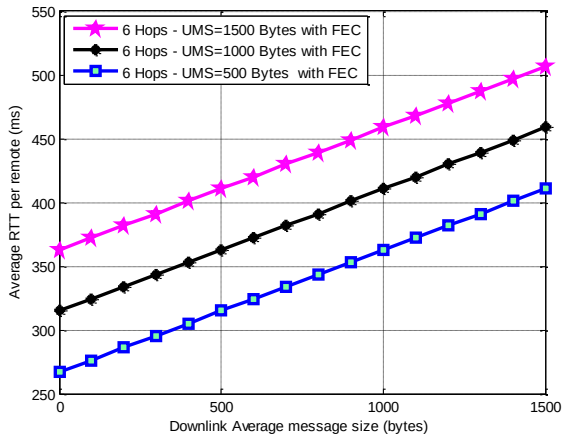


Fig. 7. Average RTT vs DMS: 6 Hops with FEC and different values of UMS

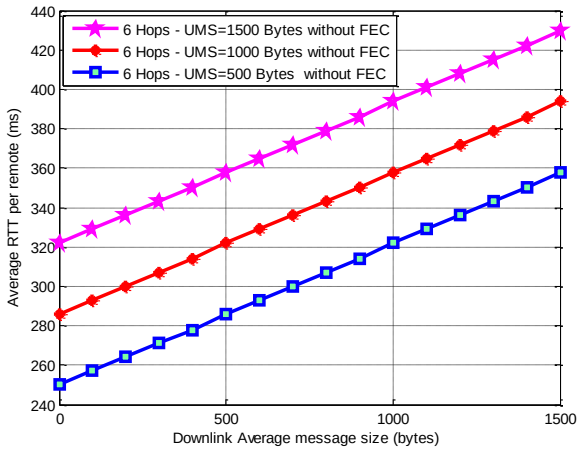


Fig. 8. Average RTT vs DMS: 6 Hops without FEC and different values of UMS

According to the results shown in Fig.7 and Fig.8, it is possible to observe that the RTT increases linearly as the DMS

increases. The increase of UMS can also promote a significant increase in RTT. In the same way, as previously discussed, the introduction of FEC can significantly affect the RTT, the results show that FEC can indeed increase the RTT.

For a fixed value of UMS= 1000 Bytes and ranging DMS from 0 up to 1500 Bytes, without FEC, the RTT for 6 hops increases to 108 msec . Whereas for FEC=3/4, RTT increases to 144 msec.

D. Average RTT per remote: Third scenario

In Fig.9 and Fig.10, the network topology in Fig.2 (c) is fully considered where the messages are transmitted from the (BS) to the RTU 3 and return to the (BS). The messages are relayed using the following RTT path, as in (13):

$$RTT(N_1 \rightarrow N_7 \rightarrow N_8 \rightarrow N_9 \rightarrow N_{10})$$

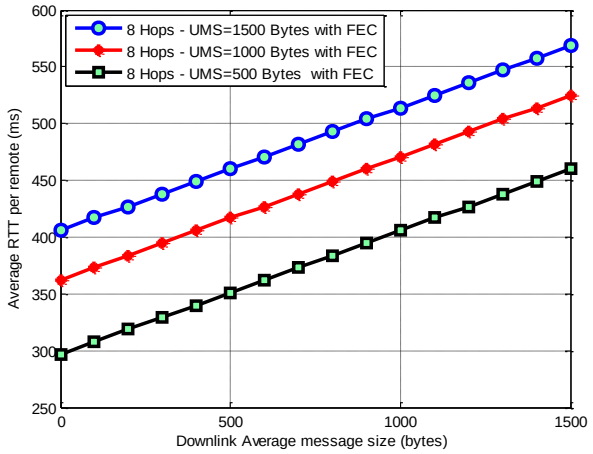


Fig. 9. Average RTT vs DMS: 8 Hops with FEC and different values of UMS

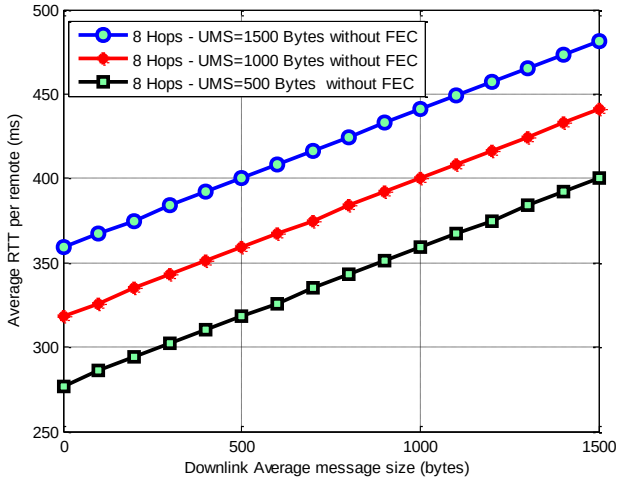


Fig. 10. Average RTT vs DMS: 8 Hops without FEC and different values of UMS

Similar to what happen in figs. 5-8, one can observe from the Fig.9 and Fig.10 for a large number of hops (8 hops) that, depending of the value of UMS, the RTT increases linearly when the DMS increases.

Once again, it can be noted that for a fixed value of UMS= 1000 Bytes and ranging DMS from 0 up to 1500 Bytes and without FEC, the RTT, for 8 hops, increases to 123 msec, whereas for FEC=3/4 it increases to 163 msec.

E. Comparison of Results

Fig.11 shows the comparison of results obtained by considering the first scenario for different UMS and DMS values and with/without FEC.

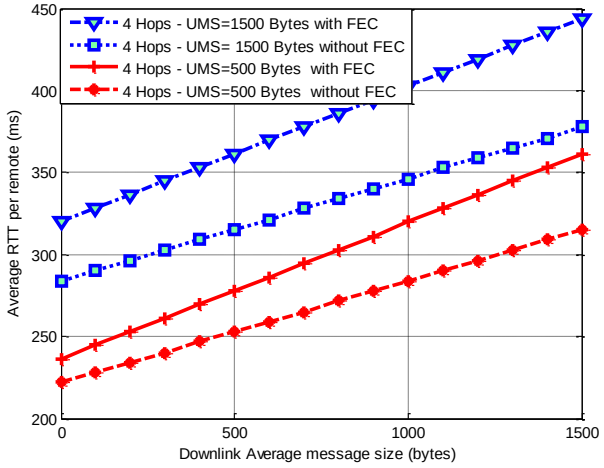


Fig. 11. The comparison of results obtained by considering a first scenario for different values of UMS and with/without FEC

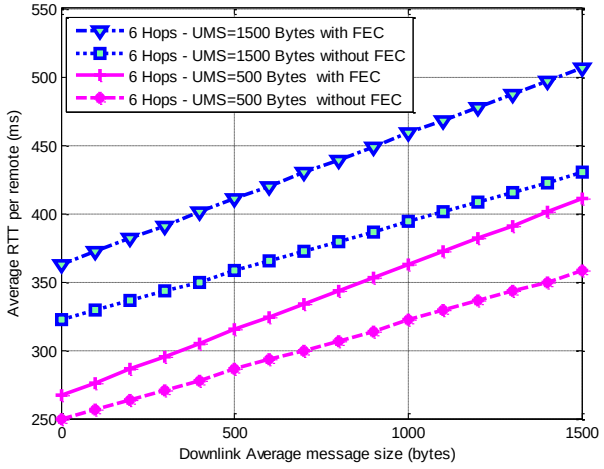


Fig. 12. The comparison of results obtained by considering a second scenario for different values of UMS and with/without FEC

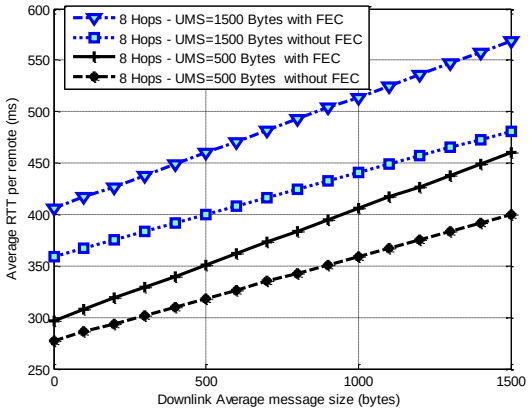


Fig. 13. The comparison of results obtained by considering a third scenario for different values of UMS and with/without FEC

The results in Fig.11 shows that the using different Downlink Average Message Size (bytes), Uplink Average Message Size (bytes) and FEC, in multi-hop wireless networks, have a significant effect on the Round-Trip Time (RTT) metric. We repeated the previous experiments with six and eight hops (second and third scenario) network respectively we obtained the same observations as illustrated in Fig 12 and Fig.13.

V. CONCLUSIONS

This paper presents a new method for measuring RTT giving a quick performance overview based on several basic parameters.

In light of the results obtained in real-time study metrics, it can therefore be concluded that RTT depends on various parameters such as the number of hops, DMS, UMS and FEC. Based on the results found, the most important concluding notes can be seen in the following:

- DMS has an effect on RTT measurements, when the DMS increases, the RTT also increases linearly in RTT path. Hence linear relationship exists between RTT and router node DMS in multi-hop wireless mesh networks.
- The offset between RTT, with and without FEC of different hop varieties, increases as the Downlink and Uplink message size increases.
- The UMS is another important factor that can affect the RTT measurements; it was observed that RTT increases as the UMS augments.
- Furthermore, it is observed that the coding technique Rate (FEC=3/4) provokes a further increase in RTT and the practical implementation of UMS, nodes and FEC, increase RTT without any influence on the linear relationship between RTT and router node DMS.

The offset between RTT, with and without FEC of different hop varieties, increases as the Downlink and Uplink message size increases.so, for relatively high error rates, it turns out to be better to send smaller packets, because when an error does occur then the entire packet containing it is lost.

REFERENCES

[1] El Miloud Ar-Reyouchi, Kamal Ghoumid, Koutaiba Ameziane, and Otman El Mrabet, “Performance Analysis of Round Trip Time in Narrowband RF Networks For Remote Wireless Communications”, International Journal of Computer Science & Information Technology (IJCSIT),5(5), pp.1-20, October 2013.

[2] Dalal, P., Sarkar, M., Kothari, N., and Dasgupta, K. ” Refining TCP's RTT dependent mechanism by utilizing link retransmission delay measurement in Wireless LAN”. Int J Commun Syst, 30,(5), 20170.

[3] El Miloud Ar Reyouchi, Kamal Ghoumid, Koutaiba Ameziane, Otman El Mrabet, Slimane Mekaoui, “Performance Analysis of Round Trip Delay Time in Practical Wireless Network for Telemanagement”, International Journal of Electrical, Computer, Energetic, Electronic and Communication Engineering , Waset, 7(11), pp.1413-1419, 2013.

[4] El Miloud Ar-reyouchi , Koutaiba Ameziane , Otman. El Mrabet , Kamal Ghoumid, “The potentials of Network Coding for improvement of Round Trip Time in wireless Narrowband RF communications”, Multimedia Computing and Systems (ICMCS), International Conference on, pp.765–770, 2014.

[5] Jain, M., Dovrolis, C. “End-to-end available bandwidth: measurement methodology, dynamics, and relation with tcp throughput”. In: SIGCOMM, ACM, 2002.

[6] Zhang, Y., Breslau, L., Paxson, V., Shenker, S.2002. “On the characteristics and origins of Internet flow rates”. In: SIGCOMM, ACM, 2002.

[7] Ratnasamy, S., Handley, M., Karp, R., Shenker, S. “Topologically-aware overlay construction and server selection”. In: INFOCOM, IEEE , 2002.

[8] Peter King “SCADA Systems – Looking Ahead“Control Microsystems White Paper, August 2005.

[9] STOUFFER, K., FALCO, J., AND KENT, K. Guide to supervisory control and data acquisition (scada) and industrial control systems security. Sp800-82, NIST, September 2006.

[10] K. Gowri Shankar “Control of Boiler Operation Using PLC- SCADA”, International Multi Conference of Engineers and Computer Scientists , Vol II, IMECS , Hong Kong , March 19-21, 2008.

[11] R. Baumann, S. Heimlicher, V. Lenders and M. May, "Routing Packets into Wireless Mesh Networks," Third IEEE International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob 2007), White Plains, NY,pp. 38-38.2007. doi: 10.1109/WIMOB.2007.

[12] Richard Draves, Jitendra Padhye Brian Zill, “Routing in multi-radio, multi-hop wireless mesh networks” Proceeding MobiCom '04 Proceedings of the 10th annual international conference on Mobile computing and networking Philadelphia, PA, USA — September 26 - October 01 pp. 114-128, 2004

[13] R. Draves, J. Padhye, and B. Zill, “Routing in Multi-Radio, Multi-Hop Wireless Mesh Networks,” ACM Annual Int’l. Conf. Mobile Comp. and Net. (MOBICOM), pp. 114–128, 2004.

[14] El Miloud Ar-Reyouchi,Youssra Chatei,Kamal Ghoumid,Ahmed Lichioui“The Powerful Combined Effect of Forward Error Correction and Automatic Repeat Request to Improve the Reliability in the Wireless Communications”, International Conference on Computational Science and Computational Intelligence (CSCI) Las Vegas, NV, USA Dec. , 2015,pp: 691-696, 2015.

[15] ETSI EN 300 113-1 V1.6.2 (2009-11), Electromagnetic compatibility and Radio spectrum Matters (ERM), Part 1: Technical characteristics and methods of measurement. European Standard. ETSI, 11/2009.

[16] ETSI EN 302 561 V1.2.1 (2009-12), Electromagnetic compatibility and Radio spectrum Matters (ERM), Land Mobile Service; Radio Equipment using constant or non-constant envelope modulation operating in a channel bandwidth of 25 kHz, 50 kHz, 100 kHz or 150 kHz; Harmonized EN covering essential requirements of article 3.2 of the R&TTE Directive. European Standard. ETSI, 12/2009.

[17] Mohammad Shorfuzzaman, Mehedi Masud and Md. Mahfuzur Rahman, “Characterizing End-to-End Delay Performance of Randomized TCP Using an Analytical Model” International Journal of Advanced Computer Science and Applications(IJACSA), 7(3), 2016.

Evaluating Predictive Algorithms using Receiver-Operative Characteristics for Coronary Illness among Diabetic Patients

Tahira Mahboob, Saman Sahaheen, Nuzhat Tahir, Mukhtiar Bano
Department of Software Engineering
Fatima Jinnah Women University, Pakistan

***Abstract*—The grouping of information is a typical method in Machine learning. Information mining assumes a crucial part to extract learning from vast databases from operational databases. In medicinal services Data mining is a creating field of high significance, giving expectations and a more profound comprehension of restorative information sets. Most extreme information mining technique relies on an arrangement of elements that characterizes the conduct of the learning calculation furthermore straightforwardly or by implication impact of the multifaceted nature of models. Coronary illness is the main sources of death over the past years. Numerous scientists utilize a few information digging methods for the diagnosing of coronary illness. Diabetes is one of the incessant maladies that emerge when the pancreas does not deliver enough insulin. The vast majority of the frameworks have effectively utilized Machine learning strategies, for example, Naïve Bayes Algorithm, Decision Trees, logistic Regression and Support Vector Machines to name a few. These techniques solely rely on grouping of the information with respect to finding the heart variations from the norm. Bolster vector machine is an advanced strategy has been effectively in the field of machine learning. Utilizing coronary illness determination, the framework presented predicts using characteristics such as, age, sex, cholesterol, circulatory strain, glucose and the odds of a diabetic patient getting a coronary illness using machine learning algorithms.**

Keywords—Artificial neural networks (ANN); Decision tree; Naïve Bayes; Logistic Regression and Clustering

I. INTRODUCTION

The improvements of system integration as well as software development techniques have made an advanced generation for the complex computer systems. The researchers have presented numerous challenges by using systems. Machine learning can be defined as a scientific field. The developed algorithms associates the real time problem based on previous statistics, and performs to resolve a real time problem under definite set of instructions and rules. Both machine learning and data mining algorithms use design formatted by means of same set of fields such as features, attributes, inputs, or variables. When an example or an instance contains the correct output (class label) then the learning process is known as supervised learning. In other words, the process of machine learning without knowing the class label of instances is called unsupervised learning. Clustering is an unsupervised learning method used for classifying data. The main objective of clustering is to describe data in an unsupervised learning ways. On the other hand, classification and regression are predictive methods. In the present research, we focus on supervised machine learning. The proposed study applies diverse algorithms such as Naïve Bayes, Decision Trees (C4.5) and Logistic Regression for Classification. Receiving Operating Characteristics (ROC) curve on the classification of algorithms has been analyzed for evaluation of predicted results on basis of data set attributes and values.

Coronary heart disease: Heart is a basic organ of our body as life is all subject to capable working of heart. If spread of blood is inefficient in the body that leads to the organs like kidney, brain tending to deteriorate and if heart stops, the end will be happen within minutes. The word Heart sickness suggests disease of heart and vein structure inside the heart. There are number of segments that extend the threat of Heart disorder, for instance, smoking, cholesterol, not exactly stellar eating schedule, hypertension, blood cholesterol, physical absence of movement, hyper weight and family history of coronary disease.

The rest of the paper is structured as Literature Survey included in Section-II followed by methodology in Section-III. Section-IV covers the implementation of the proposed study which is analyzed in Section-V. The last Section Covers the Results and Analysis along with Conclusion.

II. LITERATURE SURVEY

In this segment, we review the existing literature and confer about different aspects of data mining applications in prediction of heart diseases.

In Year 2007, Choi S.et.al [1] presents discovery strategy for heart deformities utilizing the cardiovascular sound trademark waveform (CSCW) with Information Grouping Method. The research inferred that presented framework is reasonable for the recognizable proof of patients with high/low cardiovascular danger.

In year 2012 Nambiar V. P. et.al [2] presents another system to perceive driver sluggishness by requesting the power scope of a man's Distinctions Heart Rate variability (HRV) data, which is made using Genetic Algorithm. The maker presumes that the precise level of drowsiness is hard to

choose and is not secured in the database used as a piece of this paper.

In the year 2013 M Makkiet.al [3] proposed the prediction of coronary illness using two receiving wires situated on the mid-section to screen the heart action. Synchronous with the radar estimations, the Electrocardiogram (ECG) and pulse is measured. The author presumes that the capacity to distinguish changes in the heart affirms medicinal radar as a reasonable analytic apparatus.

In year 2008 Han C. H. et.al [4] presents that strategy MCG (Magnetocardiogram) is additionally utilized for coronary illness recognition like the ECG (Electrocardiogram). The author presumes that it can't give the propelled e-Health administrations in light of immense measure of information, which may be handled and overseen.

In the year 2009 Wu W. et.al [5] presents a non-intrusive multi-channel ECG checking that is planned and executed to dissect the HRV of people amid various visual boosts. It is reasoned that a large portion of the HRV parameters changed fundamentally, and long haul negative visual jolt may render the capacity of ANS powerless which influence our body adversely. So positive visual incitement is required that helps us for more noteworthy fixation and spotlight on our undertakings, objectives and desires under the positive visual boost.

In year 2007, Manriquez A. et.al [6] presents another calculation proposed for QRS onset and counterbalance discovery in single lead electrocardiogram (ECG) records. It is inferred that the new calculations can be assessed with the PhysioNet QT database. As far as STD quality, it beats alternate calculations assessed on the same database. The location blunders in these outcomes are likewise around the resiliences acknowledged by specialists.

In the year 2014, Masethe H. D. et.al [7] exhibited that information mining calculations, for example, J48, Naïve Bayes, REPTREE, CART, and Bayes Net are connected in this examination for anticipating heart assaults. The author reasoned that the prescient precision controlled by J48, REPTREE and SIMPLE CART calculations proposes that the parameters utilized are solid pointers to anticipate the proximity of heart illnesses.

In year 2010 F. Sufi et.al [8] presents the risk of Cardiovascular Disease (CVD) related to the use of cellular telephone based computational stages, body sensors and remote correspondences is multiplying. Since cell phones have limited computational resources, existing PC based complex CVD disclosure figuring are every now and again prohibited for remote telecardiology applications Moreover, if the current Electrocardiography (ECG) based CVD recognition calculations are embraced for portable telecardiology applications, then there will be a need to handle delays because of the computational complexities of the current calculations.

In Year 2010, J. Bushra et.al [9] proposed another technique to identify the QRS Complex by wavelet based methodologies. The author investigates a non-stationary sign utilizing Gaussian wavelet and the recognized interest is indicated by relating them through the nearby extrema in wavelet change.

In the year 2009 Aardal, Ø., et.al [10] exhibited body sensor systems (BSNs) as high preparation is required in decompression that would squander profitable vitality in the asset and force obliged sensor hubs. In this paper, to analyze cardiovascular irregularity, for example, Ventricular tachycardia, a novel framework to break down and characterize compacted ECG signal by utilizing a PCA for highlight extraction and k-mean for bunching of ordinary and unusual ECG signals is proposed.

In year 2014 J. Lee et.al [11] presents that strategy ECG which analyzes cardiovascular ailments utilizing blood testing, ought to give early location to the ailment and more dependable checks. In year 2014 D. Khemphila, A., et.al [12] presents that technique to gain the clinical and ECG information, in order to prepare the Artificial Neural Network to precisely analyze the heart and foresee variations from the norm.

In year 2014 Mishra, S. K., et. al[13] presents telemetry framework to procure the Electrocardiogram (ECG) waveform and examine it utilizing a calculation created to identify cardiovascular irregularities in patients. The ECG information is sent continuously to the patient's cell phone from a Bluetooth sensor. Two methodologies are being produced to prepare the information. The Web server methodology is to send the information from the telephone to a Web server where the information will be examined and prepared and the outcomes will be sent back; this should be possible through Wi-fi or a 3G association.

In year 2014 P. Kaur et.al [14] presents in this work helps in location of heart rate, ECG variations from the norm and ensuing estimation of related infection utilizing different modules. The calculation outlined can get the information from measured document or recreate the ECG signal, form the information, and shows ECG waveform heart rate and its variations from the norm.

III. METHODOLOGY

The proposed strategy of the subsequent study is diagnosing powerlessness of patients of heart maladies. The methodology of the proposed study is presented in Fig.1.We took 303 records of patients to perform the experimentation. Thus the data set used for setting up the classifier contains 303 diabetic patient records out of which 175 records are of those having coronary sickness (positive cases) and the staying 128 records are of those not having coronary disease (negative cases). A sample of the characteristics making up each record/dataset is presented in Table1 and the attributes chosen for the prediction is presented in Table.2. The dataset is available at [*]"UCL Machine Learning Repository"

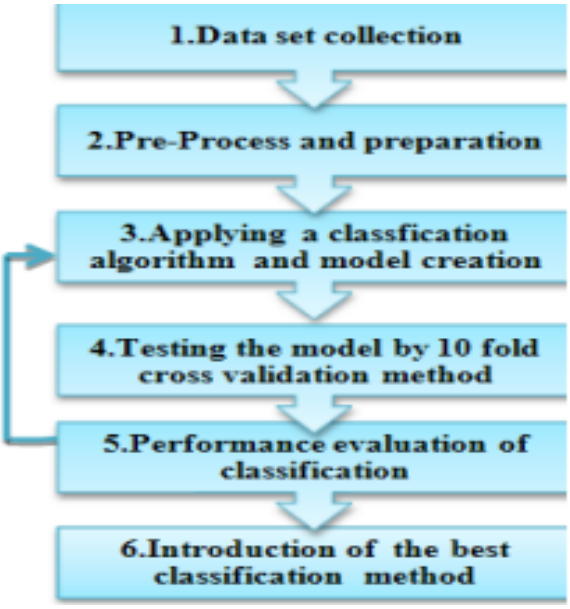


Fig. 1. Proposed Model of Data Flow

*The dataset of heart diseases is obtained from "UCL Machine Learning Repository" available at the link: "http://archive.ics.uci.edu/ml/"

TABLE I. DATA SET USED FOR DETECTION OF HEART ABNORMALITIES

Viewer

Relation: cleveland-14-heart-disease

No.	age Numeric	sex Nominal	cp Nominal	trestbps Numeric	chol Numeric	fbs Nominal	restecg Nominal	thalach Numeric	exang Nominal	oldpeak Numeric	slope Nominal	ca Numeric	thal Nominal	num Nominal
1	63.0	male	typ_a...	145.0	233.0	t	left_v...	150.0	no	2.3	down	0.0	fixed_...	(50
2	67.0	male	asympt	160.0	286.0	f	left_v...	108.0	yes	1.5	flat	3.0	normal	)50_1
3	67.0	male	asympt	120.0	229.0	f	left_v...	129.0	yes	2.6	flat	2.0	revers...	)50_1
4	37.0	male	non_a...	130.0	250.0	f	normal	187.0	no	3.5	down	0.0	normal	(50
5	41.0	female	atyp_...	130.0	204.0	f	left_v...	172.0	no	1.4	up	0.0	normal	(50
6	56.0	male	atyp_...	120.0	236.0	f	normal	178.0	no	0.8	up	0.0	normal	(50
7	62.0	female	asympt	140.0	268.0	f	left_v...	160.0	no	3.6	down	2.0	normal	)50_1
8	57.0	female	asympt	120.0	354.0	f	normal	163.0	yes	0.6	up	0.0	normal	(50
9	63.0	male	asympt	130.0	254.0	f	left_v...	147.0	no	1.4	flat	1.0	revers...	)50_1
10	53.0	male	asympt	140.0	203.0	t	left_v...	155.0	yes	3.1	down	0.0	revers...	)50_1
11	57.0	male	asympt	140.0	192.0	f	normal	148.0	no	0.4	flat	0.0	fixed_...	(50
12	56.0	female	atyp_...	140.0	294.0	f	left_v...	153.0	no	1.3	flat	0.0	normal	(50
13	56.0	male	non_a...	130.0	256.0	t	left_v...	142.0	yes	0.6	flat	1.0	fixed_...	)50_1
14	44.0	male	atyp_...	120.0	263.0	f	normal	173.0	no	0.0	up	0.0	revers...	(50
15	52.0	male	non_a...	172.0	199.0	t	normal	162.0	no	0.5	up	0.0	revers...	(50
16	57.0	male	non_a...	150.0	168.0	f	normal	174.0	no	1.6	up	0.0	normal	(50
17	48.0	male	atyp_...	110.0	229.0	f	normal	168.0	no	1.0	down	0.0	revers...	)50_1
18	54.0	male	asympt	140.0	239.0	f	normal	160.0	no	1.2	up	0.0	normal	(50
19	48.0	female	non_a...	130.0	275.0	f	normal	139.0	no	0.2	up	0.0	normal	(50
20	49.0	male	atyp_...	130.0	266.0	f	normal	171.0	no	0.6	up	0.0	normal	(50
21	64.0	male	typ_a...	110.0	211.0	f	left_v...	144.0	yes	1.8	flat	0.0	normal	(50
22	58.0	female	typ_a...	150.0	283.0	t	left_v...	162.0	no	1.0	up	0.0	normal	(50
23	58.0	male	atyp_...	120.0	284.0	f	left_v...	160.0	no	1.8	flat	0.0	normal	)50_1
24	58.0	male	non_a...	132.0	224.0	f	left_v...	173.0	no	3.2	up	2.0	revers...	)50_1
25	60.0	male	asympt	130.0	206.0	f	left_v...	132.0	yes	2.4	flat	2.0	revers...	)50_1
26	50.0	female	non_a...	120.0	219.0	f	normal	158.0	no	1.6	flat	0.0	normal	(50
27	58.0	female	non_a...	120.0	340.0	f	normal	172.0	no	0.0	up	0.0	normal	(50
28	66.0	female	typ_a...	150.0	226.0	f	normal	114.0	no	2.6	down	0.0	normal	(50
29	43.0	male	asympt	150.0	247.0	f	normal	171.0	no	1.5	up	0.0	normal	(50
30	40.0	male	asympt	110.0	167.0	f	left_v...	114.0	yes	2.0	flat	0.0	revers...	)50_1
31	69.0	female	typ_a...	140.0	239.0	f	normal	151.0	no	1.8	up	2.0	normal	(50
32	60.0	male	asympt	117.0	230.0	t	normal	160.0	yes	1.4	up	2.0	revers...	)50_1
33	64.0	male	non_a...	140.0	335.0	f	normal	158.0	no	0.0	up	0.0	normal	)50_1
34	59.0	male	asympt	135.0	234.0	f	normal	161.0	no	0.5	flat	0.0	revers...	(50
35	44.0	male	non_a...	130.0	233.0	f	normal	179.0	yes	0.4	up	0.0	normal	(50
36	42.0	male	asympt	140.0	226.0	f	normal	178.0	no	0.0	up	0.0	normal	(50
37	43.0	male	asympt	120.0	177.0	f	left_v...	120.0	yes	2.5	flat	0.0	revers...	)50_1
38	57.0	male	asympt	150.0	276.0	f	left_v...	112.0	yes	0.6	flat	1.0	fixed_...	)50_1
39	55.0	male	asympt	132.0	353.0	f	normal	132.0	yes	1.2	flat	1.0	revers...	)50_1

Undo OK Cancel

TABLE II. ATTRIBUTES TAKEN FOR DETECTION OF HEART DISEASE

Sr.#	Attribute name	Depiction/values
1	Age	Age of the patients in year
2	Sex	Sex of patients takes two value(Female ,Male)
3	chest pain(Cp)	chest pain type in the heart patients Value 1: typical angina Value 2: atypical angina Value 3: non-angina pain Value 4: asymptomatic
4	resting blood pressure(Trestbps)	resting blood pressure (in mm Hg on admission to the hospital)
5	cholesterol (chol)	Heart patient serum cholesterol in mg/dl
6	fasting blood sugar (fbs)	(fasting blood sugar > 120 mg/dl)
7	resting electrocardiographic (restecg)	latent electrocardiographic effects Value 0: normal Value 1: having ST-T wave abnormality
8	thalach	maximum heart rate achieved
9	(exang)	implementation prompted angina (1 = yes; 0 = no)
10	old peak	ST dejection made by exercise relative to rest
11	slope	the slope of the peak exercise ST segment Value 1: upsloping Value 2: flat Value 3: down sloping
12	ca	quantity of main vessels (0-3) colored by fluoroscopy
13	thal	3 = normal; 6 = fixed defect; 7 = reversible defect
14	num	analysis of core disease (angiographic disease status) Value 0: < 50% diameter narrowing Value 1: > 50% diameter narrowing

IV. IMPLEMENTATION

Data Mining Techniques Used For Predictions: The diverse information mining arrangement methods utilized as a part of our examination, i.e. Neural Networks, Decision Trees, bunching, logistic relapse, and Naive Bayes are utilized to break down the dataset of coronary illness patients targeting the prediction of coronary illnesses. As a means to evaluate the accuracy of predictive capacity of the applied algorithms the receiving operating characteristics [15] of each have been plotted followed by a comparison to find the best fit algorithms for predictions. [16] R.O.C outline is a valuable visual instrument for looking at order techniques. It demonstrates the exchange off between the genuine positive rate and the false positive rate for a given model. ROC outline depends on the restrictive probabilities affectability and specificity.

A. Neural Networks:

A simulated neural system Artificial Neural Networks (ANN) regularly called as "neural system" (NN) [17]. It is a numerical model or computational model taking into account natural neural system. As it were, it is a reproduction of organic neural framework in the field of restorative. The receiving operating characteristics have first been plotted in Fig. 2 depicting the correctness of the results using ANN algorithm and the ROC measurement observed is 0.891.

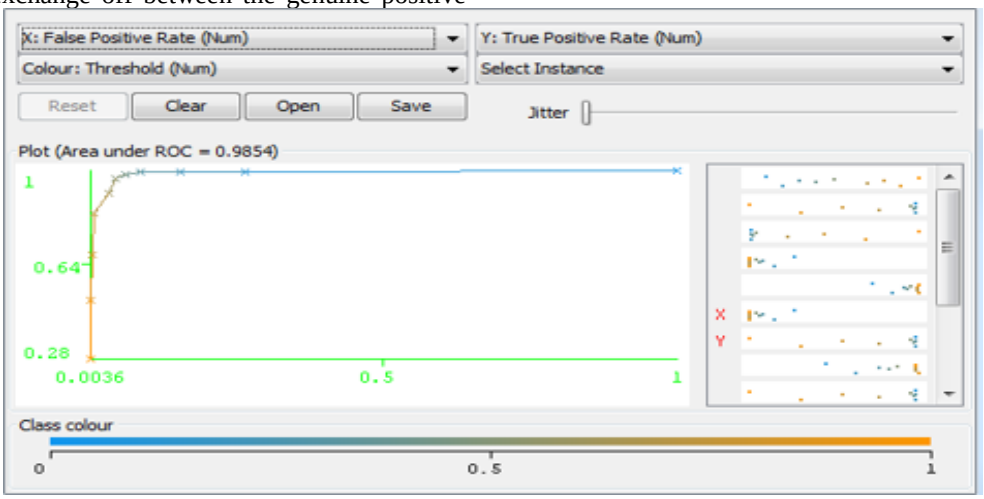


Fig. 2. ROC of Artificial Neural Networks

B. Decision Trees:

The decision tree methodology is all the more effective for the arrangement of issues. There are two phases used as a piece of this framework; building a tree and applying the tree to the dataset of coronary ailment patients. The Algorithm count uses pruning methodology to create a tree. Pruning is a system which diminishes the degree of tree by ousting over fitting data, which may prompts poor precision in predications examination.[18] [19]The DT figures recursively gatherings' data until it will sort data as sublimely as could sensibly be

normal. This methodology gives most amazing exactness on planning data sets. A sample DT result is given in Fig. 4 depicting/Classifying attribute 'Chest Pain' as 'unstable angina', 'stable angina' and "non-angina'. The receiving operating characteristics of the DT are shown in Fig. 4 giving a value of 0.952.

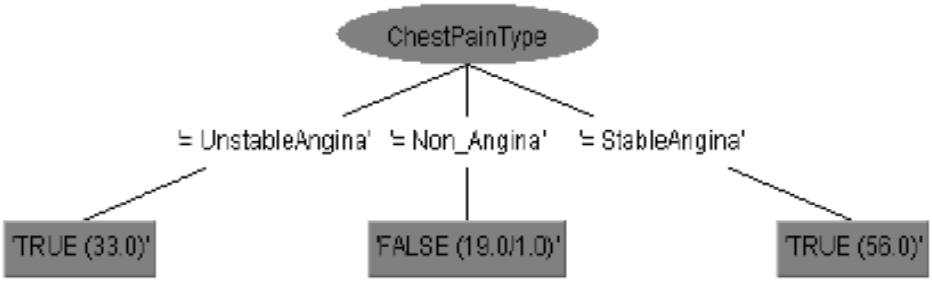


Fig. 3. Decision Tree of chest pain type

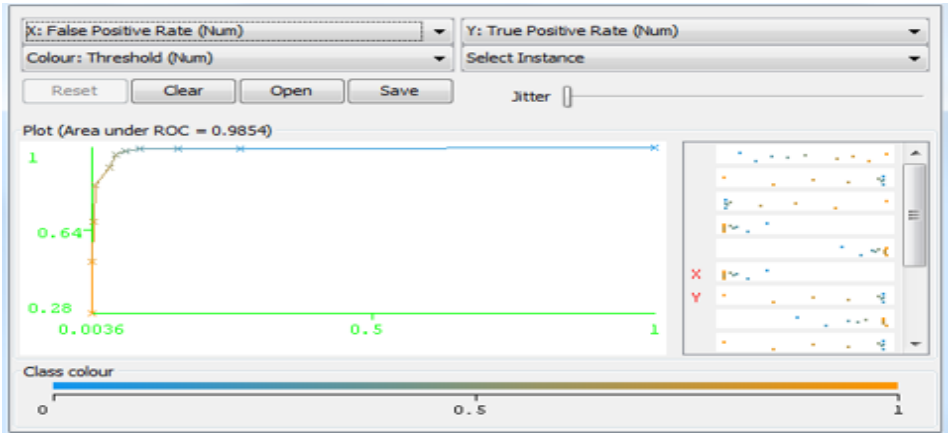


Fig. 4. ROC Decision Tree

C. Naive Bayes:

Naive Bayes classifier depends on Bayes hypothesis. This classifier calculation utilizes contingent freedom, implies as it accept that a quality worth on a given class is autonomous of the estimations of different traits. [20][21]The ROC curve for the NaiveBayes algorithm is given in Fig.5. Though very simple to implement but doesn't give a high value for ROC measurement as compared to other implemented algorithms. i.e. 0.91

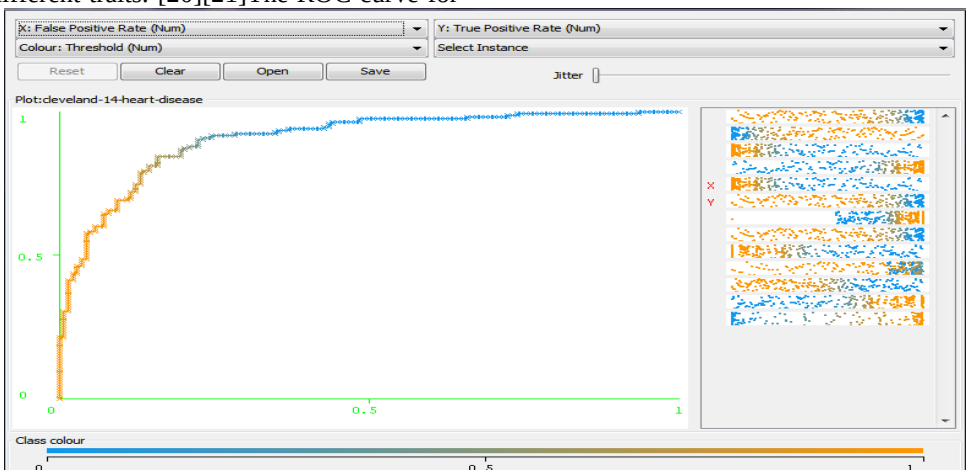


Fig. 5. ROC Naive Baye

D. Clustering:

Clustering is an information mining procedure that makes valuable group of items that have same trademark utilizing this system. Not quite the same as grouping, clustering strategy likewise characterizes the classes and place objects in them, while in order articles are doled out into predefined classes [22]. For instance in expectation of coronary illness by utilizing grouping we get bunch or we can say that rundown of

patients which have same danger variable such as different rundown of patients with high glucose and related danger component. Results predominant cluster having instances 175 in cluster 1 and 128 in cluster 2.

=== Model and evaluation on training set ===

Clustered Instances

0	175 (58%)
1	128 (42%)

E. Logistic regression:

Logistic regression is a measurable technique for examining a dataset in which there are one or more autonomous variables that decide a result. [23] The result is measured with a dichotomous variable (in which there are just two conceivable results). The ROC of the Logistic regression is presented in Fig.6 giving an ROC measurement of 0.938.

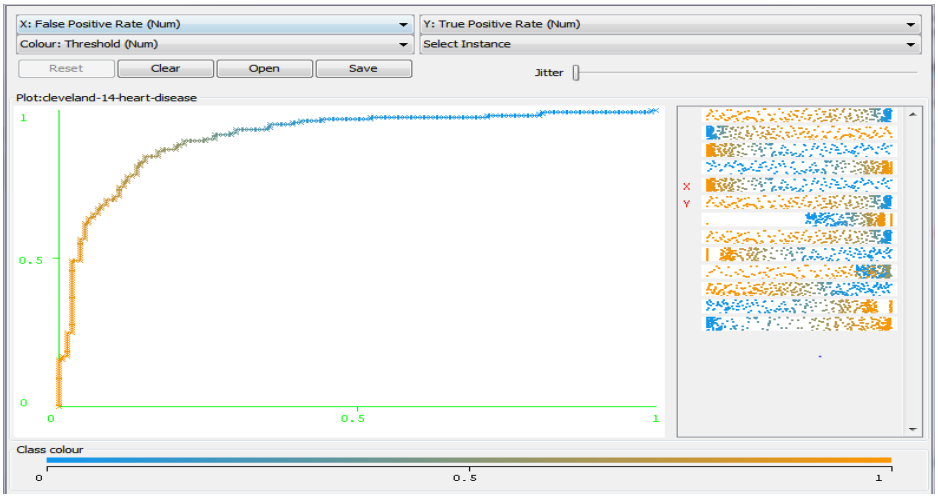


Fig. 6. ROC Logistic regression

F. Prediction Analysis:

The prediction as it name suggested is one of an information mining strategies that finds relationship between autonomous variables and relationship amongst reliant and free variables. For example, expectation examination system can be utilized as a part of offer to anticipate benefit for the future on the off chance that we consider deal is an autonomous variable, benefit could be a variable. At that point taking into account the chronicled deal and benefit information, we can draw a fitted relapse bend that is utilized revenue driven expectation. The ROC measurement for Prediction Analysis is. 0.908.

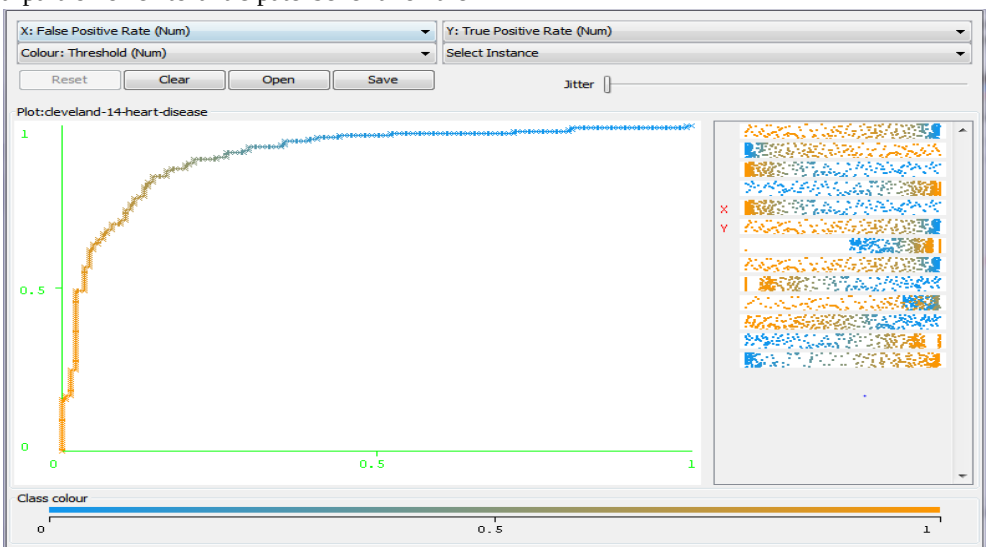


Fig. 7. ROC Prediction Analysis

G. Support Vector Machine:

The data set used for setting up the classifier contains 303 diabetic patient records out of which 175 records are of those having coronary sickness (positive cases) and the staying 128 records are of those not having coronary disease (negative cases). These records after satisfactory pre-get ready are given as commitment to set up the SVM classifier. Support Vector machines have been used as a classifier for feature selection in CAD disease in combination with genetic algorithms [24]. Also [SVM technique has been used to predict heart abnormalities as Babaoğlu, I., et al. 25]

The confusion matrix demonstrating the precision of the SVM classifier for the given information set is presented in Table 4. The confusion matrix is a Visualization apparatus utilized as a part of administered realizing which contains real and predicated grouping. Every section speaks to example in a predicated class and every line speaks to case in a genuine class.

Fig. 8 shows the comparative plots of the ROC of that servers as a basis for analysis.

TABLE III. CONFUSION MATRIX

a	b	<-- classified as
181	23	a = no
48	51	b = yes

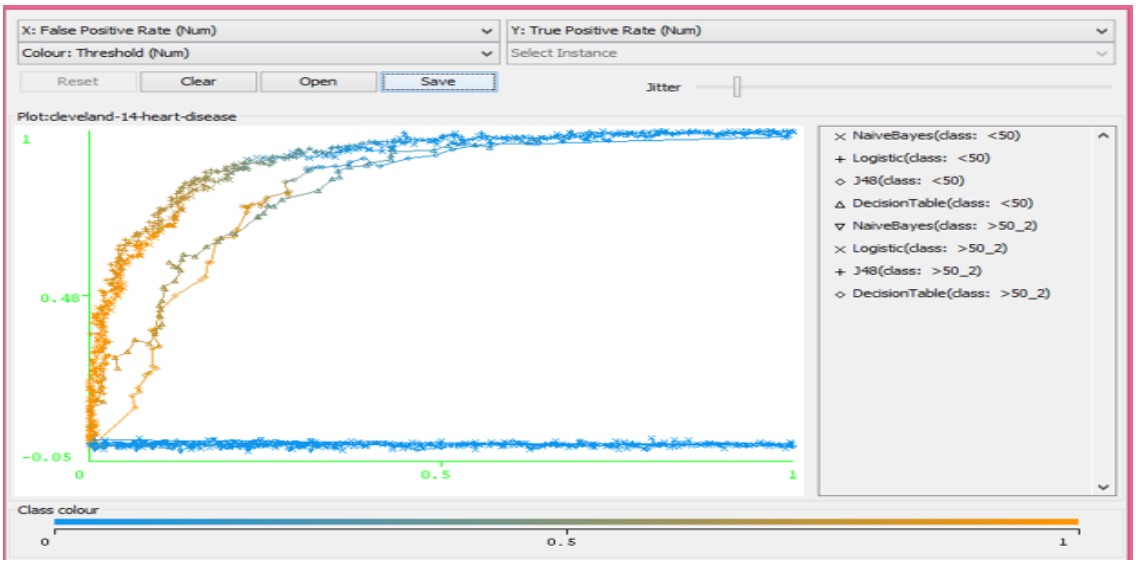


Fig. 8. Comparative ROC Curve for Various Techniques

V. ANALYSIS AND RESULTS

A comparative analysis of the ROC curves for implemented techniques has been presented in Fig. 8 where the Decision Tree algorithm produces the best curve having value of 0.952. The Accuracy measurements of the results obtained are presented in Table 4. Hence it can be concluded that the DT are best classifier for the prediction of heart disease given the dataset. The results shows that the DT algorithm applied on the data set is more accurate than the rest of the algorithm used such as Naive Bayes, Logistic Regression, ANN, The DT shows the 92% result to predict the abnormality found in the human heart. The whole model is proposed to collect data and analyzed using machine learning algorithm one by one. The value is predicted by correctly classified the instances. As the instances found that are correctly classified shows that the people who have chronic heart diseases. By predicting the results we reach the results to find the heart abnormalities in patients. As the rest of algorithms such as Naive Bayes show 83.49% correctly specified instances then decision tree specified 82.83% results which are correctly specified. ANN specifies correct parameters about 80.85%. Logistic Regression specifies correct instances 87.13% and Prediction 76.57%. A comparative graph clearly depicts the discussed results. The results show that we can find a way to specify correct instances and which algorithm is best to use.

TABLE IV. RESULT ANALYSIS ON THE BASIS OF ACCURACY

Sr.#	Algorithm	W.avg TP Rate	W.avg FP Rate	W.avg Precision	W.avg Recall	W.avg F-Measure	W.avg ROC Area	Accuracy
1.	Naïve bayes	0.835	0.171	0.835	0.842	0.841	0.91	83.5%
2.	Decision Tree	0.921	0.085	0.922	0.921	0.921	0.952	92%
3.	ANN	0.809	0.194	0.809	0.809	0.809	0.891	81%
4.	Predication	0.766	0.363	0.757	0.766	0.756	0.908	76%
5.	Logistic Aggression	0.871	0.136	0.872	0.871	0.871	0.938	84.4%

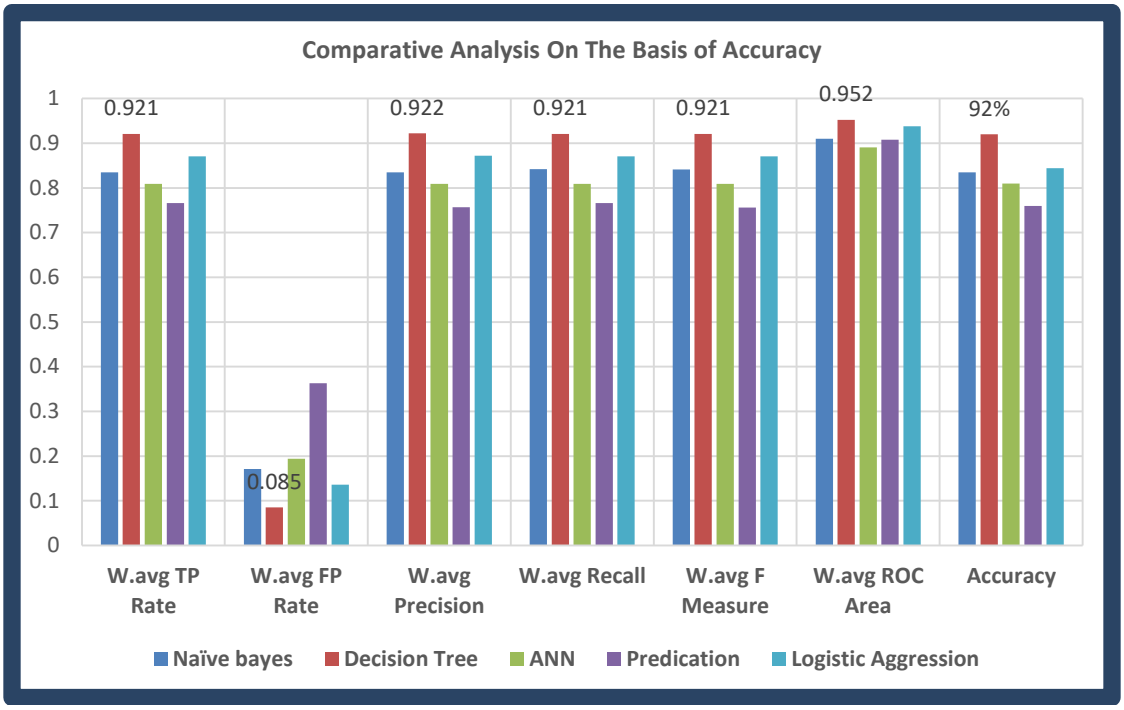


Fig. 9. Accuracy Measures Comparison of Algorithms

TABLE V. COMPARATIVE ANALYSIS OF RESULTS

Algorithms	Naïve Bayes		Decision Tree		Prediction		ANN		Logistic Regression	
Correctly Classified instances	253	83.49%	279	92.08%	232	76.57%	245	80.85%	264	87.13%
Incorrectly Classified instances	50	16.50%	24	7.92%	71	23.43%	58	19.14%	39	12.87%
Kappa Statistics	0.6661		0.8396		0.4304		0.6141		0.7391	
Mean Absolute Error	0.0738		0.0532		0.2976		0.0772		0.0775	
Root mean squared error	0.2299		0.1624		0.4449		0.2544		0.1954	
Relative absolute error	36.80%		26.56%		67.58%		38.48%		38.65%	
Root relative squared error	72.97%		51.54%		94.85%		80.74%		62.02%	
Total Number of instances	303		303		303		303		303	

VI. CONCLUSION

Application of Machine learning in analyzing the therapeutic information is a decent technique for considering the current connections between variables. From our proposed approach we have demonstrated that it recovers valuable connection even from qualities which are not immediate pointers of the class we are attempting to foresee. In our work we have attempted to foresee the odds of getting a coronary illness utilizing qualities from diabetic's determination and we have demonstrated that it is conceivable to analyze coronary illness powerlessness in coronary illness patients with sensible precision. There by the patients can be forewarned to change their lifestyle. We can easily get the required and accurate result to detect the heart abnormalities using machine learning and data mining algorithms. In future work we can extend this research further in all aspects of learning to the system. We will analyze and learn to the system for more accurate results by measuring more analysis using machine learning algorithms in the current research.

REFERENCES

[1] Choi, S., Jiang, Z., Kim, I. H., & Park, C. W. (2007, October). Cardiovascular abnormality detection method using cardiac sound characteristic waveform with data clustering technique. In Control, Automation and Systems, 2007. ICCAS'07. International Conference on (pp. 1596-1600). IEEE.

[2] Nambiar, V. P., Khalil-Hani, M., Sia, C. W., & Marsono, M. N. (2012, October). Evolvable block-based neural networks for classification of driver drowsiness based on heart rate variability. In Circuits and Systems (ICCAS), 2012 IEEE International Conference on (pp. 156-161). IEEE.

[3] Makki, M. (2013, November). Analyzing Electrocardiograms Via Smartphone To Detect Cardiovascular Abnormalities. In Qatar Foundation Annual Research Conference (No. 2013, pp. BIOSP-014).

[4] Han, C. H., Youn, C. H., & Jung, W. (2008, June). Web-Based System for Advanced Heart Disease Identification Using Grid Computing Technology. In Computer-Based Medical Systems, 2008. CBMS'08. 21st IEEE International Symposium on (pp. 343-348). IEEE.

[5] Wu, W., Lee, J., & Chen, H. (2009, August). Estimation of heart rate variability changes during different visual stimulations using non-invasive continuous ecg monitoring system. In Bioinformatics, Systems Biology and Intelligent Computing, 2009. IJCBS'09. International Joint Conference on (pp. 344-347). IEEE.

[6] Manriquez, A. I., & Zhang, Q. (2007, August). An algorithm for QRS onset and offset detection in single lead electrocardiogram records. In 2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (pp. 541-544). IEEE.

[7] Masethe, H. D., & Masethe, M. A. (2014, October). Prediction of heart disease using classification algorithms. In Proceedings of the world congress on engineering and computer science (Vol. 2, pp. 22-24)

[8] Sufi, F., Khalil, I., & Tari, Z. (2010, August). A cardiod based technique to identify cardiovascular diseases using mobile phones and body sensors. In 2010 Annual International Conference of the IEEE Engineering in Medicine and Biology (pp. 5500-5503). IEEE.

[9] Bushra, J., Ouadi, B., Eric, F., & Olivier, L. (2010, August). QRS Complex Detection by Non Linear Thresholding of Modulus Maxima. In Pattern Recognition (ICPR), 2010 20th International Conference on (pp. 4500-4503). IEEE

[10] Aardal, Ø., Brovoll, S., Paichard, Y., Berger, T., Lande, T. S., & Hamran, S. E. (2013, April). Detecting changes in the human heartbeat with on-body radar. In 2013 IEEE Radar Conference (RadarCon13) (pp. 1-6). IEEE.

[11] Lee, J., Jung, J., Lee, J., & Kim, Y. T. (2014, May). Diagnostic device for acute cardiac disease using ECG and accelerometer. In 2014 International Conference on Information Science & Applications (ICISA) (pp. 1-3). IEEE.

[12] Khemphila, A., & Boonjing, V. (2010, October). Comparing performances of logistic regression, decision trees, and neural networks for classifying heart disease patients. In Computer Information Systems and Industrial Management Applications (CISIM), 2010 International Conference on (pp. 193-198). IEEE.

[13] Mishra, S. K., Daman, R., Kumar, A., & Singh, S. (2014). Design and Development of Workspaces for Surgical Skill Development. Med-e-Tel 2014, 17.

[14] Kaur, P., & Sharma, R. K. (2014, May). Labview based design of heart disease detection system. In Recent Advances and Innovations in Engineering (ICRAIE), 2014 (pp. 1-5). IEEE.

[15] Polonsky, T. S., McClelland, R. L., Jorgensen, N. W., Bild, D. E., Burke, G. L., Guerci, A. D., & Greenland, P. (2010). Coronary artery calcium score and risk classification for coronary heart disease prediction. Jama, 303(16), 1610-1616.

[16] Zurada, J. M. (1992). Introduction to artificial neural systems (Vol. 8). St. Paul: West.

[17] Quinlan, J. R. (1987). Simplifying decision trees. International journal of man-machine studies, 27(3), 221-234.

[18] Schmid, H. (2013, November). Probabilistic part-ofispeech tagging using decision trees. In New methods in language processing (p. 154). Routledge.

[19] Rish, I. (2001, August). An empirical study of the naive Bayes classifier. In IJCAI 2001 workshop on empirical methods in artificial intelligence (Vol. 3, No. 22, pp. 41-46). IBM New York.

[20] Pattekari, S. A., & Parveen, A. (2012). Prediction system for heart disease using Naïve Bayes. International Journal of Advanced Computer and Mathematical Sciences, 3(3), 290-294.

[21] Soni, J., Ansari, U., Sharma, D., & Soni, S. (2011). Predictive data mining for medical diagnosis: An overview of heart disease prediction. International Journal of Computer Applications, 17(8), 43-48.

[22] Soni, J., Ansari, U., Sharma, D., & Soni, S. (2011). Predictive data mining for medical diagnosis: An overview of heart disease prediction. International Journal of Computer Applications, 17(8), 43-48.

[23] Steinwart, I., & Christmann, A. (2008). Support vector machines. Springer Science & Business Media.

[24] Babaoglu, İ., Findik, O., & Ülker, E. (2010). A comparison of feature selection models utilizing binary particle swarm optimization and genetic algorithm in determining coronary artery disease using support vector machine. Expert Systems with Applications, 37(4), 3177-3183.

[25] Babaoğlu, I., Fındık, O., & Bayrak, M. (2010). Effects of principle component analysis on assessment of coronary artery diseases using support vector machine. Expert Systems with Applications, 37(3), 2182-2185.

Self-Protection against Insider Threats in DBMS through Policies Implementation

Farukh Zaman, Basit Raza
Department of Computer Science
COMSATS Institute of Information Technology
Islamabad, Pakistan

Ahmad Kamran Malik, Adeel Anjum
Department of Computer Science
COMSATS Institute of Information Technology
Islamabad, Pakistan

***Abstract*—In today's world, information security of an organization has become a major challenge as well as a critical business issue. Managing and mitigating these internal or external security related issues, organizations hire highly knowledgeable security expert persons. Insider threats in database management system (DBMS) are inherently a very hard problem to address. Employees within the organization carry out or harm organization data in a professional manner. To protect and monitor organization information from insider user in DBMS, the organization used different techniques, but these techniques are insufficient to secure their data. We offer an autonomous approach to self-protection architecture based on policy implementation in DBMS. This research proposes an autonomic model for protection that will enforce Access Control policies, Database Auditing policies, Encryption policies, user authentication policies, and database configuration setting policies in DBMS. The purpose of these policies to restrict insider user or Database Administrator (DBA) from malicious activities to protect data.**

Keywords—autonomic; self-protection; insider threats; policies; DBMS

I. INTRODUCTION

Data is probably most important and valuable asset on which entire organization depends. However, it's difficult to memorize some data so these data should be kept in an organized way in a special storage location called databases. It's necessary to build a trustworthy relationship with an organization and its clients by protecting its data from possible threats. Data should protect by imposing CIA (Confidentiality, Integrity, and Availability) security model which should be guaranteed in any kind of security system [5] [34] [35] [36] [37] [38]. Without CIA security model data can be lost or destroyed. Some security threat against database management systems are:

- Misuse of sensitive data by the authenticated user
- Malware infection causing damage to data or programs
- Physical damage of database server
- Weak parameter setting or design flaws causing vulnerabilities in DBMS
- Unauthorized access of DBMS

Database threat may have initiated either in an external way or from within an organization. The external threat can be detected by imposing software tools and technologies such as Firewall, network traffic monitoring, enforcing password mechanism and penetration testing [4]. However, it's difficult to monitor insider's intent. According to CERT survey, more than 700 cases were caused by the insider threats [6]. To protect against these threats database should have some extra features of Autonomic Computing like self-protection. We first provide an introduction to Autonomic computing and its components.

Autonomic computing has the ability to self-manage its system [39] [40]. It controls all the functionality of computer systems or applications without any user involvement. Autonomic computing concept is taken from human body's autonomic nervous system, which controls human body functions such as heart rate, respiratory rate, pupillary response parts and Digestive system without the conscious input of an individual [2]. How the human body mechanisms manage itself without external involvement in many cases? The main objective of autonomic computing is to build a system that has a self-managed characteristic and make a decision on its own by using high-level functionalities when any unpredictable problem occurs. Autonomic computing framework based on autonomic components that interact with each other. The autonomic computing system has the ability to respond to any problems occur and make the system precise and available to the user. Instead of directly user input in the system, User defines general procedures and policies that guide the self-management process. IBM defines four main self-* components [7] [41] [42] [43] [44] [45].

- Self-optimization
- Self-healing
- Self-configuration
- Self-protection.

Some other extended self-* features are defined as in [8] are Self-Adaption, Self-regulation, Self-learning, Self-awareness, Self-organization, Self-creation, Self-management and Self-descriptive. When all these self-* features of self-managed apply to any system that system has the ability to protect any external or internal threats and heal itself when it is needed without any user input [9][3]. Autonomic functions and their management are automated in a control loop task called MAPE. Self-optimization consists of the system's automatic ability to configure and optimize itself to achieve top level performance against current settings, workload, and resources [9]. In DBMS environment different features are

used to achieve the best optimization. The query optimizer is used to optimize and execute the query execution plan. The Database statistic manager is used to collect statistics of database objects. Such features are already configured to obtain self-optimization in DBMS.

Self-healing is to recover the damaged part or data automatically without any human intervention in order to remain active and operating correctly [9] [45]. Self-healing is a grand-challenge to an autonomic system which first detects a problem in the system, diagnoses it, and then repairs it automatically. Self-Healing deals with lacking precision in the uncontrolled situation and recovers it according to the dynamics. Healing the system is a serious problematic situation when the information is being corrupted by a malicious attack or any insider's malicious intent or by mistake as this could lead to disastrous decisions when it comes to Military or Health database. For this, the system must be smart enough that it can detect the problem, prepare a plan against it and execute it to bring the database to a normal state.

An autonomic computing system configures its components automatically to achieve its goal [9]. In this environment, the system automatically detects changes and configures, reconfigures its components accordingly [48]. Since the adaptation needs to achieve optimal performance, the category of self-configuring is close to self-optimize. Following features provide self-configuration in autonomic DBMS: Memory components, dynamic parameter configuration, supporting objects for performance purpose, such as indexes, materialized views, partitions, etc. are all components which are used to provide self-configuration ability in the Database. Self-protection is a key component of self-managed systems capable of automatically defend against malicious attacks at runtime. A self-protecting system or application proactively identifies malicious threats and triggers necessary actions to stop them [9] [46] [47]. Security professionals used different tools and skills such as (protection filters, detectors of suspicious activity, logging mechanism & backtracking tools) to protect their systems [1].

The organization of this research paper comprises of the following sections. Section 2 discusses autonomic computing in Database Management system that mainly focuses on the self-protecting perspective. Section 3 discusses current approaches to database protection and section 4 present proposed autonomic model w.r.t self-protection. provides analysis and discussion of database protection and section 5 concludes the research and provide future directions.

II. Autonomic Computing in DBMS

In today's era Complex Databases and their manageability have become a serious concern for organizations nowadays. These databases need to be easily accessible and available to their clients. For this purpose, it requires expert Database Administrators (DBA) for their continuous monitoring, evaluation, and availability. Keeping in view the scarcity of such expert Database Administrators in the market and the cost of their hiring, the concept of the Autonomic Database Management System is introduced which is capable of managing and maintaining such databases without any human intervention [2].

A. Self-Protection in DBMS

Self-protection of the database is to protect your data from both external threats and internal threats and make available 24/7 to their clients. Experienced DBAs are being hired by organizations for continuous monitoring and availability of complex databases. As a DBA has full access to the database so he or she can easily carry out or harm organization data. The organization uses different techniques and methods to protect their information or data from the internal user, but these techniques and methods are insufficient or not enough. In this regard, the database should have some extra ability or features of autonomic computing, i.e. Self-healing, Self-protection, Self-configuration and Self-optimization to protect and manage its information without any human interventions. The autonomic computing system has the capability to respond automatically to any issue occurred and to make the system precise and available to the user.

A number of authors use different techniques and approaches to achieve database security. Data is an important asset for any organization and its security is critical for maintaining the relationship between an organization and its end users. Different techniques such as access control, encryption scheme, auditing policies, and inference control are used in database management system by a different researcher. While combining autonomic properties such as self-healing and self-protection with database security features such as access control, encryption, database auditing features, we can get the more secure DBMS without the involvement or intervention of any DBA or security engineer. Such autonomic properties are very useful for insider threat or monitoring DBA activities. Table I, presents protection techniques against different attacks and Self-protection of external threat is mostly implemented by configuring the firewall and network traffic monitoring. On the other hand, self-protection against internal threat or insider's malicious intent should achieve by obtaining best security policies [2]. Implementing these policies within a database block every attempt to compromise the state of the database. Database security achieved by user access control mechanism and by using stored procedures to manage the internal database threat. When the attacker attempts a request to change security configuration request carried to the stored procedure for verification. Fig 1 shows some critical areas need to be considered in Database Security [5] and how different researcher use different techniques and methods to mitigate these risks.

TABLE. I. PROTECTION TECHNIQUES AGAINST DIFFERENT ATTACKS

Protection Techniques	Attack type	Reference
Access Control Policies	Used for both insider and outsider attacks	[11, 12, 16, 19, 22]
Mixed Cryptographic Database	Used for both insider and outsider attacks	[13]
RSA Encryption Technique	Insider attack	[15,17]
Attributes Based Encryption	Used for both insider and outsider attacks	[3]
Hash-Based Encryption	Used for both insider and outsider attacks	[18, 28, 29]
Data Centric Approach	Insider attack	[23]
SQL Injection and insider misuse detection system	Used for both insider and outsider attacks	[20]
Auditing Method	Used for both insider and outsider attacks	[24, 27]

Hackers exploit these critical areas and security holes in a database application to gain database administrator (DBA) level grants and privileges to access sensitive data and cause a denial of service (DOS) attacks. Following are the security threats that need attention [10].

- Excessive and unused privileges: granted extra privileges to user that exceed the requirement of their job function
- Privilege abuse: authenticated user misuse authentic database privileges for illegal purposes
- SQL injunction: targets traditional database and big database [NoSQL]. Inserting malicious statement into the input field of web application and big data components.
- Malware: an advance attack that uses multiple approaches to stealing organization data. these approaches are phishing emails and malware.
- Weak audit trail and misconfigured database
- Storage media Disclosure such as backup media needs for special protection.

III. CURRENT PROTECTION APPROACHES

Database security has the main concern of computer security or information security. Security Analyst uses different security controls, i.e. (physical, procedural and technical) to protect their organization data. Protecting databases on multiple hosts and securing information within the database are done with these controls. It's all required deeper research to protect the database from malicious activities. Researcher used different method and techniques such as Access control [4] [11] [12] [14] [16], Encryption technique [3] [13] [15], Audit Trail [19] [24] [27] mechanism for Database security purposes. The Summary of these methods and techniques are as follows.

A. User Identification

User identification means to verify any user or application identity who use information or data. User identification is based on password management system and password should keep secret all times. Password management system control through the user profile. Self-protection is a key component of self-managed systems capable of automatically defend against the malicious user, attacks at runtime. A self-protecting system or application proactively identifies users, malicious threats and triggers necessary actions to stop them [9] [46] [47]. Security professionals used different tools and skills such as (protection filters, detectors of suspicious activity, logging mechanism & backtracking tools) to protect their systems [1].

B. Access Control

Jabbour, et al. [4] presents Insider threat security architecture (ITSA), of self-protection in databases against insider threats. In this architecture privileged user compromised the database state where ITSA can protect. ITSA framework consists of security policy and defense mechanism managed by the super system owner. Security policy contains system parameter and their values while built-in logic is embedded in defense mechanism in the form of stored procedures and triggers and this logic is used to protect the system parameters. Three main components of ITSA are Autonomic Access Control Enforcement (AACE), Integrated Self-Protection Capability (ISPC), and Integrated Business Intelligence Capability (IBIC). The author discussed how the same scenario can be moderated under the Insider threat security architecture framework.

Jabbour, et al. [11] present notion based self-protection framework within the database by using the policy based approach. These policies are created by the system owner and block every attempt that compromises the Database state. Each action in the database is verified by the system owner before it applied to the database. Protection is achieved by implementing stored procedures, functions and triggers that have the built-in logic of checking insider user request. When an insider or attacker wants to change database security parameters, its request for changing parameters goes through a verification process through stored procedures before the following change can be applied to the database. If the change request truly verifies set of policies, then it can be applied to the database and its audit trail is maintained in the database. If the request is not verified from stored procedures, then change request is blocked and system owner is alerted through email and audit trail is maintained. Author present four types of policies, i.e. verifying and controlling user actions, monitoring database resources, changing security policy conditions and their parameters.

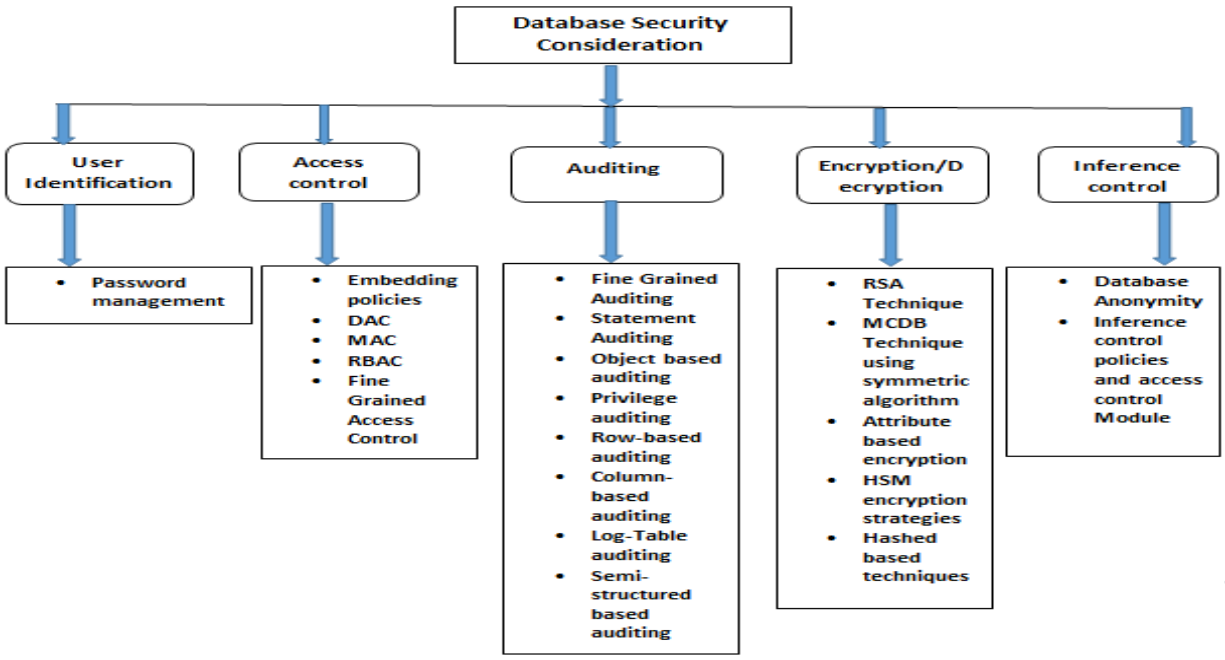


Fig. 1. Critical areas need to be considered in DBMS Security

Jabbour, et al. [12] addresses a protective framework for securing autonomic system policies. The author used two types of methodology in this framework. The first type is to partition security policies, blocks into numerous levels and then adding complexity to the entire architecture of the policies. This assists the purpose by adding alleged obscurity, which denies the potential attackers from decoding the policy's contents and directives. The second method is to insert false sense or false elements to different partitions of the policies (parameters and their values). Whose purpose is too confusing an attacker and giving a false sense of accomplishing his/her goal. K. Ahmed, et al. [14] addressed different types of a security layer, i.e. Database administrator (DBA), the System administrator (SA), Security officer (SO), Database developers and client or end user. These security layers are applied at almost all DBMS i.e. (Oracle, SQL Server, DB2, Teradata) environment. Theses security layers are responsible for implementing some well-defined security policies. The purpose of implementing these policies to ensure security features such as Confidentiality, integrity, efficiency, access control and privacy within the database.

A. Patil, et al. [16] presented Access control policy mechanism is used to secure a database against insider user. Three types of AC policies are mainly used, i.e. discretionary access control policy (DAC), Mandatory access control policy (MAC) and Role Base access control policy (RBAC). DAC based on the discretion of information creator or owner of the data. DAC used to restrict access of user on the basis of user identity and authentication. In MAC all users follow the same rule created by the Database administrator. RBAC used in a large organization where turnover rate of the employee is high. RBAC model built on the notion of role where role signifies a specific function within the organization. Each user performs a specific action which is granted to the specific role associated with it.

C. Auditing

Auditing is one of the important components in Database security infrastructure. In the database production environment in various database operations such as user login, Data manipulation language statements (DML), Data definition language statements (DDL) are needed to obtain an audit trail. Different methods and techniques are used by Researcher for auditing. The Database auditing purpose is to monitor and record user actions what he or she performs on the database.

Olumuyiwa O. Matthew et al. [24] discussed several already existing database auditing techniques such as statement auditing, privilege auditing, schema object auditing and fine-grained auditing etc. at various database environments. The author also discussed issues concerning about handling of audit trails against different database environment. According to author Database Auditing performs level by level. At first level logging (login and logoff) activities are a monitor, second level privileges check are an audit. In third level changes made to database schema are monitored, fourth level database DML activities are monitored and fifth level concerned with auditing changes made to a stored procedure, function and other codes. In next level database error is an audit and in the last level auditing any changes made to the definition of what is to be audited.

Li Yang [25] developed to extend auditing concept and technique by applying practical lab experience on security and auditing of a relational table that comprising an audit log of all commands and causes data changes on the target table. Some Common techniques of database auditing for monitoring database access control attempts, user login and logoff attempts, Data Control Language (DCL) activities, Data Definition Language (DDL) activities, and Data Manipulation Language (DML) activities. Erroneous queries should also be logged and monitored. Database auditing is implemented

through log files and audit tables. According to author security and auditing should be applied with integrated way.

Liu and Huang [26] present a framework of network-based database auditing that offers zero-impact of database performance. An agent is configured in passive mode to capture traffic flowing from the Database system and extract the audit log data which is beneficial for audit log analysis and then store this log information on another server. The author used Berkeley Packet Filter (BPF) filtering mechanisms to capture traffic and compare them against given conditions. They divided their methodology into three steps: packet filtering, the packet analyzing and data storage. Then alarm will be generated against any database anomaly or upon detection of malfunction of security regulation.

Narongrit Waraporn [27] suggested four methods implement database auditing for historical data. These methods are row-based auditing, column-based auditing, log-table auditing and semi-structure-based auditing. In row-based auditing, a separate audit table was created against each relational table. The operational table contains the last updated value while auditing table contains both static and historic data, two timestamps (start time and end time), operation type (update, delete, insert) and username. Row-based auditing caused data redundancy because the same record exists in two tables. To remove data redundancy column-based auditing is used. Column-based auditing does not contain the static data in auditing table. Column-based auditing caused null value in auditing table. The author suggests two approaches using log-table mechanisms. In the first approach extra table creates against each auditing column, while in second approach the only single audit table will be created against all operational tables. Semi-structure-based auditing also categories in two ways, i.e. Object-relational type, and XML type.

D. Fabbri et al. [31] proposed the idea of select triggers which are executed implicitly when a select query takes place on a specific object on which it is defined. Mostly none of the database management systems are implementing such a select trigger. Only Microsoft, however, is working on select query trigger and its researchers have presented their work earlier. Mostly triggers are based on the insert, update or delete commands, but the author also extends trigger in select command. It is also important to understand the action which is performed during trigger execution. The major issue of integrating select triggers in the DBMS is to handling a low overhead mechanism while ensuring the semantics are richly adequate to capture the modification of data access using SQL queries.

D. Encryption/ Decryption

In [3], Akinyele et al. present a flexible approach using attribute-based encryption (ABE) to generate self-protecting electronic medical records (EMRs), when health data is transferred on cloud servers or cell phones which are outside the trust boundaries of the healthcare organization. The EMR system ensures availability when the provider is offline. In this approach, the patient can encode each node of medical records in XML-based EMR file with attached access policy before it is transferred to the cloud storage. The Policy engine creates these access policies over electronic medical records on the basis of different user types (patient, physician, and insurance agent). Policy engines further define attribute sets, i.e. record type, patient age, and date to encode each record using attribute based encryption.

H. Kadhem, et al. [13] presented Mixed Cryptographic Database MCDB [13], a new data classification framework used to protect the databases by encrypting it in the semi-trusted scenario where data are shared among different parties using different keys. In this technique, database encryption is done over the unsecured network in an altered way that involves keeping numerous keys of different parties. In This scheme encryption is done at the client side, untrusted database and server side and it use symmetric key encryption mechanism. The purpose of keeping numerous keys by different authenticated parties that when the database is attacked by the attacker (insider or outsider) the database is not compromised. The performance of queries and security analysis is affected because of encryption Algorithms.

S. Sachdeva et al. [15] proposed negative database as extra security layers on generic databases. Negative data defined by some database security researchers as a database that contains a large amount of data consisting of bogus data and as well as real data. In this approach, author separated the information into two parts, i.e. sensitive information and non-sensitive information. Non-sensitive information directly stores in the Database while sensitive information first encrypted using RSA encryption algorithm and then convert the cipher text of sensitive information into base 32 shrink its length and then create large amount counterfeit. Now encrypted sensitive data along with counterfeit data stored in the database.

L. Bouganim et al. [18] suggest A new approach which embeds the security server inside the hardware security module (HSM). HSM is used to manage users, privileges, encryption policies and keys. HSM is responsible for all cryptographic operations and encryption keys are not exposed from this technique. Security server cannot modify or altered because it's fully embedded in the tamper-resistant Hardware Security module. The main limitation with this approach is that the Hardware Security Modules require a complex piece of software to be embedded in it. In this approach, encryption is done at the storage level, database level, and application level.

R. Jena et al. [28] proposed a cryptographic hash based function and digital timestamp technique to prevent from silently corrupting audit log files from both insider and outsider malicious user. Proposed technique will be implemented for the database system and trusted timestamp is efficiently used if logs are compromised or corrupted. The author implements their results in a high-performance engine. Audit log files comprise of log entries and each entry contains an element in a hash chain which authenticates the value of previous log entries. Two additional columns such as HashCode and Chain_ID and an additional table for digital timestamp is added. Chain_ID contain at most recent digital timestamp and it is generated by timestamp authority. Hash code based on previous values or data. If any audit log entry is tempered then database forensic analysis algorithm identifies

the tempering and regulate who, when, where and what components of audit log are tempered.

Kyriacos E. Pavlou et al. [29] developed a prototype DRAGOON to monitor the audit logs of the database and then detect malicious activities and perform forensic analysis against both insider and outsider users. They added some additional properties in DRAGOON to support information accountability in a cloud computing environment. The author used a cryptographic one-way hash function to protect silently corrupting audit log from an insider or an outsider or an unknown error in DBMS. Analyst used a series of algorithms which were designed for the forensic purpose to detect malicious activities. Extending some more features in DRAGOON architecture in database management systems increase scalability and it supports multiple databases and DBMSs. Extended DRAGOON architecture isolates four different areas of control. The first area is user application and GUIs controlled by the company itself. The second area is monitored by cloud provider where the monitored database resides (CLOUD A). The third area is monitored by cloud provider where DRAGOON resides (CLOUD B). The final area is END, which should not use cloud services. The extended DRAGOON architecture is scalable and customizable for providing a level of security and forensic analysis.

Kyriacos E. Pavlou et al. [30] highlighted the deep relations between time and the definition, Temper detection, forensic analysis of temper detection, and characterized different level of a database exploitation within the context of information accountability. Time in the context of applying information accountability and identifying time-security interactions. They categorized their audit system in three phases. The first phase is audit system execution phase second is their sub-phases and the third phase is an action performed during each phase. Transactions are hashed and cumulative associated with a cryptographically strong hash function in the first phase and the results of its digitally notarized with an external digital notarization service. In the second phase hash values are again extracted and matched from previously notarized. If the hash values are not matched from previously notarized, then these values are detected. The author introduces different forensic algorithms to detect when malicious activities occurred and what type of data has been corrupt.

E. Inference control

Inference control is a data mining technique used to attack databases where malicious user or attacker infers data from complex databases at a high level. The inference is used to find information hidden from common users. Popeea T et al. [32] presented multi-layer security to database anonymity and database security in a data warehouse which contains information of current and past employees of large companies. They mainly focus on securing communication channel, securing operating system and securing the database. They developed an engine based on java, which provides protection of both static and dynamic sensitive data. In this paper, an inference can be classified into six categories, i.e. splits queries, overlapping inferences, subsume inferences, complementary inferences, unique characteristic inferences and functional dependency inferences. To achieve high-level database security, they used mandatory access control layer, secure communication channel SSL, Ubuntu OS enhanced with MAC module and MYSQL as an open source DBMS.

Yang et al. [33] provide a secure inference control model by the trusted computing paradigm. This model entrusts the implementation of inference control to specific users' computer platforms. In this architecture, the database server is liable for the implementation of traditional access control, while the individual user's platform is allowed to handle inference control based on their own query logs in a decentralized manner. This architecture is used for complex and large databases. In traditional architecture, both access control and inference control are imposed at The Database server side. The Access control module (ACM) implements access control functionality, while the inference control module (ICM) performs a designated inference control algorithm. In the new architecture, inference control module resides at user side instead of server side. The user requests a query to inference control module (ICM), ICM transfer this query request to the access control module (ACM). ACM further check user requests against access rules and policies. If the user has granted access, then ACM return a response to ICM together with IC policy. Table II summarized the literature review with respect to protection in DBMSs.

TABLE. II. LITERATURE REVIEW SUMMARIZED

References	Research Contribution	Attack Type	Protection Techniques Used	Limitation
[11]	Policies are enforced for securing database configuration from inside user or DBA	Insider Attack	Embedding policies in DBMS	Policies based on the notion. Database configuration specific policies. No confidentiality provided as DBA can view data.
[12]	Protect security policies of autonomic system	Insider attack	Partitioning and giving a false sense by adding false elements	
[13]	Encryption of database over untrusted networks, data classification is based on data ownership data is confidential if one key compromise	Both for Insider and Outsider attacks	MCDB technique using any symmetric algorithm	Performance of queries and security analysis is affected because of encryption Algorithms.
[4]	ITSA based on security policies and defense mechanism.	Insider attacks	Security policy and defense mechanism	Works only Autonomic Access Control Enforcement, Integrated Self-

	Security policies consist of database parameter and their values Defense mechanism comprises logic encoded in a set of stored procedures.			Protection Capability, and Integrated Business Intelligence Capability
[14]	Define database security layer Each layer has some specified policies for authentication and authorization purpose	Used for both insider and outsider attack	Policies based	
[15]	Proposed extra security layer through negative database techniques Entity, attribute, and value (EAV) model is used	Insider attack	Negative DB and RSA encryption technique	More complex Taking more query execution time More costly and variable k value is fixed
[16]	Review key access control models	Both for Insider and Outsider attack	DAC, MAC, RBAC	
[17]	Protection of real world health databases to restrict access to data from internal user or outsider	Both for Outsider and Insider attack	RSA Technique	Provide application based Database security Not for generic database Security.
[3]	Implementing autonomic property to protect electronic medical records (EMRs) using attribute-based encryption scheme (ABE).	Both for inside and outside attacks	Attributes based encryption access control using RBAC and content based	Encryption and decryption time based on a number of attributes in access policies.
[18]	Review different encryption level, techniques, and methods Key management and their issues.	Both Insider and outsider attacks	HSM Encryption Strategy for key management.	HSM now requires a complicated piece of software to be embedded in it
[19]	DBMS-Layer is a most appropriate layer to protect against insider for exfiltration detection. Virtualization techniques are used to tackle provenance.	Insider attack	Role based access Profiles and threshold Provenance Embedding and virtualization Techniques	Modeling and Specification of Lineage Information Authentication and Authorization Systems and network issues
[20]	Discuss database security threats against both internal and external threats. Proposed SIIMDS to detect both internal and external attacks.	Both internal and external attacks	SQL Injection and Insider Misuse Detection System (SIIMDS)	
[21]	A large number of abnormal queries are running in same specific time caused query-flood attacks. Degrade database performance.	Insider attack	Attack detection algorithms	DB performance slow
[22]	Review all requirements of access control for the context of scalability, granularity, and situation-aware decisions.	Insider attack	RBAC approach Fine Grained Access Control	Implementation is not done.
[23]	Each activity of users is modeled on the basis of SQL commands running and data generated by that user.	Insider Attack	DATA-CENTRIC APPROACH	Performance consideration
[24]	Outlines main auditing techniques and methods Issues relating to handling of audit trail are also discussed and key important impacts of security are also highlighted	Both internal and external attacks	Auditing methods such as FGA, Statement auditing, Privilege auditing, schema object auditing	Discussed already existing auditing technique
[25]	To engage students actively, practical labs are developed to assimilate theories of database security and auditing Use of two major database products (Microsoft SQL Server and Oracle 10g	Paper used for Database Security purpose	Monitoring database access attempts, DCL activities, DDL activities, and DML activities	Some issues regarding terminology and capabilities of DB are not completely discussed in a hands-on lab.
[26]	Monitor network flowing into and of DB system and generate log information about DB Execute audit analysis through event correlation Generate alarm in case of any violation detected	Both internal and external attacks	Agent-based network monitoring Used Berkeley Packet Filter (BPF) filtering to scan packets	Network-based logging has its limitation too if DB has been encrypted, then passive packets capturing method will be invalid
[27]	Discuss different four methods to achieve database auditing. Discuss multiple audit log columns, tables for transaction logs Single audit table for transaction logs	Paper used for Database Security purpose	Row-based auditing Column-based auditing Log-Table auditing Semi-structured based auditing	Row-based auditing caused data redundancy In column-based auditing, null values in the table would lead to problem
[28]	Auditing data integrity themselves is a very serious concern Malicious activities are performed both by authorized user and as well as unauthorized user	Both internal and external attacks	Cryptographic Hash-based technique used for forensic analysis Trusted Timestamping used to prevent the log files from both internal and external	Implement in the only online transactional database. Does not produce tamper resistant audit log

			users	
[29]	Developed a prototype called DRAGOON for information accountability periodically audits database, detect malicious activities and then perform forensic analysis DRAGOON support non-cloud databases Deploy how existing prototype extending within the cloud	Both internal and external attacks	DRAGOON use cryptographic hashing technique	Concurrency issue raised when transaction data is replicated at application level Hashing occur at the application level is open design issue
[30]	Notarization and validation of database exploit the temporal semantics of transaction times database.	Insider attacks	Monochromatic forensic algorithms	
[31]	Triggers are useful to track and log any changes made on data by executing any DML commands Trigger assists row-level auditing of both DML and DDL commands Select trigger fires when a select operation takes place on the object	Internal attacks	Select trigger techniques	Some scenario's large number of false positive occur
[32]	Inference detection is done here with SSL communication channel The Re-identification algorithm is an implementation of k-anonymity	Both internal and external attacks	Split queries	Data anonymity is not fully completed

IV. Proposed Autonomic Model W.R.T Self-Protection

In this proposed model firstly we will explain how an adversary or the attacker can perform what types of malicious actions to compromise database state. In our survey adversary can be internal users or database administrator. He or she can perform the following actions to attack the database. The Attacker can change some configuration parameters of database management system that change the state of the database in a way that Database performance is slow or it's not obvious to its end user. For example, in Oracle database, database administrator changes various system configuration parameters such as disable auditing parameter or run the database in NOARCHIEVE log mode or changes some other security parameter that compromises the database health or its behavior.

In some organization, DBA with full access to the database can run any DML (update, delete or insert) or DDL (create, alter, drop, truncate) commands on sensitive data or information to change it. The DBA can also drop any database or drop any schema in the database. Audit trail used for forensic analysis, provide documentary proof of the sequence of actions that have affected at any time a specific operation. The attacker also Change or remove Audit trail information in the database. If a user is granted database privileges that exceed their job role and requirements, then those privileges can be abused or misused.

Fig 2 shows proposed autonomic model against insider threats in DBMS with respect to self-protection diagram. In this proposed model, we mainly discussed self-protection property in database management system against insider threats. Self-protection against insider threats in DBMS, previously proposed techniques is based on embedding security policies for enforcing database security configuration parameters. In our proposed architecture, we imposed CIA (Confidentiality, Integrity, and Availability) security model for building policies against these three properties.

In this model super user, build security policies for database security. These security policies are related to Access control, Database configuration parameter setting policies, password management policies and encryption policies, etc. When an insider user or DBA attempts to change in the DBMS through SQL command Prompt, the request goes for verification phase. If the request verifies set of policies, then the request will be applied in DBMS and audit trail record will also be saved in a log table, else insider request will be rejected, alert through an email is generated to the super user, notify the insider user and audit trail will be recorded. For monitoring malicious activities against internal threats we used Alert mechanisms. When any malicious activities found alert will be generated.

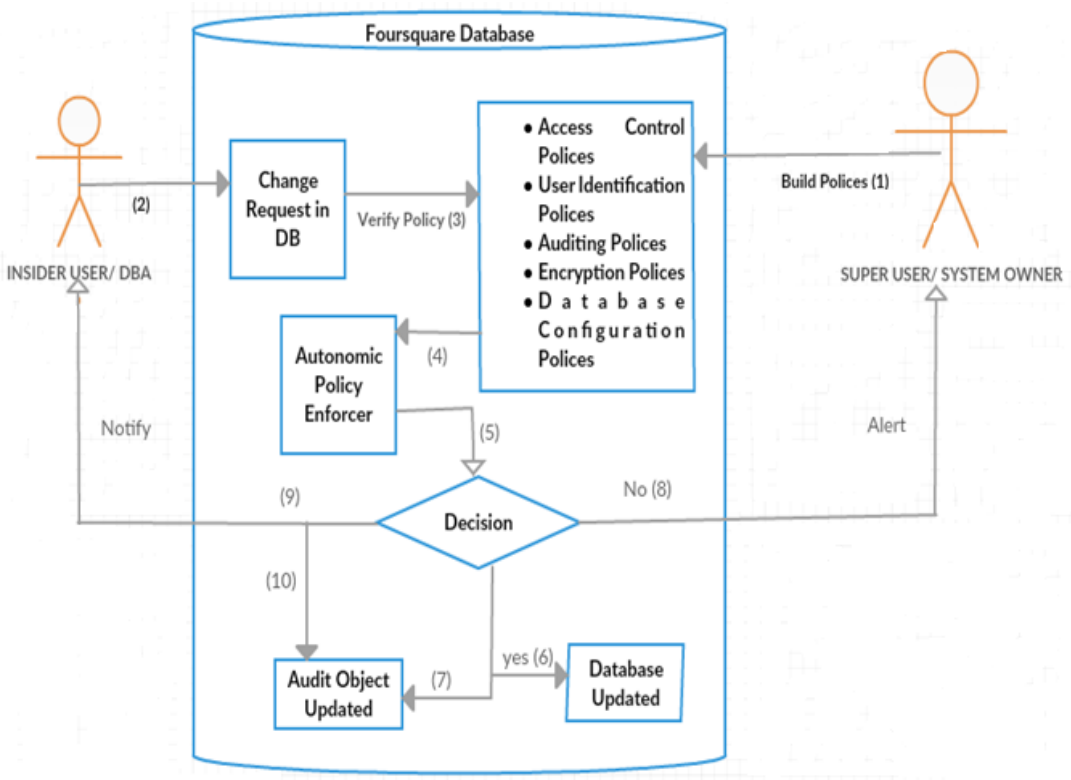


Fig. 2. Autonomic Model against Insider Threats in DBMS with respect to Self-Protection

In data confidentiality insider user or DBA has full access to view sensitive information and non-sensitive information. In our model super user-segregated information into two ways, i.e. sensitive information is non-sensitive information. Sensitive information stored in encrypted form in the database after applying the encryption function and the non-sensitive information is stored as it is in the database. The only superuser can encrypt or decrypt sensitive information. Our purpose is to make sensitive information more confidential from inside user. We evaluated our proposed Architecture with an already existing architecture based on the following criteria:

- Set of Polices is verified using a set of queries.
- Improved Autonomous property of self-protection. (Autonomic Improving Capability)
- Generation of alerts at the time of any attack in DB.

We are expecting our enhanced model provides more secure protection against insider threats in database management systems.

V. Conclusion and Future Work

Data is probably most valuable property on which entire organization depends. Database security is one of the main concerns of the researchers nowadays. This paper addressed the security threats against database management systems and how to mitigate these threats by using autonomic computing properties. The research emphasized some critical areas such as access control, encryption, auditing, accountability and inference control that need to be considered in database security. This study identified, how malicious user exploits these areas and gain DBA level access to the database and causes a denial of service attack. An autonomic model is proposed which protects data against insider threats. A number of security techniques and policies are addressed that should be used in database management system to achieve protection against the insider threats. The premise of our proposed model is to highly enforce the concept of separation of duties in an organization and also brings security. We adapted the concept of building system level policies in such a way that meets the autonomous self-protecting capabilities to defeat privileged insider users and unintentional actions. Organizations owner or super-user builds policies for database security against critical areas. The alerts can also be generated through an email against malicious activities of insider user.

As for future research, we will implement and demonstrate all above mention policies in database management system. We plan to implement access control policies at the database connection level, DML and DDL command level to achieve self-protection in DBMS. Similarly, we will implement database configuration level policies, encryption level policies and auditing level policies, etc. We believe that it would be valid and beneficial attempts to apply and demonstrate these different levels of policies in the DBMS environment to achieve security.

References

[1] Depalma, N.; Claudel, B.; Lachaize, R, "Self-Protected System: an experiment," In 5th Conference on Security in Network Architectures (SAR). Addison Wesley, Longman, England, 2006.

[2] B. Raza, A. Mateen, M. Sher, M. M. Awais, and T. Hussain, "Autonomicity in Oracle Database Management System," 2010 Int. Conf. Data Storage Data Eng., pp. 296–300, Feb. 2010.

[3] A. Akinyele, C. U. Lehmann, M. D. Green, M. W. Pagano, Z. N. J. Peterson, and A. D. Rubin, "Self-protecting electronic medical records using attribute-based encryption," Cryptology ePrint Archive, Report 2010/565, 2010.

[4] Jabbour, G. G., & Menasce, D. A, "The Insider Threat Security Architecture: A Framework for an Integrated, Inseparable, and Uninterrupted Self-Protection Mechanism," 2009 Int. Conf. Comput. Sci. Eng., pp. 244–251, 2009.

[5] L. Basharat, F. Anam and A. Wahab Muzaffar, "Database Security and Encryption; A Survey Study", International Journal of Computer Application, vol. 47, (2012), pp. 28-34.

[6] Cert, "cert insider," 2007. [Online]. Available: http://www.cert.org/insider-threat/research/database.cfm?

[7] P. Horn, "a u t o n o m I c o m p u t i n g : the information technology industry loves to prove the impossible possible. We obliterate barriers and set records with astonishing regularity. But now we face a problem springing from the very core of ou," 2011.

[8] M. R. Nami and K. Bertels, "A survey of autonomic computing systems," in ICAS '07: Proc. Third International Conference on Autonomic and Autonomous Systems. Washington, DC, USA: IEEE Computer Society, 2007, p. 26.

[9] M. C. Huebscher and J. A. McCann, "A survey of autonomic computing - degrees, models, and applications," ACM Comput. Surv., vol. 40, no. 3, 2008.

[10] P. A. Pe, "Top Ten Database Security Threats The Most Significant Risks of 2015 and How to Mitigate Them Red Flag."

[11] Jabbour, G. G., & Menasce, D. A, "Policy-Based Enforcement of Database Security Configuration through Autonomic Capabilities," Fourth Int. Conf. Auton. Auton. Syst., pp. 188–197, Mar. 2008.

[12] I. Technology and C. Science, "Securing Security Policies in Autonomic Computing Systems," 2008.

[13] H. Kadhem, T. Amagasa, and H. Kitagawa, "A Novel Framework for Database Security Based on Mixed Cryptography," 2009 Fourth Int. Conf. Internet Web Appl. Serv., pp. 163–170, 2009.

[14] K.Ahmad, "Policy Levels Concerning Database Security," no. Feb 2016.

[15] S. Sachdeva, "Implementing Security Technique on Generic Database," 2015.

[16] A. Patil and P. B. B. Meshram, "Database Access Control Policies," Applications (IJERA) vol. 2, no. 3, pp. 3150–3154, 2012.

[17] N. Batra and Pooja, "Secure Mechanism for Medical Database Using RSA," IJAIEM vol. 3, no. 7, pp. 320–327, 2014.

[18] L. Bouganim and Y. Guo, "Database encryption," in Encyclopedia of Cryptography and Security, Springer, 2010, 2nd Edition.

[19] Bertino and G. Ghinita "Towards mechanisms for detection and prevention of data exfiltration by insiders." In Proc. 6th ACM Symp. on Information, Computer, and Communications Security. pages 10–19, 2011.

[20] Asmawi A, Sidek ZM, Razak SA. " System architecture for SQL injection and insider misuse detection system for DBMS," in International Symposium on Information Technology (ITSim'2008), 2008, pp. 1 -6.

[21] A. C. Squicciarini, I. Paloscia, and E. Bertino, "Protecting databases from query flood attacks," in ICDE, 2008, pp. 1358–1360.

[22] Park, J. S. & J. Giordano, "Access Control Requirements for Preventing Insider Threats," Proc. ISI'06 LNCS 3975, pp. 529–534, Springer, 2006.

[23] S. Mathew, M. Petropoulos, H. Q. Ngo, and S. Upadhyaya, "Data-Centric Approach to Insider Attack Detection in Database Systems," Recent Advances in Intrusion Detection, 2010.

[24] O. O. Mathew and C. Dudley. "Critical Assessment of Auditing Contributions to Effective and Efficient Security in Database Systems." Int Conf on CSITA, At Royal Orchid Central Bangalore, India pp. 1-11, March, 2015.

[25] Yang, L., "Teaching Database Security and Auditing," Proceedings of the 40th ACM Technical Symposium on Computer Science Education (SIGCSE), Chattanooga TN, March, 2009.

[26] Liu, L. and Huang, Q. "A Framework for Database Auditing". Computer Sciences and Convergence Information Technology, 2009.

[27] Waraporn, N. "Database Auditing Design on Historical Data." In Proceedings of the Second International Symposium on Networking and Network Security (ISNNS '10). Jinggangshan, China, April. 2010, pp. 275-281

[28] Jena, R., Aparna, M., Sahu, C., Ranjan, R. and Atmakuri. "Ensuring Audit Log Accountability through Hash Based Techniques." International Journal of Future Computer and Communication 1.4, Dec. 2012.

[29] Kyriacos E. Pavlou and Richard T. Snodgrass, "Achieving Database Information Accountability in the Cloud" Tucson, AZ 85721–0077, USA, 2002.

[30] Pavlou, Kyriacos E., and Richard T. Snodgrass. "Temporal implications of database information accountability." 2012 19th International Symposium on Temporal Representation and Reasoning. IEEE, 2012.

[31] Fabbri, D., Ramamurthy, R. & Kaushik, R. "SELECT triggers for data auditing." Proceedings of the 29th International Conference on Data Engineering (ICDE). IEEE:1141-1152, 2013.

[32] T. Popeea, A. Constantinescu, L. Gheorghe and N. Ţăpuş, "Inference Detection and Database Security for a Business Environment.", International Conference on Intelligent Networking and Collaborative Systems, (2012), pp. 612-617.

[33] Yang, Y., Li, Y., and Deng, R.H., "New paradigm of inference control with trusted computing.", in Proceedings of the 21st annual IFIP WG 11.3 working conference on Data and applications security, Redondo Beach, CA, USA. 2007, pp. 243-258.

[34] Martin S. Olivier. "Database privacy: balancing confidentiality, integrity and availability." SIGKDD Explor. Newsl., 4(2):20–27, 2002.

[35] Von Solms R, Van Niekerk J. "From information security to cyber security." Computers & Security. 2013 Oct 31;38:97-102.

[36] Safa, N. S., Von Solms, R., & Furnell, S. "Information security policy compliance model in organizations." computers & security, 56, 70-82, (2016).

[37] Chen D, Zhao H. "Data security and privacy protection issues in cloud computing." InComputer Science and Electronics Engineering (ICCSEE), 2012 International Conference on 2012 Mar 23 (Vol. 1, pp. 647-651). IEEE.

[38] Tianfield, Huaglory. "Security issues in cloud computing." In Systems, Man, and Cybernetics (SMC), 2012 IEEE International Conference on, pp. 1082-1089. IEEE, 2012.

[39] Buyya R, Calheiros RN, Li X. "Autonomic cloud computing: Open challenges and architectural elements." InEmerging Applications of Information Technology (EAIT), 2012 Third International Conference on 2012 Nov 30 (pp. 3-10). IEEE.

[40] Patel A, Taghavi M, Bakhtiyari K, JúNior JC. "An intrusion detection and prevention system in cloud computing: A systematic review." Journal of network and computer applications. 2013 Jan 31;36(1):25-41.

[41] Maggio M, Hoffmann H, Papadopoulos AV, Panerati J, Santambrogio MD, Agarwal A, Leva A. "Comparison of decision-making strategies for self-optimization in autonomic computing systems." ACM Transactions on Autonomous and Adaptive Systems (TAAS). 2012 Dec 1;7(4):36.

[42] De Lemos, Rogério, Holger Giese, Hausi A. Müller, Mary Shaw, Jesper Andersson, Marin Litoiu, Bradley Schmerl et al. "Software engineering for self-adaptive systems: A second research roadmap." In Software Engineering for Self-Adaptive Systems II, pp. 1-32. Springer Berlin Heidelberg, 2013.

[43] Frei, Regina, Richard McWilliam, Benjamin Derrick, Alan Purvis, Asutosh Tiwari, and Giovanna Di Marzo Serugendo. "Self-healing and self-repairing technologies." The International Journal of Advanced Manufacturing Technology 69, no. 5-8 (2013): 1033-1061.

[44] Eze, Thaddeus, Richard Anthony, Chris Walshaw, and Alan Soper. "Autonomic computing in the first decade: trends and direction." In Proceedings of the Eighth International Conference on Autonomic and Autonomous Systems (ICAS). 2012.

[45] Ferreira da Silva, Rafael, Tristan Glatard, and Frédéric Desprez. "Self-healing of operational workflow incidents on distributed computing infrastructures." In Proceedings of the 2012 12th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (ccgrid 2012), pp. 318-325. IEEE Computer Society, 2012.

[46] Chen, Qian, Sherif Abdelwahed, and Abdelkarim Erradi. "A model-based approach to self-protection in computing system." In Proceedings of the 2013 ACM Cloud and Autonomic Computing Conference, p. 16. ACM, 2013.

[47] De Palma, Noel, Daniel Hagimont, Fabienne Boyer, and Laurent Broto. "Self-protection in a clustered distributed system." IEEE Transactions on Parallel and Distributed Systems 23, no. 2 (2012): 330-336.

[48] Ayala, Inmaculada, Mercedes Amor, and Lidia Fuentes. "Self-configuring agents for ambient assisted living applications." Personal and Ubiquitous Computing 17, no. 6 (2013): 1159-1169.

A Low Complexity based Edge Color Matching Algorithm for Regular Bipartite Multigraph

Rezaul Karim
Dept. of Computer Science & Engineering
University of Chittagong (CU)
Chittagong, Bangladesh

Md. Rashedul Islam
Dept. of Computer Science & Engineering
International Islamic University Chittagong (IIUC)
Chittagong, Bangladesh

Muhammad Mahbub Hasan Rony
Dept. of Computer Science & Engineering
International Islamic University Chittagong (IIUC)
Chittagong, Bangladesh

Md. Khaliluzzaman*
Dept. of Computer Science & Engineering
International Islamic University Chittagong (IIUC)
Chittagong, Bangladesh

***Abstract*—An edge coloring of a graph G is a process of assigning colors to the adjacent edges so that the adjacent edges represents the different colors. In this paper, an algorithm is proposed to find the perfect color matching of the regular bipartite multigraph with low time complexity. For that, the proposed algorithm is divided into two procedures. In the first procedure, the possible circuits and bad edges are extracted from the regular bipartite graph. In the second procedure, the bad edges are rearranged to obtain the perfect color matching. The depth first search (DFS) algorithm is used in this paper for traversing the bipartite vertices to find the closed path, open path, incomplete components, and bad edges. By the proposed algorithm, the proper edge coloring of D – regular bipartite multi-graph can be obtained in *O (D.V)* time.**

***Keywords*—matching; edge-coloring; complexity; bipartite multigraph; DFS**

I. INTRODUCTION

An edge coloring of a Graph is one of the well-known, exoteric researched topics in the arena of graph theory. Edge coloring of a graph G is used various colors, so that, the adjacent edges are obtained different colors. By using this concept of edge coloring many real-world problems can be solved. An edge coloring has applications in scheduling problems and in frequency assignment for fiber optic networks. It also used to solve the timetabling problem, register allocation, pattern matching, designing seating plans, solving Sudoku puzzles and so on.

This section provides a descriptive summary of some methods that have been implemented and tested at graph theory for solving edge coloring problems. This topic has gained importance for the purpose of efficient edge color matching in the different graphs. For example, in [1], proposed a method for edge coloring in which every $(3, \Delta)$-bipartite graph G, chromatic index $\leq 4\Delta$. This paper only considered the $(3, \Delta)$-bipartite simple graph. In [2], proposed an edge coloring method for course timetabling. One-sided interval colorings of a bipartite graph method are introduced in [3]. For any graph G with bipartite set (X, Y) where authors present upper $x'_{int}(G, X)$ for classes of bipartite graphs G with maximum degree $\Delta\ (G)$ at most 9. In particular, if $\Delta\ (G) = 4$, then $x'_{int}\ (G, X) \leq 6$ and so on. In [4], authors derived a theorem to find the closed paths from $C = M \cup N$ for matching M and N. A closed path C can be found in $O(|C|)$ time on average. This theorem helped to develop the proposed algorithm for minimal edge-coloring. This concept of open and closed path can be easily obtained from this theorem.

In [5], showed a theorem in which any edge color matching of a complete bipartite graph $K_{n,n}$ contains 18 vertexes with three colors. This method creates disjoint monochromatic cycles which together cover all vertices. The minimum number of cycles is required for this type of covering is 5. In [6], proposed an algorithm to find out two disjoint matching M_1 and M_2 for a given (X, Y) bipartite graph with set $S \subseteq X$, where, M_1 saturates X and M_2 saturates S. The problem was solved by finding and appropriate factor of the graph when $|S| \geq |X|-1$. In [7], proved a method for two bipartite graphs G and H, where, H is a fixed graph whose vertices will be shown as colors. And H-coloring of a graph G is a process of assigning colors for preserving adjacency in graph G.

In this paper, an algorithm is proposed that is developed to find a perfect color matching of a regular bipartite multigraph. This is done by dealing edge coloring with lower time complexity. For that, the proposed method is divided into some parts that are run with an independent time complexity and helps to reduce the overall time complexity.

The edge coloring of a bipartite multigraph is highly related to finding a perfect matching efficiently. To obtain the perfect edge color matching the proposed method is divided into two parts. In the first part, the Depth First Search (DFS) algorithm is used to extract the closed and opened path circuits as well as bad edges. In the second part, the bad edges are rearranged to find the perfect edge color matching.

The rest of the paper is organized as follows. The preliminaries of the graph theory are described in Section II. The proposed minimal edge color matching algorithm is introduced in the next section. Case studies are described in Section IV. Experimental results and discussions are explained in Section V. The papers are concluded in Section VI.

II. PRELIMINARIES

A Graph G is an ordered pair $G = (V, E)$ along with a set V of vertices, nodes or points simultaneously with a set E of edges, which are two elements subsets of V. On the other hand, Graph coloring problem is a process of marking out colors from definitive components of a graph subject to certain obligations. There are two terms of Graph coloring which are edge coloring and vertex coloring. In this paper, an algorithm is developed for edge coloring. A Graph is said to be regular when every vertex has the same degree. Where, a graph is said to be bipartite which has vertices and also can be separated into two disuniting sets (U, V). Multigraph is a graph which has multiple edges and all the edges have similar finishing nodes. The edges associate with a vertex inside U to another one V is termed as a closed path. On the other hand, a path in which the first and last vertices are distinct is thread as an open path. Furthermore, a matching in a graph G is a set of pairwise disjoint edges. The edge coloring of bipartite multigraph is highly related to find a perfect matching efficiently.

III. ALGORITHM FOR MINIMAL EDGE COLORING

Assume that, G (V, E) is a regular graph. To achieve this regular graph every vertex added the edges such that every vertex has degree exactly D. In the whole description the graph circuits are considered as a closed path, whereas, the set of paths may consist of opened paths and/or closed paths. So, when a path sets are considered then there may be consisting of opened paths or closed paths.

The method of Alexander Schriver's [4] is applied in this paper. In that paper, the authors derived a theorem to find the closed paths of $C=M \cup N$ for matching M and N and a closed path C can be found in $O|(C)|$ time on average. Accordingly, the regular matching from the graph is computed to extract the close and open paths. The matching is colored and removed from the graph. This process is applied recursively to extract all matching.

Let, C_1, C_2, ... are closed paths that can be constructed from the graph G. Let, VC_1, VC_2, ... are the set of vertices of the closed path C_1, C_2, ... respectively. In this paper some closed paths are found i.e., C_1, C_2, ... such that $Vc_1 \cup Vc_2 \cup .... \subseteq V$ *and* $Vc_1 \cap Vc_2 .... = O$. This means more than one closed path from a graph G can be received, where the vertices are disjoint. These received closed paths are produce two full matching. However, it is not possible to get closed paths that cover entire vertex in G. If it is possible to add two full matching of G then it may possible to obtain one or more closed path. So, after receiving $D/2$ closed path sets another set of edges are found that holds at most V edges. Note that, at most V number of edges in closed path set can be found. And to achieve these it may needs to find the closed path for several times.

Theorem 1: A perfect matching in a regular bipartite graph can be found in $O(V)$ time.

Proof: To prove this theorem, in this paper the Alexander Schriver's [4] is considered to get the perfect matching. It is proved from [4] is that it takes $O|(C)|$ time to get a closed path and edges in the closed path are equally decomposed into two matchings. If the average open path size is L then it will take O(L) times to get the opened paths. According to Alexander Schriver's, the union of the two matching construct at least one closed path. It is also possible to get more than one disjoint closed paths and/or open paths by the union of two full matching.

In this proposed method, one or more disjoint paths are taken such that each vertex reduces its degree by maximum two. This process will be continued by using DFS. After that, checking each closed path where the number of vertices must be equal to the number of edges. If this condition is not satisfied then start the DFS again to find another closed path. After getting the entire closed path, the process is terminated and starts to get the open paths. In the normal situations all the vertices in G within the closed path set couldn't be found. However, using this theorem the maximum number of closed path can be found. There are many vertices that are not visited in the graph. For that, find one or more opened path associated with these non-visited vertices and decomposes them into two matchings.

In this method, the maximum $V/2$ number of closed path or $V/2$ number of opened path can be found in each time and maximum V number odd edges can be found in the paths. Every closed or opened path will possibly be minimized by its length. Then the closed and opened path's edges are alternatively distributed into two matching. While finding the paths in this method is required $V+V/2 = O(V)$ time. Hence, $O(V)$ time is needed to find a perfect matching.

Theorem 2: A proper edge coloring of a D-regular bipartite multigraph can be found in $O(D.V)$ time.

Proof: According to Theorem 1 all closed path set can be computed by $D/2$ times and found D number of matching. However, all time these types of matching are not possible by this theorem. For this reason, two different processes are required for finding the D number of matching. In process 1, the possible closed and opened paths are extracted from the graph. And the incomplete component can be identified as bad edges. In process 2, the bad edges are rearranged at incomplete matching to find the D number of matching.

Procedure 1: Extraction of possible paths

Step1: Taking a node from graph for finding closed path

Fig. 1(a) is the processing example to describe the steps of the theorem. For finding the closed path from the graph, the searching is started from vertex 1 in Fig. 1(a).

Step 2: From the starting node a closed path using DFS is found and is removed it from the graph

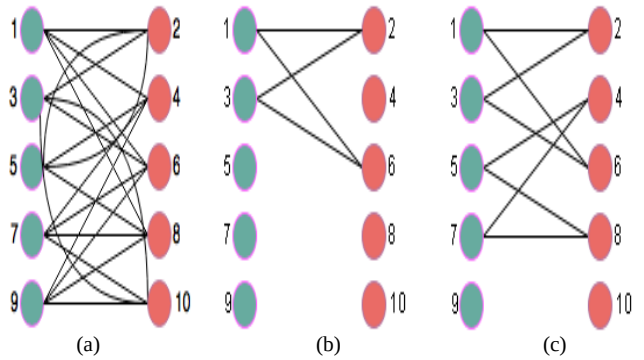


Fig. 1. Processing example: a) 4-regular bipartite multi-graph, b) first closed path with vertex 1, 2, 3, 6 c) all closed path from Fig. 1(a)

As the graph which contains the 10 vertices with degree 4. The closed and opened paths are obtained by reducing each vertex degree by two. According to step 2, from the starting node 1 the first closed path is found with the vertex 1, 2, 3, 6, 1 is shown in Fig. 1(b). By traversing the graph using the DFS the closed paths are obtained that are shown in Fig. 1(c).

Step 3: The edges in the closed path are distributed into two matching

The closed path in the Fig. 1(b) is distributed in two path set shown in Fig. 2(a) and Fig. 2(b). The closed paths that are in the Fig. 1(c) are distributed into two paths shown in Fig. 2(c) and Fig. 2(d).

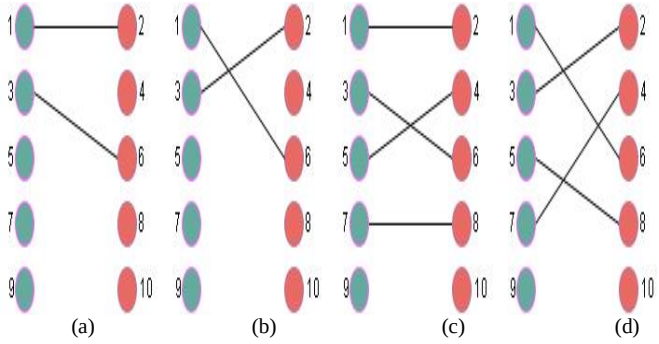


Fig. 2. The processing example of edge distribution for matching: a) and b) matching from Fig. 1(b), c) and b) matching form Fig. 1(c)

Step 4: If all the vertices are not traversed then repeat the process from Step 1 to Step 3 for non-visited vertices

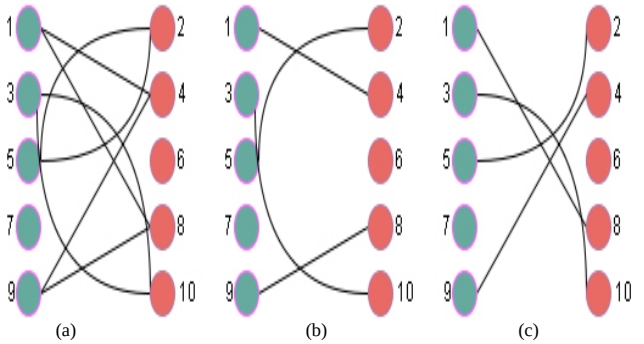


Fig. 3. Processing example of closed path extraction by repeated process: a) the closed path by repeated process, b) and c) matching from Fig. 3(a)

Since the vertices 9 and 10 in Fig. 1(a) are not visited by the DFS, Step 1 to step 3 have to be repeated. This repeating procedure traces the other closed paths that are not extracted by the previous steps shown in Fig. 3(a). The closed path in the Fig. 3(a) is distributed in two path sets that are shown in Fig. 3(b) and Fig. 3(c).

Step 5: If the remaining vertices do not make closed paths then start to find the opened paths

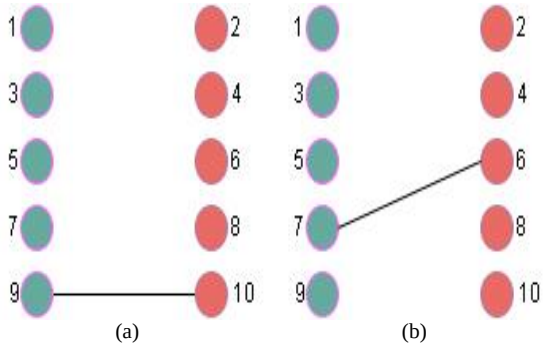


Fig. 4. The processing example of finding the opened path: a) and b) opened a paths from Fig. 1(a)

The opened path can be obtained from the vertices that are not used in the closed path. According to the Fig. 1(c), vertex 9 and 10 are not used to create any closed path. From Fig. 3(a), vertex 7 and 6 are not used to create any closed path. Those vertices are not connected by any edge. However, according to Fig. 1(a), vertex 9 and 10 as well as vertex 7 and 6 are connected by two separated edges. These separate edges are considered as opened path that are shown in Fig. 4(a) and Fig. 4(b).

Step 6: The edges of an opened path will be distributed in the matching (matching found in step 3).

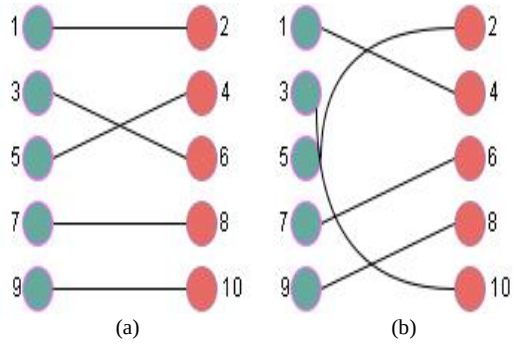


Fig. 5. The processing example of distributing opened paths into the matching: a) and b) open path in Fig. 4(a) and Fig. 4(b) are distributed into matching in Fig. 2(c) and Fig. 2(d)

The opened path shows in Fig. 4(a) and Fig. 4(b) are distributed in any two incomplete matching from Fig. 2(c), Fig. 2(d) as well as Fig. 3(b), and Fig. 3(c). In this case, the opened path in Fig. 4(a) and Fig. 4(b) are distributed in incomplete matching in Fig. 2(c) and Fig. 3(c) respectively. The remaining incomplete matching will be reconstructed in the procedure two.

Step 7: While all the vertices are traversed then switch for the next closed path set.

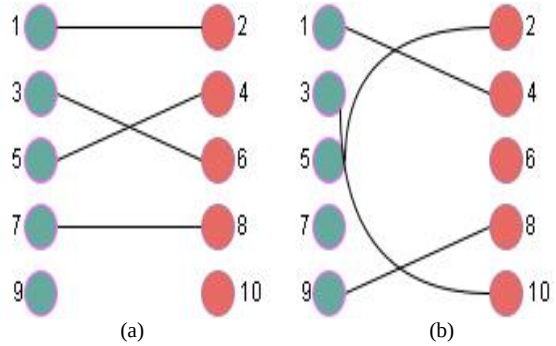


Fig. 6. The processing example of extracting components: a) components from the Fig. 2(d), b) components from the Fig. 3(b)

Step 8: Identify the component's (the vertices that have no edges) of an incomplete matching in a set

Since, the Fig. 2(d) and Fig. 3(b) are contained components, i.e., vertex 9, 10 and vertex 7, 6. That's why; Fig. 2(d) and Fig. 3(b) are considered as incomplete matching.

After *D/2* iteration, the process terminates, and gets *D/2* numbers of incomplete matching and *D/2* number of complete matchings. According to the processing example, each vertex has a degree of 4. The iteration process is performed in this processing example is two which is 4/2=2.

The algorithm in procedure 1 runs in *O (D.V)* time. This is obtained by multiplication of the number of paths being found and the total number of edges in the paths by D/2 iteration, i.e., *D/2*(V+ (V/2)) =3DV/4= O (D.V).*

After completing the procedure 1, there may exist many edges in the graph G that are not used while creating the closed and opened path are considered as bad edges.

Procedure 2: Finding the bad edges and rearranging them for perfect matching

Step 1: The remaining bad edges need to be re-distributed into an incomplete matching i.e., Fig. 2(d) and Fig. 3(b). According to the Fig. 1(a) only two edges are not used while creating the closed path that is edge(7, 10) and edge (9, 6) shown in Fig. 7(a). The bad edges are re-distributed into incomplete matching of Fig. 2(d).

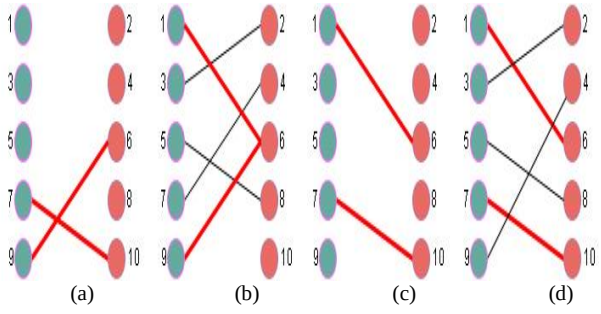


Fig. 7. The processing example of the bad edges: a) bad edges from Fig. 1(a), b) adjacent
edges after inserting the bad edge in Fig. 2(d), c) final bad edge set, d) final distribution of the bad edges

To re-distribute the bad edges of Fig. 7(a) into an incomplete matching i.e., Fig. 2(d), firstly traverse the incomplete matching by DFS to find the component. Here, the component is vertex 9 in Fig. 2(d). After that, remove the bad edge that is connected to vertex 9 from the bad edge set. This bad edge insertion may cause two adjacent edges in the matching as shown in Fig. 7(b). In this case, the bad edge that is connected to the vertex 9 is removed from the bad edge set and inserts into the other adjacent edge in the bad edge set i.e., edge (1, 6) as shown in Fig. 7(c). Finally, insert all the edges from Fig. 2(d) into Fig. 7(c) except the board edges shown in Fig. 7(d). Similarly, the bad edges are also re-distributed into incomplete matching of Fig. 3(b) is shown in Fig. 8. According to procedure one, two complete matching are found that are shown in Fig. 9(a) and Fig. 9(b) with different colors. By procedure two, another two complete matching are found that are shown in Fig. 10(a) and Fig. 10(b) with other different colors.

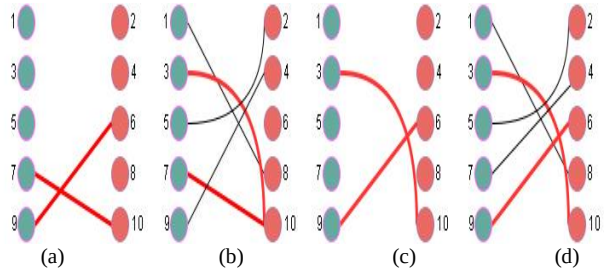


Fig. 8. The processing example of the bad edges: a) bad edges from Fig. 1(a), b) adjacent
edges after inserting the bad edge in Fig. 3(b), c) final bad edge set, d) final distribution of the bad edges

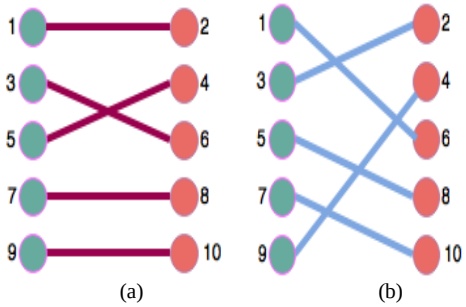


Fig. 9. The processing example of complete matching: (a) and (b) complete matching from procedure one, c) and d) complete matching from procedure one

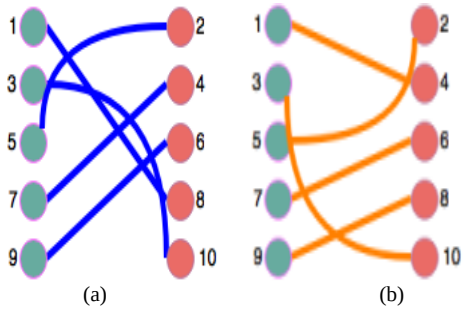


Fig. 10. The processing example of complete matching: (a) and (b) complete matching from procedure two, c) and d) complete matching from procedure two

The perfect edge coloring of Fig. 1(a) is found after combining all the complete matching in Fig. 9 and Fig. 10 that is shown in Fig.11.

Step 2: If it is unable to redistribute the bad edges in the incomplete matching then the same process can be followed for full matching and incomplete matching sequentially until re-distributed the bad edges into the D-matching's.

After *O (D.V)* times, the bad edge sets and the disjoint vertex sets are similar and re-distributed them according to the matching. After finding D-matching color the edges are colored into D-color. This part of the algorithm takes *O (D.V)* times. Because each bad edge has (2D-2) adjacent edges, so, to re-distribute each bad edge to a matching perfectly, it needs maximum (2D-2) iteration. So maximum *V* number of bad edges needs to iterate for *V*(2D-2) = 2D.V-2V = O (D.V)* times on average.

So, the overall time complexity of edge color matching algorithm is *O (D.V) +O (D.V) =O (D.V).*

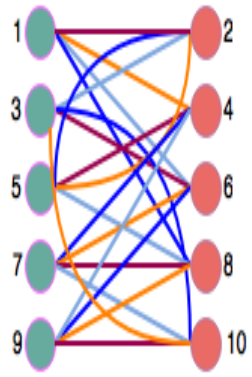


Fig. 11. The processing example of final color matching: the final edge coloring of Fig. 1 with perfect matching

IV. CASE STUDY

Case 1: In the following case study shows a bipartite graph that has 8 vertices, 16 edges and maximum degree 4. Here, the graph is 4-regular.

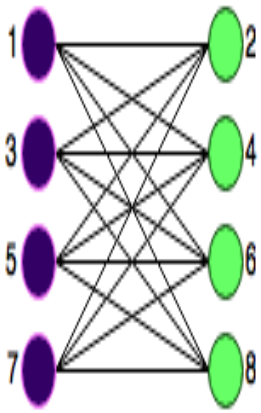


Fig. 12. The processing example of the case study: A regular bipartite multigraph with 8 vertices and degree 4

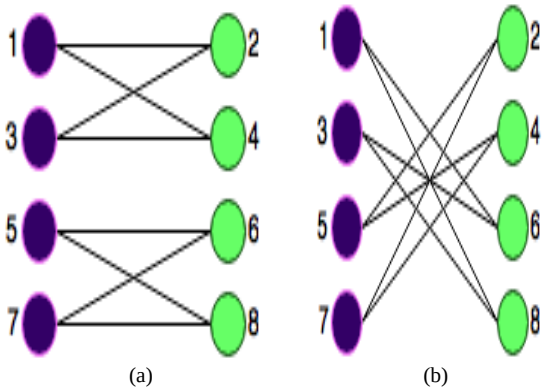


Fig. 13. The processing example of finding closed paths from Fig. 12: a) and b) the set of closed paths from Fig. 12

Fig. 13(a) and Fig. 13(b) shows two sets of the closed path from which four matching can be achieved that are shown in Fig. 14 and Fig. 15. Fig. 14 shows two matching that is achieved from the Fig. 13(a) closed path sets. And Fig. 15 shows two matching which is achieved from the Fig. 13(b) closed path sets.

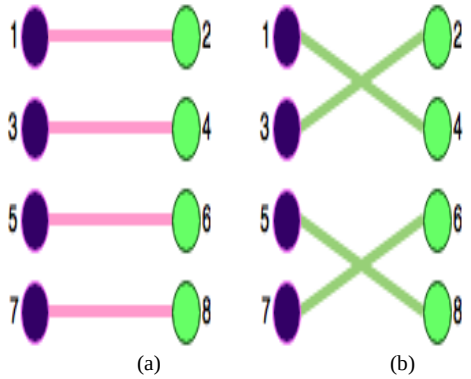


Fig. 14. Processing example of matching found: a) and b) matching from Fig. 13(a)

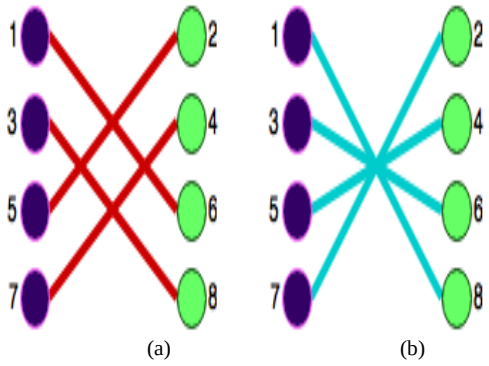


Fig. 15. The processing example of matching: a) and b) matching from Fig. 13(b)

After getting these four matching shown in Fig. 14 and Fig. 15 color the edges of each matching with separate colors to represent the edge color matching shown in Fig. 16.

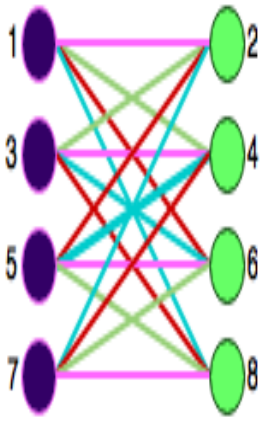


Fig. 16. Processing example final color matching: the final edge coloring of Fig. 11 with perfect matching

Case 2: In the following case study shows a bipartite graph that has 10 vertices, 20 edges and maximum degree 4 with 1 multiple edges.

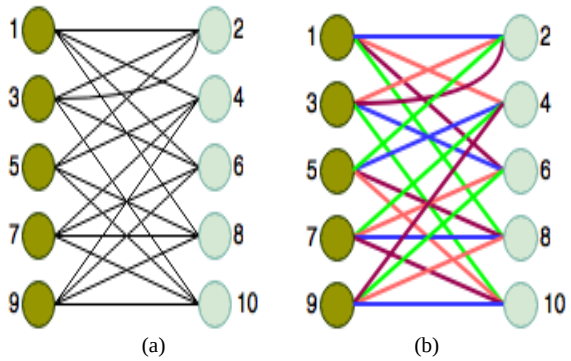


Fig. 17. The edge coloring of with perfect matching: a) input graph and b) final edge coloring with perfect matching

Case 3: This is a regular bipartite graph with no multiple edges. This graph consists of 20 edges, 10 vertices and 4 edges in every vertex, i.e., degree 4.

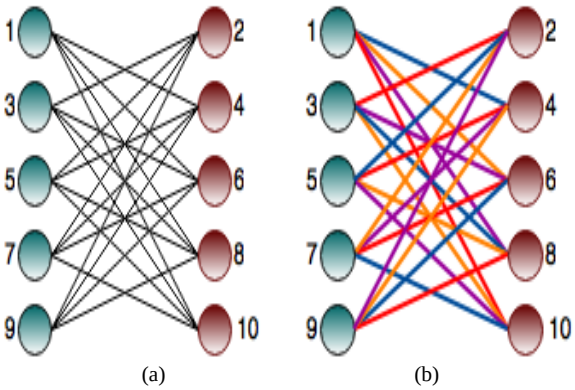


Fig. 18. The edge coloring of with perfect matching: a) input graph and b) final edge coloring with perfect matching

V. Experimental Results

In this paper, four different types of bipartite graph including multi-graphs are used to verify the proposed edge color matching algorithm. The first graph is used in the processing example and describes the proposed algorithm step by step. Other three graphs are shown in the case study, where case 1 is explained shortly with processing example. Case 2 and case 3 are shows the output only. The experimental results run time of procedure 1 that is described in Theorem 2 for the four types of the regular bipartite graph is shown in Table I. Accordingly, the experimental results run time of procedure 2 that is describes in Theorem 2 for the four types of the regular bipartite graph is shown in Table II. From the experimental result in Table I and Table II it is seen that total runtime is less than O (D.V).

TABLE. I. The Experimental Runtime for Procedure 1 Theorem 2

Input Graph	Number of vertices (V)	Degree of the graph (D)	Runtime for procedure 1
Processing example (Fig. 1(a))	8	4	16
Case 1 (Fig. 11(a))	10	4	33
Case 2 (Fig. 14(a))	10	4	34
Case 3 (Fig. 15(a))	10	4	26

TABLE. II. The Experimental Runtime for Procedure 2 Theorem 2

Input Graph	Number of vertices (V)	Degree of the graph (D)	Runtime for procedure 1
Processing example (Fig. 1(a))	8	4	0
Case 1 (Fig. 11(a))	10	4	4
Case 2 (Fig. 14(a))	10	4	4
Case 3 (Fig. 15(a))	10	4	14

VI. Conclusions

This paper has been presented an algorithm of edge coloring for finding perfect matching. This paper considered the regular bipartite multigraphs. To prove the algorithm two theorems for edge coloring is considered. The first theorem shows the perfect matching of a regular bipartite graph and the second theorem shows proper edge coloring of a D-regular bipartite multigraph. The overall time complexity of the proposed edge color matching algorithm is O *(D.V) +O (D.V = O (D.V)* times. The algorithm reduces overall time complexity. Experimental results show that the total runtime is less than *O (D.V)* time. A graph with large vertex cannot be considered with the proposed theorem. In future, this work will be extended to develop an algorithm in order to solve the large volume of the graph.

References

[1] J. Bensmail, A. Lagoutt, & P. Valicov,outt, & P. Valicov, "Strong edge-coloring edge coloring with perfect matching of (3, Δ)-bipartite graphs," *Discrete Mathematics*, Vol. 339, No. 1, pp. 391-398, 2016.

[2] H. A. Razak, Z. Ibrahim, and N.M. Hussin," Bipartite graph edge coloring approach to course timetabling," In 2010 *International Conference on Information Retrieval & Knowledge Management, (CAMP),* pp. 229-234, IEEE, March, 2010.

[3] C. J. Casselgren and B. Toft "One-sided interval edge-colorings of bipartite graphs," *Discrete Mathematics*, Vol. *339*, pp. 2628-2639, 2016.

[4] A. Schrijver, "Bipartite Edge Coloring in O(Δm) Time, **"***SIAM Journal on Computing*," Vol. 28, No. 3, pp. 841-846, 1998.

[5] R.Lang, O.Schaudt, and M. Stein, "Partitioning 3-edge-coloured complete bipartite graphs into monochromatic cycles," *Electronic Notes in Discrete Mathematics*, Vol. 49, pp. 787-794, 2015.

[6] G. J. Puleo, "Complexity of a disjoint matching problem on bipartite graphs," *Information Processing Letters*, 2016.

[7] J. Engbers and D. Galvin, "H-colouring bipartite graphs, " *Journal of Combinatorial Theory, Series B*, Vol. 102, No. 3, pp. 726-742, 2012.

AnyCasting In Dual Sink Approach (ACIDS) for WBASNs

Muhammad Rahim Baig
Department of Computer Science
CECOS University of IT and Emerging Sciences
Peshawar, Pakistan
Preston University, Peshawar, Pakistan

Sheeraz Ahmed
Department of Electrical Engineering
Gomal University
Dera Ismail Khan, Pakistan
CECOS University, Peshawar, Pakistan

Najeeb Ullah
Department of Computer Science
CECOS University of IT and Emerging Sciences
Peshawar, Pakistan

Abdul Hanan
Department of Computer Science
CECOS University of IT and Emerging Sciences
Peshawar, Pakistan

Fazle Hadi
Department of Computer Science
Preston University, Peshawar, Pakistan

Imran Ahmed
Department of Computer Science
Institute of Management Sciences, Peshawar, Pakistan

***Abstract*—After successful development in health-care services, WBASN is also being used in other fields where continuous and distant health-care monitoring is required. Various suggested protocols presented in literature work to enhance the performance of WBASN by focusing on delay, energy efficiency and routing. In this research we focus to increase the stability period and throughput, while decreasing end-to-delay. Two sink nodes are utilized and concept of AnyCasting is introduced. In this research, we have presented a scheme AnyCasting In Dual Sink (ACIDS) for WBASN and compared it with existing protocols LAEEBA and DARE. The performance of ACIDS is found to be 51% and 13% efficient than LAEEBA and DARE respectively in throughput. Results show that, the stability period of ACIDS is much greater than LAEEBA and DARE with minimum delay. Energy parameter in ACIDS is in tradeoff with the improved parameters, due to the computation of RSSI which does more processing and utilizes more energy.**

***Keywords*—Stability Period; WBASNs; Throughput; End-to-End Delay; Energy Consumption; AnyCasting; RSSI**

I. INTRODUCTION

Computer Science is now growing rapidly to operate large data and keep higher level of linking. At the current era advancement also occur in small scale networks which support higher level of access and mobility [1], [2]. Wireless Body Area Sensor Networks (WBASNs) is a new field of Wireless Sensor Networks (WSNs), started developing since 1995 [3]. The objective of this network is to establish communication associated to human body. Initially it was developed for the purpose of providing health-care services to serious patient. After the successful development in health-care services now WBASNs is also being used in other fields where continuous and distant health-care monitoring is necessary, like, players in sports, astronaut in space, environmental condition, motion detection of animals, security, etc. [4], [5].

WBASNS consists of small size sensors which have minimum energy source and processing capability. Depending upon the application, these sensors can be fixed inside the body or can be wearable sensors attached on body [2].

Data rate and power consumptions of implanted devices are low as compared to that of wearable devices [6], [7]. In medical field, different sensors are attached to human body that is residing at home or hospital. These sensor nodes detect and monitor different biological parameters of human body like glucose-rate, Electrocardiogram (ECG) temperature, heart-rate, blood-rate, heart beat and blood pressure etc. [8]. The received signals are aggregated or collected by a personal device, e.g. Personal Digital Assistance (PDA) which, act as a relay node and sent them to health treatment professionals or to some internet using applications with the help of sink node. The sink acts as a source of route between the hospital, and the WBASNs [9].

In WBASNs, sensor nodes are battery driven units that run with limited energy source [7]. It is necessary to use minimum energy for transmitting data from sensor nodes to sink. One of the major task in this network is to recharge the batteries of senor nodes, therefore it is demanded for energy efficient protocol which lessen the issue of recharging the batteries. In [4], researchers found some challenges faced by WBASNs which are as follows:

- Trade-off between communication and processing
- Bandwidth and power consumption
- High level of attenuation as compared to other WSNs applications

- Storage and energy harvesting

Usually, in WBASNs there is sink that collects data from the battery-powered sensor nodes and use unicast transmission to transmit the sensed data to the sink node; in transmission every sensor node has only one destination, which is a single sink. Every sensor nodes sense data and forward to this single sink node. Problems may arise while using single sink and transmitting data through unicasting, which are discussed in detail in letter section. Another way of using two sinks is a well approach in WSNs; in two sinks approach all the sensors in the sensing area sense the environmental information and forward to any available sink node in AnyCast manner [10]. AnyCasting is a best working technology in WSNs and none of researchers use this technology in WBASNs.

WBASNs, in the current era of sensor networks is the hot area of research, where, much more papers have been published on Quality of Service (QoS) energy efficiency, bandwidth and security, the topology of nodes in a body and transmission of data between the sensors and the sink is the focusing point. Therefore, in this study we are going to address these issues by considering the transmission technique.

Rest of the paper is further divided into sub sections which are described as follows. Section II provides some existing works. In section III motivation behind this research is mentioned. Proposed model for ACIDS protocol is presented in Sections IV. Simulation results are discussed in section V and finally the paper is concluded in section VI.

II. Related Work

Different researchers have proposed different protocols and techniques to increase performance and reliable communication. Some of them are summarized and discussed below.

In paper [1], authors propose Incremental relay-based CoCEStat protocol for Wireless Body Area Networks (InCo-CEStat). Author's uses two relay nodes for cooperation purpose to quick transmit the critical data in emergency cases in WBASNS. Relay nodes detects the data from sensors and forward to the single destination sink node in three phases with a dublicat copy to decrease the chance of packet drop. Researchers in [2], propose a protocol called DARE. This protocol is used to monitor eight patients in a ward, each comprises of seven sensors to monitor different parameters of the patient body. Five topological scenarios have been proposed where more than one sink nodes are placed in different location of the ward.

In [3], authors present a protocol name as SIMPLE. In this protocol, author uses eight sensors and single sink node in human body. Nodes lose enery in short span of time therfore, to balance energy level and stability period of the network a mathematical formula is suggested to select cluster head/forwarder in a cluster. In paper [5], authors present a protocol; THE-FAME (Threshold based Energy-efficient Fatigue MEasurment) to find fatigues of a player in a Soccer game. Each player is implanted a sensor node to sense fatigue's parameter of the player. Multiple sink nodes are attached to the different side of the game. The size of sensor kept small, sensor send data through direct communication when threshold level is reached.

Researchers in [6], present hardware and software architecture, error detection and data transmission for normal and emergency data. Authors used a cluster head (CH) to gather the parametric data. Researchers in [8] present another protocol called Effect of Packet Inter-arrival Time on the Energy Consumption of Beacon Enable MAC Protocol for WBASNS. In this protocol, researchers focus on Media Access Control (MAC) to reduce the energy consumption. Two type of nodes has been used; Fully Functional Devices (FFD) and Reduced Functional Devices (RFD). FFD can be simple node or coordinator node and RFD works only as simple node.

In [7], a protocol called LAEEBA is proposed for Wireless Body Area Networks. In this scheme authors uses both direct communication and multi-hop communication by focusing the path loss. Cost function is calculated to minimize the energy consumption. In paper [9], a protocol called M-ATTEMPT is proposed, which is threshold based routing protocol to find the link spot in the established links. In this model authors placed the sensors in decreasing order of their data rate with respect to the sink. Both single hope and multi hope communication is used for emergency and normal data. In [11], S. Ahmad et al. revised their previous work (LAEEBA), and propose new routing protocol for WBASNS named as Co-LAEEBA. New protocol is suggested for collective working and path loss factors. Knowing the residual energy and sharing the distance from sensor to the sink, cost function is introduced to select the feasible route to the destination.

In paper [12], researchers propose a routing protocol that focuses the accessibility of patient's data either by offline or online to the health care takers. According to the propose model different sensors are placed at the patients cloths (on body). Data are gathered by the accumulator and send to the medical server through Wi-Fi. Researchers in [13], take four attributes into accounts which are; distance covered, residual energy, hop-count and node criticality. Routing occur in two phase i.e. setup and operational phase. In paper [14], cost based routing protocol is proposed for Wireless Body Area Networks" is proposed. In this paper energy efficiency and cost function is focused by considering the reliability of path on basis on critical factor. In paper [15], researchers used two existing protocol: Dynamic Source Routing (DSR), and Energy efficient self-adaptive route E-DSR with tree topology to minimize energy consumption and network lifecycle as. In paper [16], authors present a protocol which improves the throughput and avoid single point of failure in WBASNS. The protocol uses Cooperative Network Coding (CNC) in many to many as multiple input multiple output (MIMO). Researchers in [17], propose a robust and fast routing protocol focusing on topological design. Because of the heterogeneity of sensor nodes and their energy consumption researcher takes advantage of suitable linear relaxations to guide a randomized fixing of the variables, supported by large variable neighborhood search. This algorithm also focuses the traffic uncertainty in the design of WBASNS by using relay nodes, and use preferably single path routing.

In paper [19], authors propose a mathematical technique that uses network topology and cross-layer optimization in WBSNs. Authors introduce multilevel primal and dual decomposition methods to solve the non- convex mixed-integer optimization problem.

Authors in [20] design a network architecture that uses WBASNS and Cloud for the purpose of sharing data. Data sharing in this model occurs in four different layers with the help of TCP/IP and Zigbee. In-order to support mobility of users, WBAN coordinators inter-operates with different local networks such as WiFi and LTE. Furthermore, adaptive streaming technique is also used in this paper to reduce packet loss.

III. Motivation

According to the literature survey, majority of the researchers [1], [3], [7]-[11] used single sink node to aggregate and send the sense data from the sensor nodes to the destination. Single sink node collects data from all sensors and forward to the destination. Some problems may arise for which the researchers did not focus which are: first, the overall topology will fail when the single sink node die or stop working for some time for any reason. Second, all sensors send their data to this sink which makes an over burden for it to aggregate and quick delivery. Over burden problem leads to delay between packets, especially for emergency data which needs quick delivery. The performance factor increase when more number of packets will send in less time and received error free to the medical caretaker. This over burden on sink leads to decrease in delivery ratio.

In [17], authors presented a new scheme to make fast and robust design for WBASNS. In this scheme authors prefer to use a greater number of relay nodes in the body, almost equal to normal nodes, leads to increase the network cost creates delay and decrease delivery ratio, which the author did not focused. The placement of sink in exact location by considering the LoS and NLoS is important. Misplacement of sink node leads to disconnection of sensor from the sink node when the body parts move. According to the figure representation correct placement of sinks are not mentioned. In [14], author placed the sink at wrist that leads to disconnect the previously established link when the arm moves back and forth due to NLoS. To overcome the limitation in above discussion we proposed ACIDS scheme for WBASNs. In ACIDS, the performance is increasing by taking care of delivery ratio, stability period and end-end-delay.

IV. ACIDS Proposed Protocol

In this section, we presents our proposed routing protocol ACIDS for WBASNS with AnyCasting technique to improve the performance of routing scheme by using two sinks.

A. Network Topology

In ACIDS we placed two sinks named as S1 and S2 with eight sensors. S1 is placed at right lumbar and S2 is at left lumbar to consider NLoS communication in case of body parts movement i.e. arm. Both cases of moving back and forth the arm is in LoS communication with lumbar, therefore it is a suitable location we selected. Node 1and 4 transmit data direct to S1 and S2 in order to consider the NLoS issue (as discussed above). The other nodes which are in direct range to either sinks transmit data in direct communication and the rest nodes transmit through multi-hop communication by selecting a forwarder node. Topology of our proposed protocol ACIDS is mentioned in figure 1.

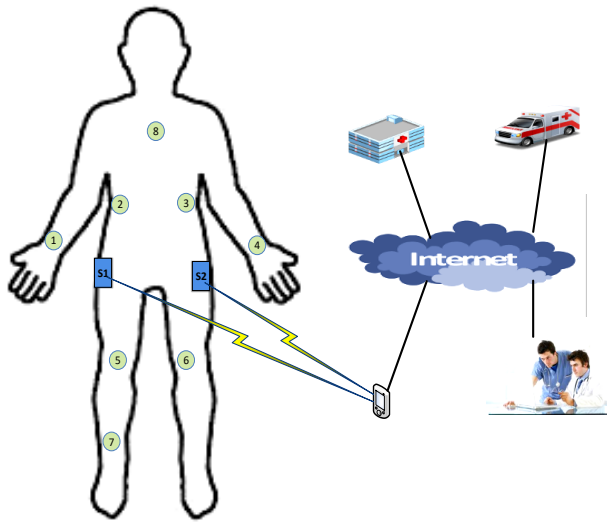


Fig. 1. Schematic diagram for ACIDS Protocol

B. Initialization Phase

In this phase, two sinks initially broadcast a “Hello” message to all sensor nodes that contains Sinks Id and location. Receiving this message sensor nodes save location and Id of Sinks in their routing table. Sensor nodes also broadcast some packets that contain the node Id, node energy level and their location. In this way each node is aware about their neighbors.

C. Selection of Next-Hop Phase

ACIDS computes forwarder node on the basis of threshold Residual Energy (RE). ACIDS fix a threshold value of RE is equal to 0.1 J.

RE >= Threshold……where, Threshold = 0.1J..... (1)

Here, if more than one nodes RE greater than threshold, than forwarder node will be selected through Received Signal Strength Identification (RSSI), [18]. RSSI describe relation between the received power and the transmitted power of wireless signal and the distance among them. ACIDS follow the RSSI model of [19] that is mentioned below in equation (2).

$$Pr = Pt.\left(\frac{1}{d}\right)^{n} \quad \text{..... (2)}$$

Here, Pr, is the wireless signal received power. Pt, shows the power transmitted by the wireless signal, and distance between sending and receiving node is denoted by d. Transmission factor between sending and receiving node is denoted by n, whose value depends on propagation

environment. According to this equation, the node is selected for forwarder node whose Pr value is maximum.

TABLE I. SIMULATION PARAMETERS USED

Parameters	Values
DC current (RX)	18 mA
DC current (TX)	10.5 mA
Minimum supply voltage	1.9 V
ERr	36.1nJ/bit
ERt	16.7nJ/bit
ERamp	1.97nJ/bit
Wavelength (λ)	0.135 m
Frequency (f)	2.5 GHz
Initial Energy (Eo)	0.7 J
do	0.15

D. Routing and Energy Consumption Phase

Another advantage of using two sinks is that, majority of nodes comes in direct range. In direct communication packets are delivering without any delay. According to the proposed protocol ACIDS, nodes 1and 4 send data directly to S1 or S2 to consider the NLoS of arm movement. The nodes which are in direct range to either S1 or S2 will send data directly, other nodes AnyCast to S1 or S2 through relay node in multi-hop communication. Relay node will be selected until it comes in direct range to either sink. The communication flow of the ACIDS protocol is mentioned in figure 2.

Energy consumption in multi-hop (E_m) communication is mentioned below [5]:

$$ER_{t\text{-}m}(k, d) = n \times (ER_{cr} + ER_{amp}) \times k \times d^2 \ldots\ldots\ldots\ldots (3)$$

$$ER_{r\text{-}m}(k) = (n\text{-}1) \times (ER_{cr} + ER_{amp}) \times k \ldots\ldots\ldots\ldots\ldots (4)$$

$$ER_{total\text{-}m} = ERt\text{-}m + ER_{r\text{-}m} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (5)$$

Energy consumption in direct (E_D) communication is [5]:

$$ER_{t\text{-}D}(k, d) = (ER_{cr} + ER_{amp}) \times k \times d^2 \ldots\ldots\ldots\ldots\ldots\ldots (6)$$

$$ER_{total\text{-}D} = ER_{t\text{-}D} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (7)$$

Here, ER_t and ER_r are the energy required for transmission and receive by sender and receiver. In equation k, represent the size of bits and d, is the distance between nodes and sinks. ER_{cr} is the required energy to run the electronic circuit of the receiver and transmitter. ER_{amp} is the required energy to amplify k bits to the distance d. in the above equation 3 and 4 n is the number of nodes needed to reach the sinks and d^2, is the loss of energy while transmitting data through transmission channel. The energy parameters depend on the hardware used. Energy parameters used in ACIDS are mentioned in the table 1.

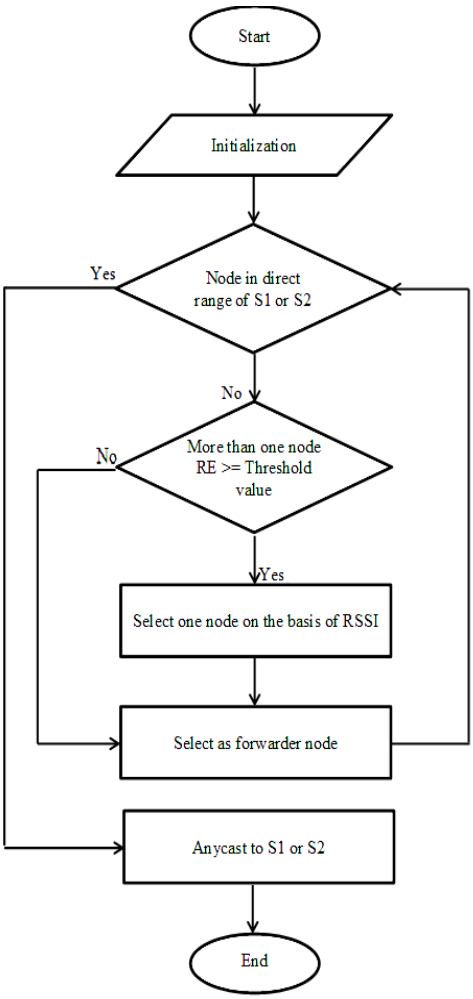


Fig. 2. Flow Chart for ACIDS Protocol

V. RESULTS AND DISCUSSIONS

In this paper we have conducted a series of simulations to measure the performance of our proposed protocol ACIDS. We compared it with the existing protocols: DARE and LAEEBA. All three protocols are evaluated for same key performance parameters: Stability Period, End-to-End Delay, Throughput and residual energy. Energy Parameter used in simulations is enlisted in table 1.

A. Stability Period

Stability Period is the time duration of operation of network till the first node dies. According to figure 3, performance of ACIDS is much better than LAEEBA and DARE. Table 2 shows the numerical values of dead nodes for all the three protocols starting from time 1000 till time 10000. First node of ACIDS dies at time 7200, while LAEEBA loses its first node at time 2100 and DARE at time 4250. The

increased performance of ACIDS is because of availability of more than one sink. Each sink comes near to nodes which helps them to send data to the nearest sink node. Another obvious reason for better performance is the location of sinks. The wrist sensors nodes direct communicate to S1 and S2 without NLoS issue. These factors also improve the performance of ACIDS up to the 10000 times. All nodes of LAEEBA and DARE are dead at time 8000, while last node of ACIDS dead in time 10000.

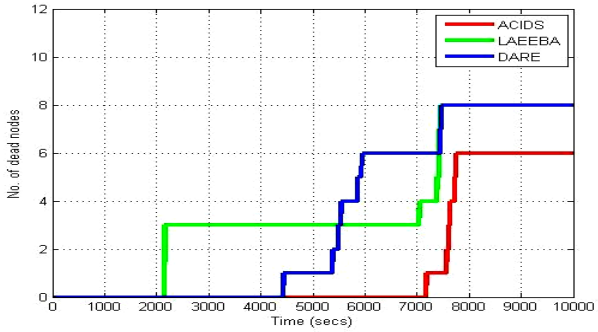


Fig. 3. Stability Period vs Time

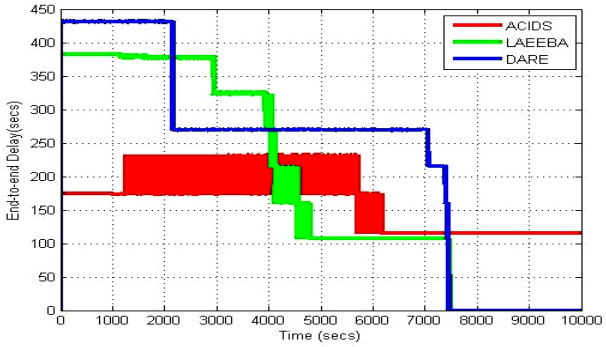


Fig. 4. End-to-End Delay vs Time

B. End-to-End Delay

The term delay refers to the time lag between the source nodes and the destination sinks. Minimizing delay in WBASNS is an important factor. The data sense from the sensors is to be evaluated by medical specialists, so it is important to receive data in time. It will be obtained if the sinks gather data in time. Figure 4 shows that the performance of ACIDS protocol, as compared to LAEEBA and DARE protocol is much better. The average delay values are mentioned in table 3, which shows that ACIDS takes an average of 162.34, which is 27% efficient than DARE. LAEEBA takes an average of 167.97. The evaluated percentage value in table 5.2 shows that the ACIDS is 2% efficient than LAEEBA. The achieved performance is due to the availability of AnyCasting between two sink nodes. The sensor nodes have two option to send their data. The LAEEBA and DARE protocol uses single sink which waste their time in calculation for gathering data from more than one sensor nodes. The availability of two sink in ACIDS makes its performance better because each sink gather few of the sensors data. Another valid reason for achieving better performance as compared to LAEEBA and DARE is: majority of the sensors in ACIDS comes in direct communication to either S1 or S2 and the distance between sensors and sink become lessen. It is obvious from the literature survey that delay will decrease in direct communication as compared to Multi-hop communication if the distance is lessens.

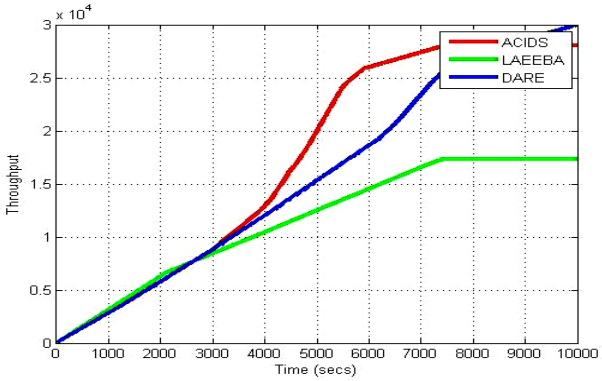


Fig. 5. Throughput vs Time

C. Throughput

Throughput is the successful delivery of packets from sensors to the sinks in per unit time. The unit here is taken as second starts from 1000 to 10000. Two type of links used in ACIDS one is from sensors to the forwarder node and one is from forwarder to sinks. The link from forwarder to the sink transmits more data as compared to the other link. So, in ACIDS more number of sensors come near to sinks, which transmit data directly, which leads to increase the throughput. Figure 5, shows a clear difference in performance among three protocols. Results show that the performance of ACIDS is better than that of other two. The average throughput value of ACIDS according to table 4 is 18790, which is 51% better than LAEEBA and 13% than DARE.

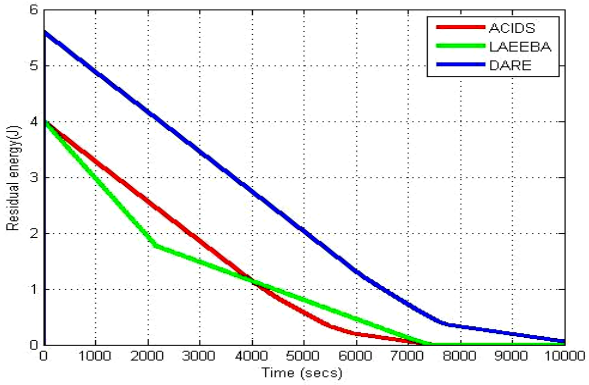


Fig. 6. Energy Consumption vs Time

D. Net Energy Consumption

For clear evaluation, we kept same initial energy for all protocols that is 0.7 Joule. According to the figure 6 and table 5, it shows that the energy level of ACIDS protocol is better than LAEEBA protocol but not efficient than DARE protocol. The energy parameter in ACIDS comes in tradeoff with delay, throughput and stability period. In ACIDS our aim was to improve above said parameters. We proved our aim by

showing efficient results in these parameters. However, ACIDS shows an improved performance as compared to LAEEBA.

The improved performance of DARE in energy is just because that DARE is focused to improve the energy parameter of WBASNS, in ACIDS we focused to improve the throughput, end-to-end delay and stability period. Another reason of depletion in energy is the use of RSSI in ACIDS. Our proposed protocol ACIDS do extra calculation to find the forwarder node when two or more nodes have the residual energy more than threshold range of energy. This extra calculation leads in depletion of energy more than DARE protocol. Tie situation in selecting the forwarder node is the focusing point, unfortunately the researchers did not focused.

TABLE II. DEAD NODES VS TIME (SECONDS)

Protocol	1000s	2000s	3000s	4000s	5000s	6000s	7000s	8000s	9000s	10000s
LAEEBA	0	0	3	3	3	3	3	8	8	8
DARE	0	0	0	0	1	6	6	8	8	8
ACIDS	0	0	0	0	0	0	2	6	6	8

TABLE III. END-TO-END DELAY VS TIME (SECONDS)

Protocol	Improvement		1000s	2000s	3000s	4000s	5000s	6000s	7000s	8000s	9000s	10000s
	%Age	Average										
LAEEBA	75%	167.97	382.7	380.7	327.3	268.9	106.7	106.7	106.7	0	0	0
DARE	100%	221.52	431.8	433.3	269.4	270.7	270.1	269.3	270.6	0	0	0
ACIDS	73%	162.34	174.6	173.5	232.3	173.5	232.5	174.6	115.6	115.6	115.6	115.6

TABLE IV. PACKET DELIVERY RATIO VS TIME (SECONDS)

Protocol	Improvement		1000s	2000s	3000s	4000s	5000s	6000s	7000s	8000s	9000s	10000s
	%Age	Average										
LAEEBA	100%	12479.8	3167	6296	8515	1.057e+04	1.261e+04	1.465e+04	1.667e+04	1.744e+04	1.744e+04	1.744e+04
DARE	138%	17246.3	2800	5777	8806	1.199e+04	1.532e+04	1.861e+04	2.339e+04	2.719e+04	2.859e+04	2.999e+04
ACIDS	151%	18790	2819	5782	8829	1.289e+04	2.004e+04	2.6e+04	2.74e+04	2.805e+04	2.805e+04	2.805e+04

TABLE V. RESIDUAL ENERGY VS TIME (SECONDS)

Protocol	Performance		1000s	2000s	3000s	4000s	5000s	6000s	7000s	8000s	9000s	10000s
	%Age	Average										
LAEEBA	100%	0.888	2.965	1.928	1.479	1.139	0.7988	0.4586	0.1182	0	0	0
DARE	2.239%	1.989	4.881	4.169	3.457	2.745	2.03	1.314	0.7188	0.3272	0.1931	0.05909
ACIDS	1.091%	0.969	3.281	2.569	1.857	1.147	0.5748	0.1954	0.06131	0	0	0

VI. CONCLUSION & FUTURE WORK

Abundant of research papers have been published that focus different parameters like; energy consumption, topology, throughput, QoS, end-to-end delay, stability period. Different researchers proposed protocols to enhance the performance of WBASNS by focusing the above said parameters. In this research thesis we focused to increase the stability period, decrease the end-to-delay, and increase the throughput. In order to achieve our goal we proposed a complete topology with mathematical work. In this research we implement eight sensors on the human with two sinks. The main idea was to implement AnyCasting by using two sink nodes with the concept of AnyCating. In mathematical work, we have also RSSI to select forwarder node in case of tie situation.

We compared our protocol ACIDS with two existing protocol LAEEBA and DARE through Matlab simulator. Before simulation the performance of LAEEBA and DARE have been analyzed. Performance of these three protocols has been analyzed for the individual parameter separately in different time (seconds) started from 1000s to 10000s.

Simulation results shows that the performance of ACIDS is more efficient than LAEEBA and DARE in throughput, stability period and end-to-end delay. The increased efficiency of ACIDS is due to the availability of more than one sink node instead of one. The sensors send their either S1 or S2 in AnyCast manner. The other reason of efficiency in ACIDS is topological factor; in ACIDS majority of the nodes comes in direct range to sink. This leads to send more data in less delay manner. In this research, we also proved that the energy

consumption of ACIDS comes in tradeoff with above said parameters. We got improvement in above said parameters but not gained better result in energy efficiency. It is proved that ACIDS is better than LAEEBA and not efficient than DARE. The reason is, DARE focused more on energy efficiency and we have focused to improve the said three parameters and got efficient result. The other reason is we introduce RSSI, for which the protocol do more processing and utilize more energy.

In future, our focus will be on energy conservation in this design consideration; and try to implement energy harvesting concept. This concept will improve the performance in great extant. Energy harvesting is a concept where sensors obtain energy from external sources like thermal energy, solar system, wind power etc. Sensors store and use this energy for their processing purpose. As the WBASN have application driven protocols, different applications are introducing in new researches, so we are willing to implement this protocol to new application. Besides the above discussed issues, there are still other challenges existing in WBASNS that are still unresolved. Among them some are; security, data management, scalability and constant monitoring etc.

REFERENCES

[1] Yousaf, S., Ahmed, S., Akbar, M., Javaid, N., Khan, Z. A., &Qasim, U. (2014, November). Incremental relay-based co-cestat protocol for wireless body area networks. In Broadband and Wireless Computing, Communication and Applications (BWCCA), 2014 Ninth International Conference on (pp. 113-119). IEEE.

[2] Tauqir, A., Javaid, N., Akram, S., Rao, A., & Mohammad, S. N. (2013, October). Distance aware relaying energy-efficient: Dare to monitor patients in multi-hop body area sensor networks. In Broadband and Wireless Computing, Communication and Applications (BWCCA), 2013 Eighth International Conference on (pp. 206-213). IEEE.

[3] Nadeem, Q., NadeemJavaid, S. N. Mohammad, M. Y. Khan, SohabSarfraz, and M. Gull. "Simple: Stable increased-throughput multi-hop protocol for link efficiency in wireless body area networks." In Broadband and Wireless Computing, Communication and Applications (BWCCA), 2013 Eighth International Conference on, pp. 221-226. IEEE, 2013.

[4] Nadeem, A., Hussain, M. A., Owais, O., Salam, A., Iqbal, S., &Ahsan, K. (2015). Application specific study, analysis and classification of body area wireless sensor network applications. Computer Networks.

[5] Akram, S., Javaid, N., Tauqir, A., Rao, A., & Mohammad, S. N. (2013, October). THE-FAME: THreshold Based Energy-Efficient FAtigueMEasurement for Wireless Body Area Sensor Networks Using Multiple Sinks. In Broadband and Wireless Computing, Communication and Applications (BWCCA), 2013 Eighth International Conference on (pp. 214-220). IEEE.

[6] Bahanfar, S., Darougaran, L., Kousha, H., &Babaie, S. (2011). Reliable communication in wireless body area sensor network for health monitoring.arXiv preprint arXiv: 1112.0393.

[7] Khan, Z. A., Rasheed, M. B., Javaid, N., & Robertson, B. (2014). Effect of Packet Inter-arrival Time on the Energy Consumption of Beacon Enabled MAC Protocol for Body Area Networks. Procedia Computer Science, 32, 579-586.

[8] Ahmed, S., Javaid, N., Akbar, M., Iqbal, A., Khan, Z. A., &Qasim, U. (2014, May). LAEEBA: Link Aware and Energy Efficient Scheme for Body Area Networks. In Advanced Information Networking and Applications (AINA), 2014 IEEE 28th International Conference on (pp. 435-440). IEEE.

[9] Javaid, N., Abbas, Z., Fareed, M. S., Khan, Z. A., &Alrajeh, N. (2013). M-ATTEMPT: A new energy-efficient routing protocol for wireless body area sensor networks. Procedia Computer Science, 19, 224-231.

[10] Fazl-e-Hadi, Abid Ali Minhas, " EAA: Energy Aware Anycast Routing in Wireless Sensor Networks", Journal of Engineering and Applied Sciences (JEAS), Vol: 30 No., January-June 2011.

[11] Ahmed, S., Javaid, N., Yousaf, S., Ahmad, A., Sandhu, M. M., Imran, M., ...&Alrajeh, N. (2015). Co-LAEEBA: Cooperative link aware and energy efficient protocol for wireless body area networks. Computers in Human Behavior.

[12] Dinkar, P., Gulavani, A., Ketkale, S., Kadam, P., & Dabhade, S. (2013). Remote health monitoring using wireless body area network. International Journal of Engineering and Advanced Technology (IJEAT) ISSN, 2249, 8958.

[13] Kumari, J., & Prachi. (2015). An Energy Efficient Routing Algorithm for Wireless Body Area Network. IJWMT, 5(5), 56-62. Doi:10.5815/ijwmt.2015.06.

[14] Kaur, H. P., & Goyal, K. (2015). Cost Based Efficient Routing for Wireless Body Area Networks.

[15] Liu, S., Wei, X., & Zhao, M. (2016). Routing Design and Simulation of Body Area Network Based on Node Energy Consumption Control Strategy.

[16] Arrobo, G. E., & Gitlin, R. D. (2011, August). Improving the reliability of wireless body area networks. In Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE (pp. 2192-2195). IEEE.

[17] D'Andreagiovanni, F., & Nardin, A. (2015). Towards the fast and robust optimal design of Wireless Body Area Networks. Applied Soft Computing,37, 971-982.

[18] Xu, J., Liu, W., Lang, F., Zhang, Y., & Wang, C. (2010). Distance measurement model based on RSSI in WSN. Wireless Sensor Network, 2(08), 606.

[19] Zhou, Y., Sheng, Z., Mahapatra, C., Leung, V. C., & Servati, P. (2017). Topology design and cross-layer optimization for wireless body sensor networks. Ad Hoc Networks.

[20] Hassan, M. M., Lin, K., Yue, X., & Wan, J. (2017). A multimedia healthcare data sharing approach through cloud-based body area network. Future Generation Computer Systems, 66, 48-58.

Exploreing K-Means with Internal Validity Indexes for Data Clustering in Traffic Management System

Sadia Nawrin
Dept. of Computer Science and Engineering
East West University
Dhaka, Bangladesh

Md Rahatur Rahman
Simplexhub Ltd.
Dhaka,
Bangladesh

Shamim Akhter
Dept. of Computer Science and Engineering
East West University
Dhaka, Bangladesh

***Abstract*—Traffic Management System (TMS) is used to improve traffic flow by integrating information from different data repositories and online sensors, detecting incidents and taking actions on traffic routing. In general, two decision making systems-weights updating and forecasting are integrated inside the TMS. The models need numerous data sets for making appropriate decisions. To determine the dynamic road weights in TMS, four (4) different environmental attributes are considered, which are directly or indirectly related to increase the traffic jam– rain fall, temperature, wind, and humidity. In addition, peak hour is taken as an additional attribute. Usually, the data sets are classified by instinct method. However, optimum classification on data sets is vital to improve the decision accuracy of the TMS. Collected data sets have no class label and thus, cluster based unsupervised classifications (partitioning, hierarchical, grid-based, density-based) can be used to find optimum number of classifications in each attribute, and expected to improve the performance of the TMS. Two most popular and frequently used classifiers are hierarchical clustering and partition clustering. K-means is simple, easy to implement, and easy to interpret the clustering results. It is also faster, because the order of time complexity is linear with the number of data. Thus, in this paper we are going to demonstrate the performance of partition k-means and hierarchical k-means with their implementations by Davies Boulder Index (DBI), Dunn Index (DI), Silhouette Coefficient (SC) methods to outline the optimal number classifications (features) inside each attribute of TMS data sets. Subsequently, the optimal classes are validated by using WSS (within sum of square) errors and correlation methods. The validation results conclude that k-means with DI performs better in all attributes of TMS data sets and provides more accurate optimum classification numbers. Thereafter, the dynamic road weights for TMS are generated and classified using the combined k-means and DI method.**

Keywords—Traffic Management System (TMS); Data Clustering; K-means; Hierarchical Clustering; Cluster Validation

I. INTRODUCTION

A new low cost, flexible, maintainable, and secure internet-based traffic management system with real time bi-directional communication was proposed and implemented (in [1][2][3][4]) to assist and reduce the traffic situation. To determine the dynamic road weights in TMS, four (4) different environmental attributes - rain fall, temperature, wind, and humidity are considered. Rainfall is one of the most influential weather attributes to determine the road congestion in metro city, as the road segments are submerged due to the heavy rains, and makes slower traffic movements. The heat released from the engines, air-conditioners of the traffic stacked vehicles, may raise the overall temperature of the area. Thus, the current temperature helps to classify traffic congestion status of a particular road segment. Gusts of wind have direct influence on road safety and that pushes to slower vehicle movement. In addition, temperature, wind and humidity have direct influence to predict the future rainfall in a particular area. Peak hour is one of the most influential attributes to cause traffic congestion in metro cities. Thus, these four (4) environmental attributes and peak hour have direct or indirect relationship on traffic congestion as well as vehicle movement and influence to choose them as decision making parameters.

The value of these attributes (features) are intelligently crawled by search engine, with metadata indexing (title, description, keyword etc.), directly from the multiple data feeds (like web site, RSS feeds, web service etc.) from the web page in [5]. Crawled data are simplified (structured) and stored in a historic table. However, the number of attributes can be changed according to the system requirements. We collect more than two (2) years or 750 days (1/12/2006 to 20/12/2008) data of five features from the web page in [5].

Initially, decision tree (DT) [1] [2] [3] was used to classify road weights and weighted moving average analytic was implemented to estimate or predict feature values in DT [28][29] based system and achieved 16.45% accuracy. However, the model data sets were classified by instinct method. Cluster based classifications (K-means, Locality-Sensitive Hashing (LSH) etc.) can be used to find optimum number of classifications in each feature and can improve the performance of the TMS. With this hypothesis, we implement two unsupervised clustering techniques partition k-means and hierarchical k-means. There are several methods (internal/external) to measure similarity between two clustering steps and used to compare how well different data clustering algorithms perform on a set of data. Only internal methods - Davies Boulder Index (DBI), Dunn Index (DI), and Silhouette Coefficient (SC) - are used to choose the optimum number of classification, as they do not have any external information. Subsequently, the optimal classes are cross-validated by using statistical analytics - correlation and Within Sum of Square (WSS) errors.

Results highlight that Dunn Index (DI) performs better for both partition k-means and hierarchical k-means algorithms by

providing minimum Sum of Square Error (SSE) for all environmental attributes. However, the optimum numbers of classifications are generated by both algorithms, for each environmental attribute, differs in their numbers. Both algorithms are compared by computing the correlation values on their optimal number of clusters for each attribute. The correlation values of partition k-means algorithm are higher than the correlations of hierarchical k-means algorithm for all attributes. The validation results conclude that the combination of the k-means with Dunn Index performs better and provides more accurate optimum classification number(s) on environmental data set. Thereafter, the dynamic road weights for TMS are generated and classified with these combined algorithms.

II. Related Works

Integrating intelligence technologies in transportation system including intelligent and effective route planning to reduce travel time, reliable estimation of traffic congestion, accident and/or hazard detection etc., can help to reduce both fuel consumption and the associated emission of greenhouse gases. However, this kind of Intelligent Transportation System (ITS) requires collecting and modeling tremendous amount of continuous data from all road segments, in different time domains, for everyday in a year, and is a complex task. In addition, analytical decision making on optimum route planning requires high data processing and centralized computation. Data mining techniques, especially clustering, are involved to shape the unstructured data to a structural formulation and make easier decision making system for ITS problems.

Traffic flow data is used in [31] to detect the traffic status and predict the traffic patterns from historical database. Two different data mining techniques-cluster analysis and classification analysis are used in the historical data prediction model. Classified road features are used to estimate traffic flows in [32]. Functional Data Analysis (FDA) is used in [33] to analyze the daily traffic flow. A comparative study on different data mining techniques to classify traffic congestion is done in [34]. It examines J48 Decision Tree, Artificial Neural Network, Support Vector Machine (SVM), PART and K-Nearest Neighborhood to classify future traffic status and concludes J48 Decision Tree algorithm has the best performance.

In our previous works, traffic management data attributes were worked with DT (decision tree) [1] [2] [3] (Fig.1) and Neural Network (NN) [4]. NN performs better than DT. However, these works did not perform any recognized data mining or classification technique to the environmental data sets. Rather, they classified data according to the intuitive guesses. Thus, the proposed TMS is suffering from optimal data classification strategies.

There are many available methods/techniques used to classify data sets. In [12], optimal cluster numbers are determined based on the intra-cluster and inter-cluster distance measurements. They use Davies-Bouldin index and Dunn's index methods for classifying both synthetic and natural images.

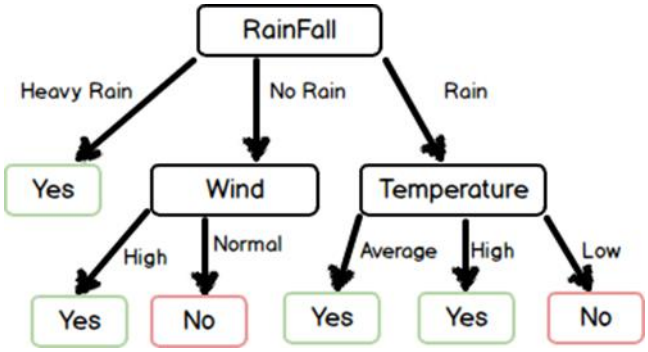


Fig. 1. Decision Tree Using ID3 Algorithm

Paper [13] evaluates the performance of three clustering algorithms (hard k-means, single linkage, and a simulated annealing) and determining the number of clusters using four methods-Davies-Bouldin index, Dunn's index, Calinski-Harabasz index and index I. Paper [14] compares three clustering algorithms- agglomerative hierarchical clustering k-means algorithm, bisecting k-means algorithm and standard k-means. Results indicate that the bisecting k-means technique performance better than other two.

In [15], authors discuss and compare the various clustering methods to find the best and fix the optimal number of clusters over three (3) structured datasets. They use three (3) different clustering algorithms- hierarchical, k-means, PAM and three (3) internal optimal clustering methods- connectivity, silhouette and dunn.

It is common and popular to apply hierarchical or partition clustering on classification problems [16]. K-means is simple, easier to implement and provide linear order complexity. Thus, partition k-means and hierarchical k-means algorithms are used to classify the TMS data sets and their optimum classification numbers are determined by three (3) different cluster validity indexes- Davies Boulder Index (DBI), Dunn Index (DI), Silhouette Coefficient (SC).

III. Classification Techniques

There are many industrial problems identified as classification problems. For examples, stock market prediction, weather forecasting, bankruptcy prediction, medical diagnosis, speech recognition, character recognitions to name a few [6-10]. Classifications are typically classified into three broad categories- supervised, unsupervised and reinforce learning [11]. Supervised learning is used when the data class label are known. Unsupervised learning (cluster analysis) is applicable on unknown class label datasets. Reinforcement learning is the problem of getting an agent to act in the world to maximize its rewards. In this paper, TMS data sets have no class label thus falls in unsupervised learning category. This section describes the algorithms and methods- those are used for clustering in this paper. Notations and their descriptions are listed in Table I.

A. Hierarchical Clustering

Hierarchical clustering constructs a hierarchy of clusters (dendrogram). Dendrogram is a process that captures whether the order in which clusters are merged (bottom-up view) or

clusters are split (top-down view). There are two variant of hierarchical clustering methods (in fig. 2.): i) Agglomerative Hierarchical clustering algorithm (HAC) or AGNES (bottom-up approaches), ii) Divisive Hierarchical clustering algorithm (HDC) or DIANA (top-down approaches). In this paper, we implement the divisive hierarchical cluster to classify the feature data, as it has less computational cost compare to AGNES. We stop our iteration when optimal clustering number is reached.

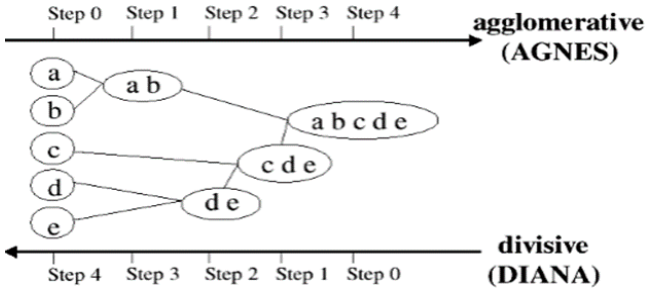


Fig. 2. Hierarchical clustering structure

1) Divisive Hierarchical Clustering Algorithm: Division Hierarchical Clustering Algorithm (HDC) or DIANA (DivisiveANAly) [17] is a variant of hierarchical clustering. It starts evaluation from the top with all data in one cluster (fig. 3) and then split using flat clustering algorithm such as k-means clustering.

Algorithm:

a. Initially all items belong to one cluster C_i=0.

b. Split C_i into sub-clusters, C_{i+1} and C_{i+2}.

c. Apply K-mean on C_{i+1} and C_{i+2}.

d. Increment the value of i.

e. Repeat steps b, c and d until the desired cluster structure is obtained.

Node 0 containing the whole data set

C_1=2 input nodes 1-2.

C_2=3 input nodes-> 2- 4 (1 spilt into 2 sub group-3 and 4).

C_3=4 input nodes ->3-6 (1 spilt into 2 sub group-5 & 6).

Do until C_{kmax}not reached where C_{kmax} is maximum number of clusters.

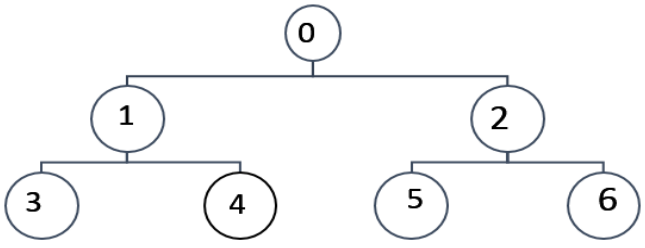


Fig. 3. Splitting node in DIANA

B. Partitional Clustering

Partitional clustering determines a flat clustering into k clusters with minimal costs. It partitions data set into k clusters and assigns the object to their nearest centers. Here (in fig. 4), k is the number of centroids.

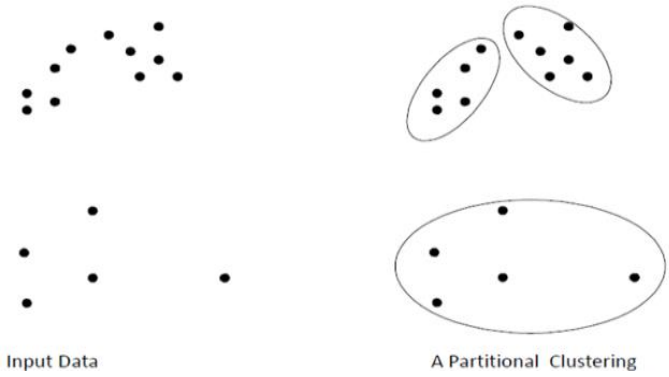


Fig. 4. Partitional clustering

1) K-means Clustering Algorithm: K-means clustering [18][19][27] aims to partition data into k clusters. K-means is the most popular non-hierarchical iterative clustering algorithm (Fig.5). The basic idea of k-means is to start with an initial partition and assign data objects to cluster so that the squared error decreases.

Algorithm:

a. Randomly initialize k center from the set of data point $\{X^d=x_1^d, x_2^d, x_3^d \ldots x_n^d\}$.

b. Assign each point to their nearest center using Manhattan distance measure.

$$\min_{0\le i<n_c}(d(x^d, v_i^d)) \quad (1)$$

c. Compute the centroid for each cluster by averaging the data objects belonging to the cluster, assign it as a new cluster center.

$$v_{inew}^d = \frac{1}{n_i}\sum_{k=1}^{n_i} d(x^d, v_{iold}^d) \quad (2)$$

d. Re-assign all the data points to its new center.

e. Repeat b, c and d steps until all the cluster centers do not change anymore otherwise stop.

TABLE. I. LIST OF SYMBOLS AND THEIR DESCRIPTION

SL No	Symbol/Notation	Description
1.	n_c	Number of total cluster
2.	C_i	i^{th}cluster
3.	d(x,y)	Manhattan distance between two data element
4.	n_i	Number of element in the i^{th} cluster
5.	v_i	Value of the center of the i^{th}cluster
6.	$d(v_i,v_j)$	Distance between two center
7.	S_i	Variance of i^{th}cluster
8.	C_{kmax}	Maximum number of cluster
9.	d	No of dimension

IV. OPTIMAL CLUSTERING METHODS

Clustering validity indexes [20][21][22][23] are usually defined by combining compactness and separability of the clusters. Compactness measures closeness of cluster elements. A common measure of compactness is variance. Separability indicates how distinct two clusters are. Basically, there are two types of validity techniques used for clustering evaluation-external criteria and internal criteria [30]. External criteria are used for categorized data clustering. No internal information is needed for internal criteria. It evaluates the quality of clusters, using only the data and without referencing to the external information. There are so many methods to measure the quality of the clustering-Davies-Bouldin index, Dunn index, CH index, Elbow method, X-means clustering, Information Criterion Approach, Information Theoretic Approach, Silhouette method, and cross-validation. The used TMS data do not have any external information and thus influences to use internal measure or criteria for clustering validation.

1) Davies-Bouldin Index: Davies Bouldin (DB) index [20][21] measures the average similarity between each cluster and its most similar one. Lower value of DB Index indicates that clusters are tight compact and well separated which reflects better clustering. The goal of this index is to achieve minimum within-cluster variance and maximum between cluster separations. It measures similarity of cluster (R_{ij}) by variance of a cluster (S_i) and separation of cluster (d_{ij}) by distance between two clusters (v_i and v_j). The formulae of DB index are-

$$S_i = \frac{1}{n_i - 1} \sum_{x \epsilon c_i} d(x, v_i)^2 \quad (3)$$

$$d_{ij} = d(v_i, v_j) \quad (4)$$

$$R_{ij} = \frac{S_i + S_j}{d_{ij}} \quad (5)$$

$$R_i = \max_{0 \le j < n_c, i \ne j} (R_{ij}) \ , i = 1 \cdots \cdots n_c R_i \ge 0 \quad (6)$$

$$DB = \frac{1}{n_c} \sum_{i=1}^{n_c} R_i \quad (7)$$

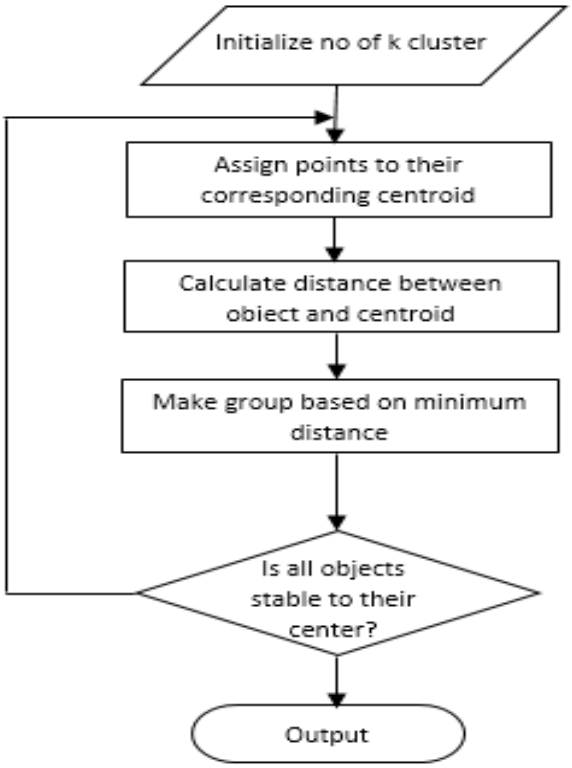


Fig. 5. Flow chart of K-mean Algorithm

2) Dunn Index: The value of Dunn index (DI) [21] is expected to large if clusters of the data set are well separated. If the dataset has compact and well-separated clusters, the distance between the clusters is expected to be larged and the diameter of the clusters is expected to be smaller. The clusters are compact and well separated by maximizing the inter-cluster distance while minimizing the intra-cluster distance. Large value of Dunn index indicates the compact and well-separated clusters. The formulae of Dunn index are-

$$D = \frac{\min_{0 \le i < n_c, 0 \le j < n_c, i \ne j} (d(C_i, C_j))}{\max_{0 \le k < n_c} (diam(C_i))} \quad (8)$$

Where,

$$d(C_i, C_j) = \min_{x \in C_i, y \in C_j} (d(x, y)) \quad (9)$$

$$diam(C_i) = \max_{x, y \in C_i} (d(x, y)) \quad (10)$$

3) Silhouette Coefficient: Silhouette Coefficient (SC) [22] [23][24] shows- how well the objects can fit within the cluster. It measures the quality of the cluster by ranging between -1 and 1. A value near to one (1) indicates that the point x is affected to the right cluster. There are two terms- cohesion and separation. Cohesion is intra clustering distance, and separation is distance between cluster centroids. A(x) is the average dissimilarity between x and all other points of its cluster. B(x) is the minimum dissimilarity x and its nearest cluster. A cluster which has a value near -1, indicates that the point should be affected to another cluster. The formulae of SC are-

$$a(x) = \frac{1}{n_i - 1}\sum_{y \in C_i, y \neq x} d(x,y) \quad (11)$$

$$b(x) = \min_{j, j \neq i}[\frac{1}{n_j}\sum_{y \in C_j} d(x,y)] \quad (12)$$

$$SC = \frac{1}{n_c} \cdot \sum_i \{\frac{1}{n_i}\sum_{x \in C_i}\{\frac{b(x)-a(x)}{\max[b(x),a(x)]}\}\} \quad (13)$$

V. Cluster Validation Methods

A. Correlation: An effective clustering algorithm needs a suitable measurement of similarity or dissimilarity. Correlation (in Fig. 6) computes the similarity matrix and incident matrix (also called occurrence matrix) to measure the correlation between the data and its cluster [25]. Higher value of correlation indicates that the points belong to the same cluster (very close to each other), and reflects good clustering. The formula of correlation is-

$$r = \frac{\sum_{i=1,j=1}^{n}(d_{ij}-\bar{d})(c_{ij}-\bar{c})}{\sqrt{\sum_{i=1,j=1}^{n}(d_{ij}-\bar{d})^2}\sqrt{\sum_{i=1,j=1}^{n}(c_{ij}-c)^2}} \quad (14)$$

Here,

r =correlation of the data and its cluster,

Distance matrix, D= $\{d_{11}, d_{22}, d_{33}, \ldots, d_{nn}\}$,

Incident matrix C= $\{c_{11}, c_{22}, c_{33}, \ldots, c_{nn}\}$,

$\bar{d}$ =mean of the distance matrix,

and $\bar{c}$ =mean of the incident matrix.

B. Distance Matrix : It is also called similarity matrix, an nxn two dimensional matrix -where n is the number of elements in a data set. d(x, y) distance or dissimilarity between objects x and y. Fig. 7 represents distance matrix.

$$d(x,y)=|x-y| \quad (15)$$

C. Incidence Matrix: An incidence matrix is a matrix that shows the relationship between two classes of objects. It is an nxn matrix where n is the total number of data set. If the object x and the object y belongs to the same cluster then Ixy=1 and if the object x and the object y belongs to the different cluster then Ixy=0.

D. Manhattan Matrix : Manhattan distance is the absolute distance between two points. Let, the objects x = $(x_1, \ldots, x_d)$ and y = $(y_1, \ldots, y_d)$ then the Manhattan distance between the two objects is,

$$d(x,y) = \sum_{i=1}^{d}|x_i - y_i| \quad (16)$$
$$where, d = dimension$$

In this work, we use Manhattan distance as a distance measurement technique.

E. Within Sum of Square Error (WSS) : WSS is also called Sum of Squared Error (SSE) [26]. Sum of Square Error (SSE) or within sum of square cluster error (WSS) is widely used for criteria measuring. The value of SSE is high, indicates high error, which means poor quality cluster. Good clustering aims for minimum value of SSE. The formula of within Sum of Square Error is-

$$SSE\ or\ WSS = \sum_{k=1}^{n_c}\sum_{x_i \in C_i} d(x_i, V_i) \quad (17)$$

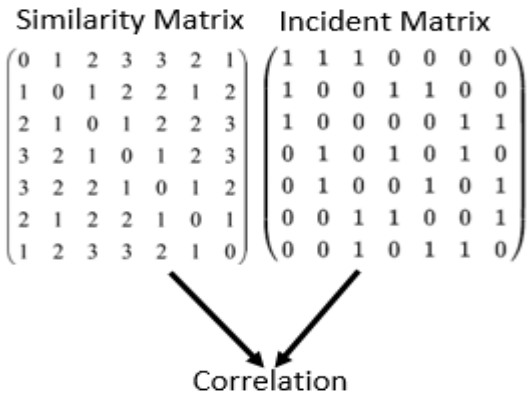


Fig. 6. Correlation

$$\begin{bmatrix} 0 & & & & & \\ d(2,1) & 0 & & & & \\ d(3,1) & d(3,2) & 0 & & & \\ \vdots & \vdots & \vdots & & & \\ d(n,1) & d(n,2) & \cdots & \cdots & 0 \end{bmatrix}$$

Fig. 7. Distance Matrix

VI. Experimental Results

Based on the above algorithms and methods, data are formulated to determine the optimal classes in each feature, road weight and verify better algorithm. Experiments in Table 2 and 3 are generated from the 750 days (1/12/2006 to 20/12/2008) collected data from [5] and presented in Fig. 8.

1) Results of Divisive Hierarchical Method: the sum of square error (SSE) of all features using divisive hierarchical cluster with Davies-Bouldin index, Dunn index and Silhouette index are presented in Table II. Shaded block (in Table II) indicates the minimum value of SSE. This table represents the optimal cluster size of each feature using three methods and also presents that Dunn index minimizes the SSE values in all cases. Thus, we conclude that Dunn index performs better for HDC to find optimal cluster. Thus, the optimal classes of each feature using HDC are – Rainfall (k=2), Temperature (k=2), Wind (k=3), Humidity (k=5) and Peak hour (k=4).

TABLE. II. OPTIMAL NUMBER OF CLUSTER AND VALUE OF SSE OF RAINFALL, TEMPERATURE, WIND, HUMIDITY AND PEAK HOUR USING HDC WITH DB, DUNN AND SC INDICES

Feature / Method	Rainfall		Temperature		Wind		Humidity		Peak hour	
	Optimal k	***SSE***	***Optimal k***	***SSE***	***Optimal k***	***SSE***	***Optimal k***	***SSE***	***Optimal k***	***SSE***
DB	2	25051.32	2	3690.40	2	9301.24	3	31152.06	3	99507.77
Dunn	2	25051.32	2	3690.40	3	4850.62	5	11231.18	4	49786.87
SC	2	25051.32	2	3690.40	3	4850.62	3	31152.06	2	183452.98
Optimal k	2		2		3		5		4	

TABLE. III. OPTIMAL NUMBER OF CLUSTER AND ITS VALUE OF SSE OF RAINFALL, TEMPERATURE, WIND, HUMIDITY AND PEAK HOUR USING K-MEAN WITH DB, DUNN AND SC INDICES

Feature / Method	Rainfall		Temperature		Wind		Humidity		Peak hour	
	Optimal k	***SSE***	***Optimal k***	***SSE***	***Optimal k***	***SSE***	***Optimal k***	***SSE***	***Optimal k***	***SSE***
DB	3	9574.97	2	3690.40	2	9301.24	3	23146.38	2	183452.98
Dunn	3	9574.97	3	2106.56	4	4657.67	6	6339.761	5	49541.539
SC	2	25051.32	2	3690.40	2	9301.24	3	23146.38	2	183452.98
Optimal k	3		3		4		6		5	

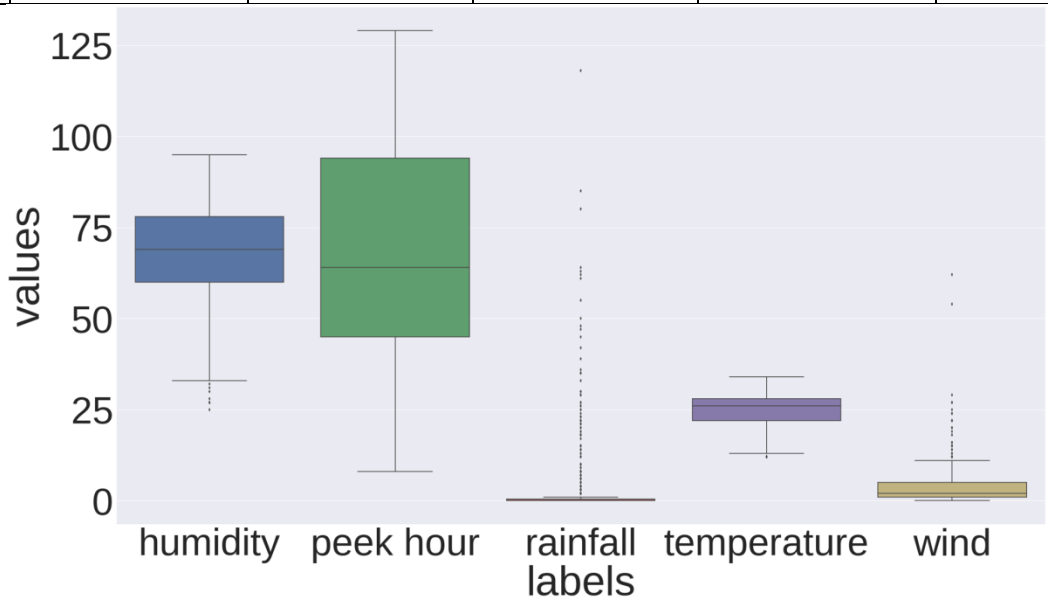


Fig. 8. Collected data from ACCU Weather[5]

TABLE. IV. COMPARISON THE CORRELATION BETWEEN TWO ALGORITHMS

Algorithm / Feature	Hierarchical		K mean	
	Optimal k cluster Using Dunn index	***Correlation***	***Optimal k cluster Using Dunn index***	***Correlation***
Rainfall	2	-0.801	3	-0.789
Temperature	2	-0.736	3	-0.721
Wind	3	-0.580	4	-0.405
Humidity	5	-0.555	6	-0.521
Peak hour	4	-0.639	5	-0.578

TABLE. V. CLUSTER SIZE AND DUNN INDEX VALUE OF THE ROAD WEIGHT

No of cluster k	**Dunn index value**
2	0.08
3	0.09
4	0.10
5	0.11
6	0.11
7	0.13
8	0.12
9	0.12

2) Result of K-means Clustering Method: The sum of square error (SSE) [26] of all features using k-means clustering algorithm with Davies Bouldin index, Dunn index and Silhouette index are presented in Table III. Shaded block (in Table III) indicates the minimum value of SSE. Table III reflects that Dunn index provides minimum value of the SSE in all features. Thus, we conclude that Dunn index performs better for k-means algorithm to find optimal cluster numbers. The optimal classes of each feature using k-means are – Rainfall (k=3), Temperature (k=3), Wind (k=4), Humidity (k=6) and Peak hour (k=5).

3) Comparison of HDC and K-means: Hierarchical clustering and K-means clustering are compared by computing the correlation on their optimal cluster numbers in each feature. It is clear from Table IV that the correlations of K-means are higher than the correlations of HDC, for all features. Thus, we conclude k-means performs better than HDC.

TABLE. VI. SAMPLE ROAD WEIGHT CLUSTERING RESULT

Data	**Rainfall**	**Temperature**	**Wind**	**Humidity**	**Peak hour**	**Road weight**
1	0	1	0	3	1	0
2	0	1	0	4	3	5
3	0	0	0	3	2	2
4	0	0	0	3	3	5
5	0	2	1	4	0	4
6	0	2	2	4	1	6
7	0	2	0	3	1	0
8	0	2	1	4	3	3
9	0	2	0	3	0	4
10	0	2	1	5	0	4

4) Optimal Cluster of Road Weight : From the previous experiments it is clear that k-means with Dunn index performs better for all features. Thus, for the classification of the road weight k-means with Dunn index can be chosen. Table V shows the no of cluster of road weight and Dunn index value of that corresponding cluster. This table represents that maximum value of Dunn index achieves in k=7. Thus, the optimal cluster size of road weight is seven (7) and there should be seven (7) different type of classes for road weight updates. Table VI presents some sample experimental results of road weight updates.

VII. CONCLUSION AND FUTURE WORKS

In this section, we summarize our work. The features data are collected from the external feeds (like web site, RSS feed, web service etc.) for classifying data. We cluster the data using two approaches (partition k-means and hierarchical k-means) and find the optimal number of clusters for each feature using Davies-Bouldin index, Dunn index and Silhouette coefficient. Thereafter, conclusion has been drawn which algorithm is better for which feature data and then find the optimal number of clusters of road weights with the input of the measured five (5) feature clusters.

In future, we can also measure validity of the classes by other probabilistic and statistical methods. Dunn index method needs lots of computational cost. Improvement on the computation cost and error of the cluster building procedure can be reduced using other statistical models. At present, we are not considering other characteristics of environmental and road status such as: accidents, road works, etc. Roads and Highway authorities in Bangladesh does not provide/publish any road construction, maintenance status and thus, these attributes will be considered in our future research direction.

Online multi data feeds capability supports the proposed model to be connected with different Social Medias (facebook, twitters etc.), and collects necessary information (mishap, disaster situations), and uses analytical tools to make proper decisions. However, special consideration is required on internet securities as all of the information is available on the internet. Recently, deep learning (DL) techniques are also used to solve unsupervised clustering problem. Interpolation of deep leaning is much complex than k-means. In addition, deep learning works with multi-layer data representation and sometimes degrades the performance due to the limited amount of data. It addresses over fitting problem also. Thus, a comparative study with simple k-means and DL is required and will be applied in near future.

Still, the proposed TMS is in construction phase and cover small road networks. City level broader area will be considered in near future. A GSM and GPS based micro controller with different embedded sensors is in developing phase. This device will help to collect real time environmental data at an instant time.

ACKNOWLEDGEMENTS

East West University (EWU), Bangladesh is gratefully acknowledged for providing article processing charge (APC) to cover the costs of publication.

REFERENCES

[1] Rahman, M. R. and Akhter, S. 2015. Real time bi-directional traffic management support system with GPS and websocket, Proc. of the 15th IEEE International Conference on Computer and Information Technology (CIT-2015). Liverpool. UK. 26-28 Oct. 2015.

[2] Rahman, M. R. and Akhter, S. 2015. Bidirectional traffic management with multiple data feeds for dynamic route computation and prediction system. *International Journal of Intelligent Computing Research (IJICR)*. Special Issue. Volume 7. Issue 2. ISSN: 2042 4655. Mar/2015. http://infonomics-society.ie/ijicr/

[3] Rahman, M. R. and Akhter, S. 2015. Bi-directional traffic management support system with decision tree based dynamic routing, *Proc. of 10th International Conference for Internet Technology and Secured Transactions, ICITST 2015.* London. United Kingdom. December 14-16. 2015.

[4] Akhter,S. Rahman, M.R., and Islam M. A. 2016. Neural Network (NN) based route computation for bi-directional traffic management system. *International Journal of Applied Evolutionary Computation- special issue on Emerging Research Trend in Computing and Communication Technologies.* Volume 7. Issue 4.

[5] AccuWeather. (2016, March 7). Retrieved from http://www.accuweather.com/en/bd/dhaka/28143/january-weather/28143?monyr=1%2F1%2F2016&view=table

[6] Moghadassi, F. Parvizian, and S. Hosseini. 2009. A new approach based on artificial neural networks for prediction of high pressure vapor-liquid equilibrium. *Australian Journal of Basic and Applied Sciences*. Vol. 3. No. 3. pp. 1851-1862. 2009.

[7] Asadi, R., Mustapha, N., Sulaiman, N. and Shiri, N. 2009. New supervised multi layer feed forward neural network model to accelerate classification with high accuracy. *European Journal of Scientific Research*. Vol. 33. No. 1. 2009. pp.163-178.

[8] Pelliccioni, R. Cotroneo, and F. Pung 2010. Optimization of neural net training using patterns selected by cluster analysis: a case-study of ozone prediction level", *8th Conference on Artificial Intelligence and its Applications to the Environmental Sciences*, 2010.

[9] Kayri, M. and Çokluk, Ö. 2010. Using multinomial logistic regression analysis In artificial neural network: an application, *Ozean Journal of Applied Sciences*. Vol. 3. No. 2. 2010.

[10] Khan, U., Bandopadhyaya, T. K. and Sharma, S. 2009. Classification of Stocks Using Self Organizing Map. *International Journal of Soft Computing Applications*. Issue 4. pp.19-24.

[11] Hagan, M.T., Demuth, H.B., and Beale, M. 1996. *Neural Network Design.*PWS publishing company, Boston, Massachusetts.

[12] Ray, S., and Turi, R., H. 1999. Determination of number of clusters in K-means clustering and application in color image segmentation, *Published in the 4th International Conference on Advances in Pattern Recognition and Digital Techniques*. Pg. 137-143. Culcutta, India.

[13] Maulik, U., and Bandyopadhyay, S., "Performance Evaluation of Some Clustering Algorithms and Validity Indices", Published in Journal IEEE Transactions on Pattern Analysis and Machine Intelligence archive Volume 24 Issue 12, Pg.1650-1654, December 2002.

[14] Steinbach, M., Karypis, G., and Kumar, V., "A Comparison of Document Clustering Techniques", Published in the 6th ACM SIGKDD, World Text Mining Conference, (2000).

[15] Begum, S., F., Kaliyamurthie, K., P., and Rajesh, A., "Comparative Study of Clustering Methods over Ill-Structured Datasets using Validity Indices", published in Indian Journal of Science and Technology, Vol 9(12), March 2016.

[16] Reddy, C.K. and Vinzamuri, B., "A Survey of Partitional and Hierarchical Clustering Algorithms", Data Clustering: Algorithms and Applications. CRC Press 2014, ISBN 978-1-46-655821-2 (pg.88-91).

[17] Hierarchical-clustering-algorithm available at: https://sites.google.com/site/dataclusteringalgorithms/hierarchical-clustering-algorithm,15-5-2016.

[18] Ray, S., and Turi, R., H., "Determination of Number of Clusters in K-Means Clustering and Application in Colour Image Segmentation", Published in the 4th International Conference on Advances in Pattern Recognition and Digital Techniques (Pg. 137-143).

[19] K-mean Algorithm Available at: http://www.academia.edu/3438357/K_Means_Algorithm_Example,12-4-2016.

[20] Kovács, F., Legány, C. and Babos, A., "Cluster Validity Measurement Techniques" ,AIKED'06 Proceedings of the 5th WSEAS International Conference on Artificial Intelligence, Knowledge Engineering and Data Bases (2006), ISBN:111-2222-33-9 (pg. 388-393).

[21] Davies, D.L. and Bouldin, D.W., "A Cluster Separation Measure", IEEE Transactions on Pattern Analysis and Machine Intelligence, 1(2), 224–227.

[22] Dunn, J.C., "A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters," J. Cybernetics, vol. 3, 1973, (pg. 32-57).

[23] Liu, Y., Li, Z., Xiong, H., Gao, X., and Wu, J.,"Understanding of Internal Clustering Validation Measures", Proceeding ICDM '10 Proceedings of the 2010 IEEE International Conference on Data Mining ISBN: 978-0-7695-4256-0 (Pg. 911-916).

[24] Zoubi, M., and Rawi, M., "An efficient approach for computing silhouette coefficients," Journal of Computer Science 4,(2008) (pg.252–255).

[25] Correlation Definition available at:
www.cse.buffalo.edu/~jing/cse601/fa12/materials/clustering_basics.pdf, 23-4-2016.

[26] Sum of square error at:https://hlab.stanford.edu/brian/error_sum_of_squares.html,15-5-2016.

[27] Pelleg, D., Moore, A.W., "X-means: Extending K-means with Efficient Estimation of the Number of Clusters", Proceeding ICML '00

Proceedings of the Seventeenth International Conference on Machine Learning, ISBN: 1-55860-707-2 (Pg.727-734).

[28] ID3 Decision Tree Algorithm - Part 1 at: http://www.codeproject.com/Articles/259241/ID-Decision-Tree-Algorithm-Part

[29] https://datajobs.com/data-science-repo/Decision-Trees-[Rokach-and-Maimon].pdf

[30] Liu,Y., Li,Z., Xiong,H., Gao,X.,Wu,J., "Understanding of Internal Clustering Validation Measures", Proc. of IEEE International Conference on Data Mining,2010 http://datamining.rutgers.edu/publication/internalmeasures.pdf

[31] Wang, Y., Chen, Y., Qin, M. and Zhu, Y., "Dynamic traffic prediction based on traffic flow mining," in Proceedings of the 6th World Congress on Intelligent Control and Automation (WCICA 0'6), vol. 2, pp. 6078–6081, Dalian, China, June 2006.

[32] Caceres, N., Romero, L. M. and Benitez, F. G., "Estimating traffic flow profiles according to a relative attractiveness factor", Procedia—Social and Behavioral Sciences, vol. 54, pp. 1115–1124, 2012.

[33] Guardiola, I. G., Leon, T. and Mallor, F., "A functional approach to monitor and recognize patterns of daily traffic profiles", Transportation Research, Part B: Methodological, vol. 65, pp. 119–136, 2014.

[34] Mirakhorli, A., "A Comparative Study: Utilizing Data Mining Techniques to Classify Traffic Congestion Status", UNLV Theses, Dissertations, Professional Papers, and Capstones. Paper 2197.

An Extensive Survey over Traffic Management/Load Balance in Cloud Computing

Amith Shekhar C
Asst. Professor
Dept. of IS&E
GM Institute of Technology, Davangere,
Karnataka, India

Dr. Sharvani. G S
Associate Professor
Dept. of CS&E
RV College of Engineering, Bengaluru,
Karnataka, India

***Abstract*—Cloud Computing (CC) is all about carrying out processing in other's system. There are various vendors who provide CC services. The basic algorithm that should be met to access CC services is a need for steady internet connection. As everything is done online the traffic across the internet is to be managed efficiently so that the transmission delay can be minimized and better quality of service can be given to the customers. The network should not be too congested at any moment of time. Hence the traffic management becomes a crucial factor for the better performance of the CC network. This paper addressed the most valuable terms and topics concerning the load balance/traffic management in cloud computing. Also, the paper is meant to discuss the study analysis of the recent researches in load balance of CC. From the study analysis the current research gap is addressed with a future scope of research to overcome the research gap.**

Keywords—Cloud Computing; Load Balance; Traffic Management

I. INTRODUCTION

The quick advancement of storage and processing technologies and the accomplishment of the Internet, computing resources have turned out to be less expensive, more intense and more pervasively accessible than any time in recent time. This technological pattern has empowered the acknowledgment of another computing model called Cloud Computing (CC), in which resources (e.g., storage and CPU) are given on rented use and discharged by clients through the Internet in an on-demand mold. In a CC environment, the conventional part of service supplier is partitioned into two: the infrastructure vendors who oversee cloud stages and rent resources as per a utilization based estimating the model, and service providers, who lease resources from one or numerous infrastructure suppliers to serve the end clients. The rise of CC has had a huge effect on the Information Technology (IT) industry. In recent years, where the organizations, for example, Amazon, Microsoft, and Google are using CC to give all the more powerful, stable and cost-proficient cloud stages, and business ventures try to reshape their business models to pick up advantage from this paradigm [1 2].

The load in a cloud is be founded on memory required, CPU (Virtual Machine) limit which will accommodate for finishing of client job. Cloud is an innovation given distributed environment so sharing of work among various resource supportive to enhance the use of comprehensive resources and achieve great execution. Also, load balancing is a procedure of guaranteeing the consistent balance of workload on the pool framework hub or processor. Load balancing mechanism predominantly characterized in two types: Dynamic and Static balancing. In static balancing earlier availability of resources is required, so moving of load not relies on upon current condition of resources. This balancing helps in handling low variation load. In Dynamic balancing is done as per the load variation. For this ongoing communication with the network is required which can build the traffic all through a network. Thus, like static, element additionally check the present condition of resource it is possible that they are loaded [3].

The current traffic analysis and management techniques can't be effectively reached out to the data centers. The thickness of connections at the data centers is much higher than the thickness of connections at the undertaking network. This presents the issues for the current techniques to be connected to gauge traffic over the CC system. A significant portion of the current strategies is utilized for the hard level of traffic management. These techniques are fit for investigating traffic over the higher number of hosts. For a particular server may have a few thousand servers; this is the place the current techniques fails. Since these strategies are intended for the hard networks, that accepts the stream designs of sensible in the Internet and undertaking networks. Also, with regards to the cloud system, it is unrealistic to allow under traffic variation. The frameworks conveyed on cloud stage should be adaptable since the information traffic over the framework may change at any time [4 5]. The traffic management choices are frequently made in a unified way. This prompts to high multifaceted nature and poor scalability. This paper presents the essential aspects of the load balancing/traffic management in the cloud and also highlights survey of the existing researches in load balancing along with the research gap. Also, significant future line of research is addressed.

The section wise discussion of this paper is categorized as Section II - conceptual description of Cloud computing along with architecture, business model, and cloud types. In section III the existing methods for traffic management are mentioned, then recent research survey of load balance/traffic management are discussed in section IV. The research gap in the existing research is addressed in section V while the future study analysis is given in section VI. Finally, the conclusion is discussed in section VII.

II. BASIC CONCEPTS OF CC

The evolution at the end of the twentieth century to the present day facilitates of pervasive computing; the web has changed the computing. It has gone from the idea of parallel computing to distributed computing to matrix computing and as of late to CC. Also, the possibility of CC has been around since long ago; it is a developing field of software engineering. CC can be characterized as a computing environment where computing needs by one gathering can be outsourced to another gathering, and when need emerges to utilize the computing force or resources like database or messages through the web [6]. CC is a late pattern in IT that moves computing and information far from desktop and compact PCs into substantial data centers. The primarily favorable position of CC is that clients don't need to pay for infrastructure, its establishment, required labor to handle such infrastructure and support.

A. Cloud Computing

In recent years, CC accomplished milestones of turning point past the conviction. The existing researches will supports to increasing cloud issues. In this realized that cloud is an interconnected fast network which permits incredible facility such as elasticity, on-demand resource provisioning, usage of resource given ordering, pays for what you require. Achieving proper resource balance for fulfillment of client's requirement is main task in cloud services. In CC, clients get to the information, applications or whatever other services with the assistance of a browser to the devices utilized and the client's area. The Cost is lessened to an exceptional level as the infrastructure and need not be obtained for periodically escalated computing issues [7].

CC gives a few striking elements that are unique about conventional service computing, which is given below:

- *The Multi-Tenancy:* In a cloud environment, services possessed by numerous suppliers are co-situated in a data center. The execution and management issues of these services are shared among service suppliers and the infrastructure supplier. The layered engineering of CC gives a characteristic division of responsibilities: the proprietor of every layer just needs to concentrate on the particular functionalities connected with this layer.
- *Resources pooling:* - The infrastructure supplier offers a pool of computing resources that can be powerfully assigned out to numerous resource users. Such dynamic resource task ability gives much adaptability to infrastructure suppliers for dealing with their particular resource use and working costs. For example, an Infrastructure as a Service (IaaS) provider can influence VM relocation mechanism to accomplish a high level of server combination, henceforth expanding resource usage while minimizing cost, for example, cooling and power utilization.
- *Network Access and Geo-Distribution:* - Clouds are by and large open through the Internet and utilize the Internet as a service conveyance network. Consequently, any devices with Internet network like cell phone, a PDA or a tablet can get to cloud services. Also, to accomplish high network execution and localization, a large number of today's clouds comprise of data centers situated at numerous areas around the world. A service supplier can without much of a stretch influence geo-distribution to accomplish greatest service utility.
- *Service Based:* - The CC receives an operation model of service. Subsequently, it puts a solid accentuation on service management. In a cloud, every Infrastructure as a Service (IaaS), Platform as a Service (PaaS) and Software as a Service (SaaS) supplier offers its service as indicated by the Service Level Agreement (SLA) consulted with its customers. The SLA assures service with basic target of each supplier.
- *Provisioning of Dynamic Resources:* - This is the main point in CC is that computing resources can be acquired. In comparison with the conventional model that arrangement resource as per provision of dynamic resources, peak demands permits service providers to obtain resources given the present demand, which can extensively bring down the operating cost.
- *Utility-based Valuing:* - CC utilizes a usage based estimating model. The correct valuing plan may change from service to service. For instance, a SaaS supplier may lease a virtual machine from an IaaS provider on every hour premise. Then again, a SaaS provider that gives on-demand customer relationship management (CRM) may charge its customers given the quantity of customers it serves. Utility-based evaluating brings down service working cost as it charges customers on a for every utilization premise.
- *Self-Organization:*- The resources can be allocated and de-allocated based on its demand; the service suppliers are engaged in dealing with their resource utilization as per their needs. Besides, the automatic management of resources highlight yields high agility that empowers service provider to react rapidly to fast changes in service demand.

B. Architecture of CC

The architectural diagram of a CC environment can be separated into four layers: the equipment/server farm layer, the infrastructure layer, the stage layer and the application layer, as appeared in Fig. 1.

- *Infrastructure Layer (IL):*- This layer is also virtualization layer (VL), the IL makes a pool of storage and computing resources by dividing the physical resources utilizing virtualization technologies, for example, Xen and VMware. The IL is a basic segment of CC, since many key components, for example, dynamic resources, are just made accessible through virtualization technologies.
- *Platform Layer (PL):*- This is kept above the IL; the PL comprises of working frameworks and application systems. The reason for the PL is to minimize the burden of conveying applications specifically into VM

units. For instance, Google App Engine works at the PL to give API support to actualizing database, business and storage logic to execute web applications.

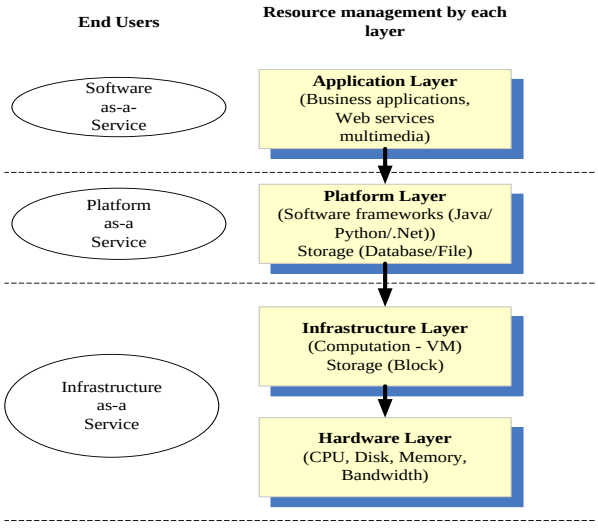


Fig. 1. CC Architecture

- *Hardware Layer (HL):-* This layer is used to manage the physical resources for the cloud, also includes routers, physical servers, cooling and power frameworks. Practically speaking, the HL is normally executed in data centers (DC). A DC more often than not contains a huge number of servers that are sorted out in racks and interconnected through routers, switches. The issues related with HL incorporate hardware setup, adaptation to non-critical failure (fault tolerance), cooling resource and traffic management.
- *Application Layer (AL):-* The significance of AL in a high-end hierarchy is that AL comprises of the real cloud applications. The cloud applications can influence the programmed scaling highlight to accomplish better execution, accessibility, and lower working expense.

In comparison with the conventional allocations, the cloud-based scaling can offer better performance and low operating cost. The design particularity permits CC to support an extensive variety of utilization necessities while decreasing maintenance and management overhead.

C. Business model of CC

The CC utilizes a service based business model or HL and Pl resources are given as services on an on-request premise. The each layer of the design depicted in the above unit be executed as a support of the layer above. The client perspective of every layer is discussed as Software as service (SaaS), Infrastructure as a service (IaaS), and platform as service (PaaS).

- *Infrastructure as a Service (IaaS):-* This is an on-request provisioning of infrastructural resources like Virtual machines(VMs). The cloud service provider can offer IaaS. For example Amazon, Flexiscale and GoGrid.
- *Platform as a Service:-* This is a PL resource, composed of operating system support and programming improvement. Example: Google App Engine, Force.com, and Microsoft Windows Azure.
- *Software as a Service:-* This offers service to on-demand applications by using the Internet. Example: SAP Business ByDesign, Rackspace, and Salesforce.com.

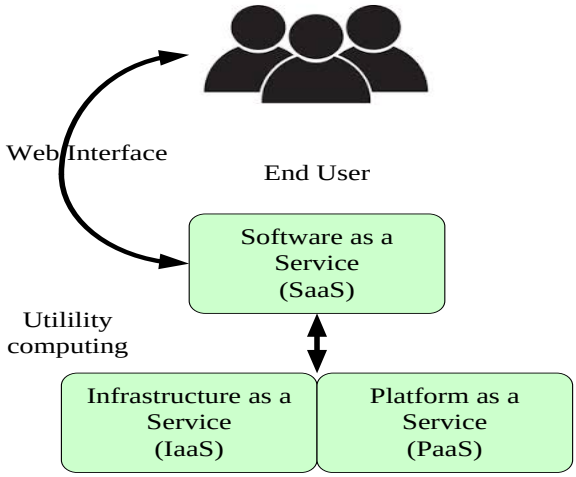


Fig. 2. Business Model of CC

The business model of CC is represented in Fig. 2. As indicated by the layered design of CC, it is totally conceivable that a PaaS provider runs its cloud over IaaS provider's cloud. Presently, the IaaS and PaaS providers are frequently parts of a similar association (Salesforce and Google). This is the reason PaaS and IaaS suppliers are regularly called the cloud or infrastructure suppliers.

D. Types of Cloud

In the cloud, there exist some issues moving an enterprise application to the cloud environment. Hence service providers are focusing on minimization of operation cost, while others may need high reliability and security. Thus, there are different types of clouds, each with its benefits and drawbacks:

- *Public Cloud (PC):-* A cloud in which specialist organizations offer their resources as the service platform to the world. The PC offers a few key advantages to providers, including no initial investment for infrastructure and moving of risks to the infrastructure provider. The control of data, system and security settings of PC can hamper their viability in numerous business situations.
- *Private Cloud (PrC):-* This is an internal cloud are intended for restrictive use within the single organization. A PrC might be constructed and managed with the by the external providers and cloud. A PrC offers the most astounding level of control over execution, security, and reliability. The frequently used

conventional server doesn't offer significances like low cost.

- *Hybrid Cloud (HC):-* An HC is a combination of PC and PrC that tries to address the issues of every approach. In an HC, unit of infrastructure service keeps running in PrC while the rest of the part keeps running out on the PC. The HC offers more adaptability than both PC and PrC. HC give more control and security over application information than PC, while as yet encouraging on-request service contraction and expansion. The drawback is that designing of HC requires precisely deciding the best split amongst PC and PrC parts.
- *Virtual Private Cloud (VPrC):-* This is a PrC which solves the limitations of PC and PrC. A VPrC is a platform running on top of PC. The principle change is that a VPrC influences virtual private system (VPN) mechanism that permits service providers to develop an own particular topology and security settings like firewall rules. VrPC is a more all holistic design since it virtualizes servers and applications, as well as the basic applications too. In most of the organizations, VPrC gives consistent service.

For most service-based organizations, selecting the correct cloud model is reliant on the business situation. In the scientific computation, applications are placed on PC for cost-effectiveness. Apparently, certain clouds will be more significant than others.

III. Existing Methods for Traffic Management in Cloud

- The data analysis in the data center is more needful for the current cloud environment. Many web applications depend on data traffic analysis to offer client data optimization [8]. The current network operators need to have an understanding of data traffic in a network so that proper data management can be followed to avoid traffic. Currently, there are a few difficulties for existing data traffic analysis and management strategies in Internet Service Providers (ISPs) systems and organization to reach out to web data centers.
- The links density is much higher than the density of ISPs or undertaking systems, which puts forth the most exceedingly bad situation for existing techniques.
- Most of the existing techniques can figure traffic between hundreds of end server hosts, yet even a flexible data centers can have a few thousand servers.
- The existing methods more often than not expect some flow pattern based designs that are sensible in the internet and organizational network systems, yet the applications conveyed on data centers, i.e., use of Map Reduce permanently change the traffic flow pattern.
- There is more tightly coupling in application's utilization of the network, storage resources, and computing than what is seen in different settings.

Right now, there is very little work on traffic analysis and management. Some of them are addressed below.

E. Virtual Private Network (VPN):

The simplest technique in which an organization can get to a CC application is through the Internet/VPN. It needs a little data synchronization between the cloud have and the undertaking data centers. The utilization of VPN puts just a little effect on the network enterprise, but it can bring about technical aspects that ought to be settled in both specialized as far as the agreement with the CC provider. The VPN is very complex to configure, and its maintenance is costly. The principle drawbacks of utilizing the VPN on the cloud for traffic service are in the cloud VPNs are difficult to scale, nearby VPN customers won't work the vast majority of the circumstances, records like to what extent the association was utilized, who got to your server are difficult to keep up. The VPN, however, is the simplest technique for an endeavor to get to the cloud organize.

F. Microsoft Azure Traffic Manager

This is an organizational Azure cloud service, or website might keep running on various data centers over the world. To control the traffic distribution to organizations predefined endpoints one can make utilization of the Traffic Manager. The Traffic Manager applies an insightful approach motor to DNS questions for the area name of your Internet resources. The arrangement of Traffic Manager enhances the accessibility of basic applications, responsiveness for performance applications. It permits overhauling and performing QoS maintenance without downtime. The execution of an extensive complex framework is upgraded by traffic distribution.

G. Cloud Network Management (CNM) Model

The CNM model demonstrates utilizes an arrangement of specialists sending redesigns to the administrators in the cloud about their execution. Every operator contains an arrangement of articles called Management Information Base (MIB) that stores the execution and other related data. To give better QoS of cloud services the system director ought to know about the present status of the chief in the group, their CPU, stockpiling and system usage, what numbers of examples of a virtual machine are assigned, etc. With a specific end goal to guarantee appropriate service every one of the messages in the CNM model is recognized. The CNM model is a half of unified and decentralized service. The CNM model is a half and half of brought together and decentralized approach. The provisos in the SNMP demonstrate prompt to the improvement of CNM model. It evacuates or minimizes a couple of pitfalls of the previous model. It gives upgraded security than the SNMP display [9].

CNM model display upgrades the system execution: utilization of less number of bundles decreases the jitter it which like this improves the system execution.

- The arrange traffic is diminished
- There are no surveying issues.
- Ensures the safe correspondence

- The concept of virtualization aides in the quicker recuperation in the setting of disappointments.
- It accommodates better security.

A system is always breaking down in light of the parameters, for example, execution, security, speed, adaptability. The execution of a cloud system ought to be broken down from both specialist co-ops' view furthermore from the client's viewpoint. Specialist organization's view point: The expert organization might be more worried about the foundation execution of the cloud arrange. Cloud expert organization needs to screen the capacity, VMs, and system traffic. Client's perspective: The clients have a tendency to choose the system execution on the premise of how the applications function; the speed of the system, availability of the remote information, and unwavering quality. In the wake of concentrate the above apparatuses, unmistakably dynamic traffic service is productive and a savvy decision for traffic service related issues over the distributed computing systems than the equipment arrangements.

H. VeriSign Traffic Management (VTM) Services

The Dynamic Traffic Management (DTM) benefit makes it conceivable to deal with the traffic in the system given continuous data. It permits for all intents and purposes boundless routes for manage based customization of association's traffic. VeriSign offers traffic service units, for example, failover, geo-allocation, weighted load adjusting, DTM. The Lua scripting dialect is solely accessible as a part of the VTM. Verisign gives dynamic traffic administration benefits over the cloud arrange. The worldwide associations can screen their system traffic effortlessly, the traffic examples can be checked, and generally in light of the fact that it is alert, the application downtime can stay away from by and large. This is a product apparatus to oversee traffic over the system.

- Low cost: When contrasted with equipment arrangements, it offers to bring down working expenses.
- Since it is conveyed all-inclusive, it permits simple versatility
- As it is a product bundle, it can be effortlessly transmitted
- It is reasonable for basic web-based services as a result of its upgraded accessibility and execution.
- Speed: the information can be conveyed to the goal much quicker as it offers the least dormancy or slack time.

The traffic over the cloud system is not the same at constantly, it differs. For instance, traffic over the VPN to get to the cloud may be diverse at various times. At the point when the traffic is less, it won't bring on any issues however if the traffic over the system is all the more; then, the system gets to be distinctly congested, and the odds of utilization downtime are high because of bottlenecks. In the distributed computing situation downtime is not acknowledged. The applications ought to run easily constantly [10]. Henceforth traffic management assumes an essential part. The system ought to have the capacity to scale up and downsize contingent upon the necessities. The hardware arrangements get to be distinctly costly because when the traffic low.

IV. SURVEY OF EXISTING RESEARCHES

In this section, various existing researches towards the traffic management in CC and load balance.

The idea of Value of Service (VoS) based task scheduling for a CC system was represented in Tunc et al. [11]. The author presented another time-based value metric to empower scheduling algorithms considers the arrival task time and also task completion value and power utilization for a given period. Authors have analyzed the other existing scheduling algorithms of VoS. In this, an author has focused to implement proposed new time-of-use VoS metric, actualize it to work in a real-time framework, and build real VMs to execute benchmark applications. Authors have analyzed the proposed VoS system framework performance and out forms significant results.

Exploiting Geo-Distributed Clouds for an E-Health Monitoring System with Minimum Service Delay and Privacy Preservation is presented in Shen et al. [12]. In the framework, the schemes of resource allocation empower the distributed cloud servers to agreeably allocate the servers to the asked for clients under the load adjust condition. In this manner, the service delay for customers is minimized. The traffic shaping (proposed) algorithm changes over the client health and non-health data traffic that the ability of traffic analysis is to a great extent reduced. Authors demonstrate the proposed traffic shaping algorithms effectiveness as far as security maximization and minimization of service delay.

The work of Abranches and Solis [13] portrayed an algorithm based on response time and traffic demands to scale containers on a CC System. The presented algorithm depends on the web requests characterization and a PID (Proportional - Integral-Derivative) controller. The proposed mechanism was evaluated with a continuous arrangement got from an operational huge web framework in a controlled infrastructure. The outcomes demonstrate that the proposition accomplishes the normal response times allocating a lower container than other works.

The idea for versatile traffic management over the cloud data centers was presented in Assi et al. [14]. This novel innovation brings new difficulties, generally in the conventions that oversee its hidden challenges. The Traffic designing in cloud data centers is the major difficulties that have pulled in consideration from the research group, especially since the legacy protocols utilized in data centers offer constrained and un-versatile traffic management. Many supported for the utilization of VLANs as an approach to give adaptable traffic management; be that as it may, finding the ideal traffic split among the VLANs is the outstanding NP-Complete VLAN task issue. The extent of the hunting space of the VLAN assignment issue is tremendous, notwithstanding for little size systems.

In Brindha et al. [15] Agent-Based Bidirectional Bidding Mechanism for Efficient Scheduling of Real-Time Tasks in CC. In this a bidirectional declaration based bidding system to allow tasks and resources progressively. Likewise, it comprises

of three stages, i.e., basic matching, forward declaration bidding and in reverse declaration bidding stage. The scalability is fundamental when data scheduling is performed in progressively including VMs which gives flexibility. Authors have planned computation rules for both forward and in reverse declaration bidding system which helps for selecting contractors. The analysis is performed for both Google cloud and synthetic workload. The examination comes about demonstrate that agent-based scheduling that gives better execution contrasted with existing strategies.

In the work of Doyle et al. [16] Load Balancing the Cloud for Carbon Emissions Control (Stratus framework) is examined. The work analyzes the power cost, carbon discharges, and normal service request for time for an assortment of situations. The choice concerning how to adjust the different components will rely on upon SLAs, government legislation, and the cost of carbon on exchanging plans. Utilizing this data and the specifics of the cloud the administrator can run the cloud in the most attractive mold. The way of the service will figure out whether a cloud provider can execute this calculation while fitting in with service level understandings.

An enhanced Genetic algorithm utilizing populace decrease for load adjusting in CC is exhibited in Patel et al. [17]. Essentially cloud depends on utilize standard pay situation recognized by client's services. Be that as it may, for every single fulfilling, that services cloud needs some predefine necessity conditions to take after which influence distinctive parameters like reaction time, resource usage, adjusting load, ordering of resources and also employments and so on. Authors work concentrated on the use of resources and reaction time given the genetic algorithm, yet Authors changed that genetic algorithm with the assistance of fractional populace decrease technique that will fulfill the demand of client services.

A work of Ran et al. [18] has said the balancing backhaul stack in different cloud radio get to organize. This work proposed to adjust the information to be transmitted on backhauls of the remote radio heads (RRHs) to diminish the data transfer capacity of backhaul required by the RRHs or accomplish better execution with given backhaul limit uniquely in contrast to pressure strategies.

The work concentrating on work process scheduling for multi-inhabitant CC situations was introduced in Rimal et al. [19].Multi-tenure is one of the key components of distributed computing, which gives adaptability and economic advantages to the end-clients and a group of service providers by having a similar cloud stage and its basic foundation with the disengagement of shared system and register resources. Be that as it may, resource service with regards to multi-inhabitant distributed computing is getting to be distinctly a standout amongst the most complex undertaking because of the intrinsic heterogeneity and resource isolation. The proposed calculation is contrasted and the best in class calculations, i.e., First Come First Served (FCFS), EASY Backfilling, and Minimum Completion Time (MCT) planning strategies to assess the execution. Facilitate, a proof-of-idea analysis of genuine logical work process applications is performed to show the versatility of the CWSA, which confirms the adequacy of the proposed arrangement.

An intriguing Dynamic Fault-Tolerant Scheduling (DFTS) Mechanism for Real-Time Tasks in CC is presented in Soniya et al. [20]. In the current framework, Primary Backup (PB) model is utilized, yet it doesn't contain any dynamic resource assigning instrument. Authors propose a dynamic resource allotting system with adaptation to internal failure to enhance resource use. Authors fuse a reinforcement covering system and proficient VM relocation technique for outlining novel mechanism perform in distributed computing. The proposed show goes for accomplishing both adaptations to non-critical failure and high resource usage in the cloud. The analysis was analyzed utilizing irregular engineered workload, and Google cloud follow logs to test the effectiveness of the proposed model.

In Vascak et al. [21] an Agent-Based CC Systems for Traffic Management is delineated. The requirement for a safe and financially productive traffic represents a test for making traffic service frameworks. This work manages to combine three fundamental ideas, in particular, cloud-based advancements, operator based methodologies and fuzzy based cognitive maps to tackle the limitation. The proposed framework was straightforwardly tried on a play area, and acquired outcomes were investigated using a few chose the algorithm. At last, some further conceivable outcomes of potential use and future research are specified.

A critical work of Sundar Rajan et al. [22] presented a Workflow Scheduling in CC Environment utilizing Firefly Algorithm. The work scheduling mechanism plays a key part in getting most extreme advantage from the resources that are given. Another critical component to being considered about distributed computing is Load adjusting. This controlling of fill guarantees that each elite machine does the extremely same measure of work at any quick of time. To ensure this, we need to prescribe on utilizing fill controlling. Here in this archive, author prescribes heuristic algorithm known as Firefly algorithm for compelling fill controlling in thinking handling. This foundation depends on the travel conduct of the fireflies which go searching for the nearest conceivable greatest choices. Author utilize Firefly algorithm to plan the occupations and subsequently uniformly disseminate the heap and like this diminish the general finish time.

A Traffic-Aware Task Allocation for Cooperative Execution in Mobile CC is said in Wang et al. [23]. In this author, designs a task allocation mechanism to the mobiles with the objective that the aggregate traffic brought is least, while the imperative of the resources on mobile is not damaged. Given that, author additionally intends to minimize the aggregate number of mobiles took an interest in undertaking execution. A dynamic programming is proposed to finish the over two destinations separately. The simulation comes about exhibit the adequacy of the proposed plots in minimizing the traffic and the quantity of required mobile devices.

A work of Zhang et al. [24] has an effect on The Modeling of Big Traffic Data Processing Based on CC. The progress of research for data collection mechanism and huge data

processing model has a vital commitment to the improvement of urban traffic information. The control of urban transportation relies on upon the successful handling of continuous traffic perception information, which is typically the way of information escalated. The author studied an extensive number of floating car data processing (FCD) traffic control in the distributed computing environment, to take care of the issue of the new distributed computing innovation to take care of the issue of urban traffic control framework. The research comes about demonstrate that the distributed computing innovation, for example, HBase and Map Reduce can give huge information stream figuring of generous utility, for example, adaptability and ongoing count execution can be precisely by the proposed data storage, service, and parallel preparing model. The appropriateness and practicability of distributed computing are assessed for two runs of the mill information figuring errands of urban traffic checking, in particular, FCD question, FCD outline.

V. Existing Research GAP

The existing traffic measurement and analysis methods cannot be easily extended to the data centers. The density of links at the data centers is much higher than the density of links at the enterprise network. This makes the worst case scenario for the existing methods to be applied to measure traffic across the CC network. Most of the existing methods are used for the enterprise level network data traffic management. These methods are capable of analyzing traffic across few hundred hosts. But a modular data center may have several thousand servers; this is where the existing methods fail. Since these methods are designed for the enterprise networks, they assume the flow patterns that are reasonable in the Internet and enterprise networks. But when it comes to cloud network it is not possible to assume a pattern because the data traffic will be varying. The variation is not linear nor is it predictable. The systems deployed on cloud platform needs to be scalable since the data traffic across the system may change anytime. The traffic management decisions are often made in a centralized manner. This leads to high complexity and poor scalability.

The existing system is witnessed with the abundant literature for load balancing techniques over cloud environment. The majority of the techniques are found to adopt an approach that uses highly complex and sophisticated design where various probabilities of normalizing traffic congestion were not focused effectively. Hence, there is a need of adopting a technique that can cost-effectively lowered down the overheads of the traffic using multi-tenancy technique, efficient scheduling, and distributed, etc.

VI. Continuation of Future Research

The further research can be followed to solve the existing issues of the traffic management in CC. In future, the work can be extended as:

- A review of research can be done, and the significant standard techniques of load balancing deployed in the existing system and explore its research gap.
- To design a novel and a simple algorithm of multi-cloud tenancy for the purpose of redirecting the traffic based on its requirement to reduce the massive load of task scheduling is need to be analyzed.
- Also, develop a novel mathematical model for addressing the routing and scheduling issues considering virtual machines over cloud are too considered.
- A framework with a novel class of algorithms for distributed load balancing system over a large number of a cloud environment for ensuring resource availability is needed to design.
- The performance of the designed system can be measured with the existing system of load balancing.

VII. Conclusion

CC is a platform that is cost-effective and also provides faster means of data transmission. Hence traffic management across the CC network becomes critical. In order to reduce the transmission delays, few of the above-described solutions are used. Managing the network traffic based on the real time scenario will be a more appropriate solution. The network must be deployed in a scalable manner. This not only helps in easier traffic management but also provides better Quality of Service (QoS) for the customers. This paper has discussed the significant factors related to the traffic management and load balance in cloud computing.

With this paper, we have discussed the existing research gap which addresses that no such work is presented with better load balance in cloud with dynamic variation in the load. Also, the existing techniques are cost effective. The future scope of research to solve the load balance issues were addressed with proper resource allocation.

References

[1] Armbrust, Michael, "A review of cloud computing." Communications of the ACM 53.4, pp.50-58, 2010.

[2] Armbrust, Michael, et al. "Above the clouds: A berkeley view of cloud computing.", 2009.

[3] Zhang, Qi, Lu Cheng, and Raouf Boutaba. "Cloud computing: state-of-the-art and research challenges." Journal of internet services and applications 1.1, pp.7-18, 2010.

[4] T OGRAPH, B., and Y. RICHARD MORGENS. "Cloud computing." Communications of the ACM 51.7, 2008.

[5] Maguluri, Siva Theja, R. Srikant, and Lei Ying. "Stochastic models of load balancing and scheduling in cloud computing clusters." INFOCOM, 2012 Proceedings IEEE. IEEE, 2012.

[6] Rimal, Bhaskar Prasad, Eunmi Choi, and Ian Lumb. "A taxonomy and survey of cloud computing systems." INC, IMS and IDC (2009): 44-51.

[7] Weiss, Aaron. "Computing in the clouds." Computing 16 (2007).

[8] Fayoumi, Ayman G. "Performance evaluation of a cloud based load balancer severing Pareto traffic." Journal of Theoretical and Applied Information Technology 32.1 (2011): 28-34.

[9] Berl, Andreas, et al. "Energy-efficient cloud computing." The computer journal 53.7 (2010): 1045-1051.

[10] Marston, Sean, et al. "Cloud computing—The business perspective." Decision support systems 51.1 (2011): 176-189.

[11] Tunc, Cihan, et al., "Value of Service Based Task Scheduling for Cloud Computing Systems" ,ICCAC, 2016.

[12] Shen, Qinghua, et al. "Exploiting geo-distributed clouds for a e-health monitoring system with minimum service delay and privacy preservation." IEEE journal of biomedical and health informatics 18.2 (2014): 430-439.

[13] de Abranches, Marcelo Cerqueira, and Priscila Solis. "An algorithm based on response time and traffic demands to scale containers on a Cloud Computing system." Network Computing and Applications (NCA), 2016 IEEE 15th International Symposium on. IEEE, 2016.

[14] Assi, Chadi, et al. "Towards scalable traffic management in cloud data centers." IEEE Transactions on Communications 62.3 (2014): 1033-1045.

[15] Brindha, SK Jeya, J. Angela Jennifa Sujana, and T. Revathi. "Agent based bidirectional bidding mechanism for efficient scheduling of real time tasks in cloud computing." Electrical, Electronics, and Optimization Techniques (ICEEOT), International Conference on. IEEE, 2016.

[16] Doyle, Joseph, Robert Shorten, and Donal O'Mahony. "Stratus: Load balancing the cloud for carbon emissions control." IEEE Transactions on Cloud Computing 1.1 (2013): 1-1.

[17] Patel, Ronak R., et al. "Improved GA using population reduction for load balancing in cloud computing." Advances in Computing, Communications and Informatics (ICACCI), 2016 International Conference on. IEEE, 2016.

[18] Ran, Chen, Shaowei Wang, and Chonggang Wang. "Balancing backhaul load in heterogeneous cloud radio access networks." IEEE Wireless Communications 22.3 (2015): 42-48.

[19] Rimal, Bhaskar P., and Martin Maier. "Workflow Scheduling in Multi-Tenant Cloud Computing Environments."IEEE, 2015.

[20] Soniya, J., J. Angela Jennifa Sujana, and T. Revathi. "Dynamic Fault Tolerant Scheduling Mechanism for Real Time Tasks in cloud computing." Electrical, Electronics, and Optimization Techniques (ICEEOT), International Conference on. IEEE, 2016.

[21] Vascak, Jan, Jakub Hvizdo, and Michal Puheim. "Agent-Based Cloud Computing Systems for Traffic Management." Intelligent Networking and Collaborative Systems (INCoS), 2016 International Conference on. IEEE, 2016.

[22] SundarRajan, R., V. Vasudevan, and S. Mithya. "Workflow scheduling in cloud computing environment using firefly algorithm." Electrical, Electronics, and Optimization Techniques (ICEEOT), International Conference on. IEEE, 2016.

[23] Wang, Xiumin, et al. "Traffic-aware task allocation for cooperative execution in mobile cloud computing." Communications in China (ICCC), 2016 IEEE/CIC International Conference on. IEEE, 2016.

[24] Zhang, Dongbo, Yanfang Shou, and Jianmin Xu. "The modeling of big traffic data processing based on cloud computing." Intelligent Control and Automation (WCICA), 2016 12th World Congress on. IEEE, 2016

Real-Time H.264/AVC Entropy Encoder Hardware Architecture in Baseline Profile

Ben Hamida Asma[1], Dhahri Salah[2]
[1,2] Laboratory of Electronic and Micro-Electronic (LAB-IT06),
Faculty of Sciences of Monastir,
University of Monastir, Tunisia

Zitouni Abdelkrim[3]
[3]College of Education in Jubail,
University of Dammam,
Saudi Arabia

***Abstract*—In this paper, we present a new hardware architecture of an entropy encoder for an H.264/AVC video encoder. The proposed design aims to employ a parallel module at a pre-encoding stage to reduce a critical path. Additionally, the arithmetic table elimination method is used to eliminate the memory cost. Besides, the reduction in the size of VLC tables offers area saving. This architecture is synthesized on an FPGA Virtex IV. The simulation results show that this design can operate up to 234 MHz, which allows processing a 4CIF video format in real time.**

Keywords—H.264/AVC; CAVLC; Exp-Golomb

I. INTRODUCTION

The entropy encoder is the last part of an H.264/AVC encoder. H.264/AVC identifies two types of entropy coding methods, which are the Context-Based Adaptive Variable Length Coding (CAVLC) and the Context-Based Binary Arithmetic Coding (CABAC) [1]. In a baseline profile, only the CAVLC is utilized as an entropy coder mode with Exponential-Golomb (Exp-Golomb) codes. The CAVLC produces coding with higher efficiency than the conventional VLC coding. However, the CAVLC adds a high computational complexity due to context-adaptive characteristics.

Some work has presented the VLSI architecture of the CAVLC encoder to improve the performance of the entropy encoder. However, most work has focused only on how to increase the throughput of the CAVLC encoder. For instance, the pipelining architecture is usually used [2, 3, 4]. The work in [2] proposed a two-stage pipeline architecture. This method could reduce the time needed to process a block until reaching half of the mean time but it involved double memory size to store all syntax element information. In [3], the parallel coding of level and run-before sub-module encoders was applied. Moreover, the authors in [5] tried to increase the throughput by scanning the coefficient in parallel. However, it clearly doubled the area cost.

To reduce this area cost, [5] put forward optimized coefficient token (coeff-token) VLC Look-Up Tables (LUTs) into 9-bit words instead of storing 16-bit words. An arithmetic manipulation of encoding levels was exploited in [6] to eliminate some of the large size of conventional VLC LUTs.

On the other hand, some work has concentrated on designing a low-power CAVLC encoder. For instance, the authors in [7] used the side information-aided and symbol look-ahead techniques to minimize memory access.

This paper presents full hardware architecture of entropy coding, which contains Exp-Golomb and CAVLC encoders for an H.264/AVC baseline profile. To improve the timing performance, parallel coding modules are introduced at the pre-coding stage. To decrease the cost memory, an arithmetic table elimination technique is exploited to encode level and run-before sub-module encoders instead of using conventional VLC LUTs. Furthermore, the optimized coeff-token VLC and total-zero LUTs are applied to reduce the memory size as well.

This paper is organized as follows. Section 2 introduces both CAVLC and Exp-Golomb entropy encoding algorithms. The proposed architecture designs of the CAVLC and the Exp-Golomb are illustrated respectively in sections 3 and 4. Finally, the conclusion is drawn in section 5.

II. ENTROPY CODING ALGORITHM IN H.264

In the baseline profile, H.264 uses two tools for entropy coding: the CAVLC coding and the Exp-Golomb one, as presented in Fig.1. The residual information (quantized coefficients) is coded using the CAVLC, while the other data are coded utilizing the Exp-Golomb.

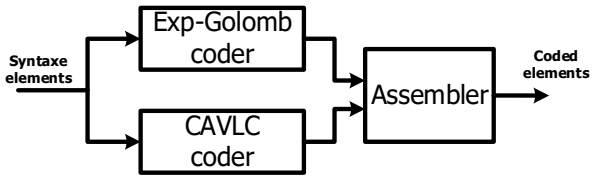


Fig. 1. Block diagram of entropy coder in baseline profile

A. CAVLC algorithm

The CAVLC is the entropy encoding used to encode the residual information in 4x4 or 2x2 blocks, which are generated by the quantification step [1]. Each block must be firstly scanned in a zigzag order to produce five main syntax elements. The latter were defined in [1] as:

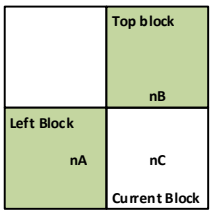
- The coeff-token represents two values : the total number of non-zero coefficients (total-coeff) and the number of trailing ones (TT1s) in the block. The trailing ones (T1s) are non-zero coefficients whose values

are+/- 1 at the end of the zigzag sequence. Each block has at most three T1s.

- The signs of T1s are the coefficients with absolute value equal to one from zero to three bits wide. They represent the signs of the T1s coefficients in the reverse order.
- The levels are the values of each non-zero coefficient in the block, other than the T1s case. They are taken in the reverse order.
- The total-zeros is the total number of zero coefficients before the last non-zero coefficient in the zigzag sequence.
- The run-before represents the runs of zeros before each non-zero coefficient in the reverse order.

After that, these syntax elements will be encoded into five sequentially coding steps. The coeff-token, run-before and total-zero steps are encoded through different VLC LUTs. The CAVLC encoder steps are depicted in Fig.3.

- In step 1, the coeff-token are encoded using four VLC LUTs, based on the number of the total coefficients in the left block (nA) and the upper block (nB) of the current block (the context-adaptive notion), as shown in Fig.2.
- In step 2, each T1s is encoded with its corresponding bit sign in a reverse order. The positive sign is represented by '0', and the negative sign is represented by '1'.
- In step 3, the level values of the 4x4 block are encoded in a reverse order using seven VLC LUTs selected by the total-coeff and TT1s. The choice of the VLC LUTs to encode each level depends on the magnitude of the last encoded level (the context-adaptive notion).
- In step 4, 15 VLC LUTs are utilized to encode the total zeros, indexed by the total-coeff value.
- In step 5, the run-before is coded with codewords taken from seven VLC LUTs selected by zero-left values, which is the total number of the remaining zero coefficients.



(a)

```
If (nA = nB =0) then
nC =0;
Else
nC =(nA +nB)/2;
End if;
```

(b)

Fig. 2. Context-adaptive notion at coeff-token coding step (a) Data dependence (b) Correspond pseudo-code

B. Exp-Golomb algorithm

The Exp-Golomb coding is performed on two stages as provided in Fig.3.

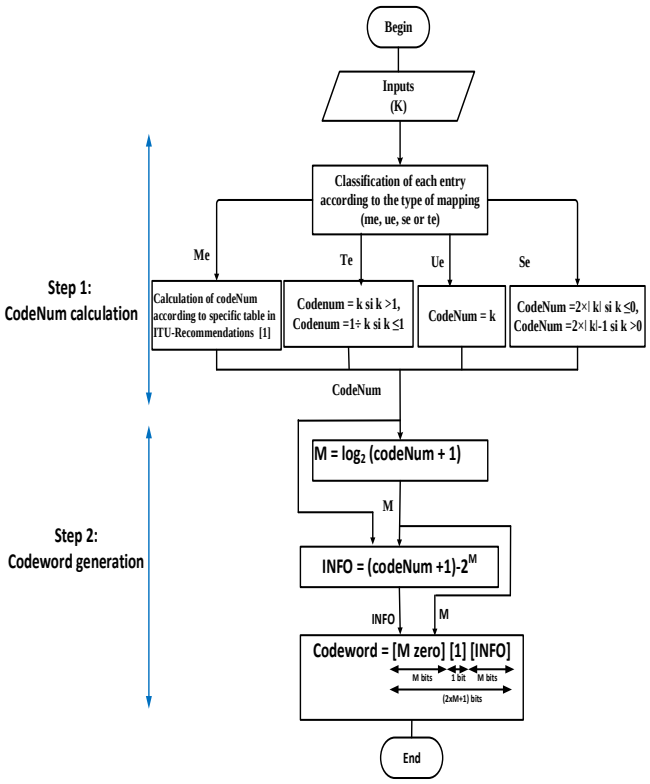


Fig. 3. Diagram of Exp-Golomb algorithm

- Firstly, each syntax element to be coded with the Exp-Golomb noted k is mapped to a non-negative integer named "codeNum." Based on the statistical characteristic, each syntax element is represented by a codeNum in various ways [1].
- If a syntax element is always larger than zero or equal to zero and if the most frequently occurring values are the lower ones, the applied process will be called "unsigned Exp-Golomb (ue) coding". The value of the corresponding codeNum is the same value of the unsigned element.
- If a syntax element is signed and the expectation value is zero, the applied process will be named "signed Exp-Golomb (se) coding". The value of the corresponding codeNum is mapped to the syntax element value k as follows:
  - CodeNum = 2|k| when ($k \leq 0$)
  - CodeNum = 2|k| - 1 when ($k > 0$)
- If an unsigned element has different statistical characteristics from the ue, its corresponding codeNum is then mapped to its value in a special way, as indicated in ITU-T recommendations [1]. The applied process is called "mapped Exp-Golomb (me) coding."
- If an unsigned element has 1 as the largest possible value, then "the truncated Exp-Golomb (te) coding" will be applied ; i.e., the bit representing the syntax element is the inverted value of the element.

Secondly, the codeNum parameter is mapped to coded string bits. The latter has the following generic form:

$$\{\text{M-zeros, 1, M-bit INFO}\} \quad (1)$$

where M and INFO are given by equations 2 and 3.

$$M = \text{floor}(\log2\,[\text{codeNum} + 1]) \quad (2)$$

$$INFO = \text{codeNum} + 1 - 2M \quad (3)$$

III. Proposed CAVLC Architecture

The suggested design processes each 4x4 block through two sequential stages. The pre-coding stage produces the Syntax Elements (SEs) to be encoded from the residual input frames, and the encoding stage translates each SE into a related codeword length and codeword value. In the following subsections, both stages are described.

A. Pre-encoding CAVLC stage architecture

The pre-encoding architecture is depicted in Fig.4. It has five main modules and four Random Access Memories (RAMs). The main modules are depicted in the figure below:

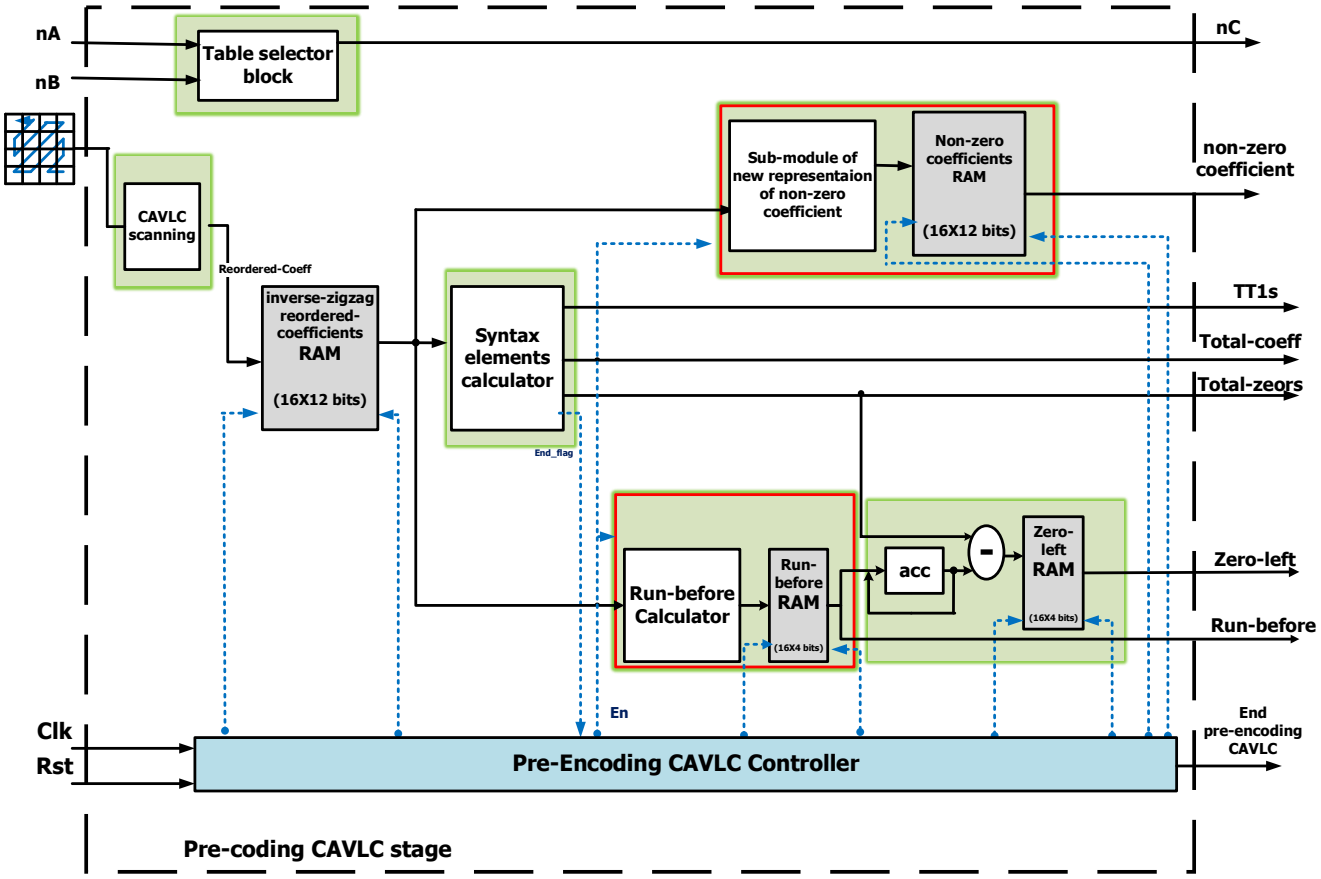
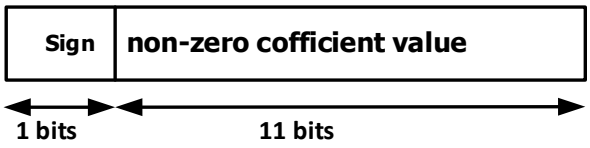


Fig. 4. Pre-encoding CAVLC architecture

The zigzag module is responsible for ordering in an inverse zigzag order the residual information coming from the quantification process. After that, the zigzagged reordered coefficient is stored in a first memory called "inverse-zigzag reordered-coefficient RAM." This module is not included in the CAVLC modules, but it is required for its correct operation.

The generator module of syntax elements has as an input the reordered coefficients. This module generates the first syntax elements to be produced, which are the TT1s, the total-coeff, and the total-zeros. When the values of these syntax elements are calculated, the next two modules, shown in red squares, start to be processed. Both modules are independent. Consequently, they are processed in parallel.

The parallel module on the top is responsible for storing the T1s and the level values into a "non-zero coefficient RAM" memory. The total number of levels and TT1s represent all total non-zero coefficients. Each non-zero coefficient is saved with a new format that represents the absolute value of the non-zero coefficient in 11 bits and the sign bit in the 11^{th} bit, as illustrated in Fig.5. This format allows simplifying the level encoding process.



(a)

```
If non-zero_coeff(11) ='1' then
 non-zero_coeff_RAM (conv_integer(i)) <= '1'& (b"00000000000" -   non-zero_coeff_RAM (10 downto 0));
Else
  non-zero_coeff_RAM(conv_integer(i)) <= non-zero_coeff;
End if;
```

(b)

Fig. 5. (a) New representation of non-zero coefficient and (b) its correspondent pseudo-code

The second parallel module is formed by combinatorial circuits and two RAMs needed for storing each run-before and zero-left syntax element, respectively. First, this module permits calculating the different run-before values. After that, each calculated run-before value will be put into the "Run-before RAM" memory. When all the run-before values are detected and stored, the controller enables the process of the next module. This latter calculates the set of zero-left values and stores them into a "Zero-left RAM "memory. The zero-left value is initially equal to the total-zeros, and then this value is decremented with the accumulation of run-before values. The mathematical relationship between the zero-left and the run-before is shown below.

$$\text{Zero-left (i)} = \text{Total-zeros} - \sum \text{Run-before} \qquad (4)$$

It is worth noting that the size of all used memory is 16 elements, which is the maximum number of non-zero run-before and zero-left coefficients per 4x4block. Besides, the use of the inverse-zigzag reordered-coefficient, Run-before and Zero-left RAM memories is required for bitstream correctness.

The nC is also generated at this stage by a combinatorial circuit shown in Fig. 6. It selects the appropriate VLC LUTs for coeff-token coding.

The controller at the pre-encoding stage is in charge of defining the control unit of the different RAMs and synchronizing the various modules. When the end-pre-encoding signal is set active, all the syntax elements will be ready to be encoded.

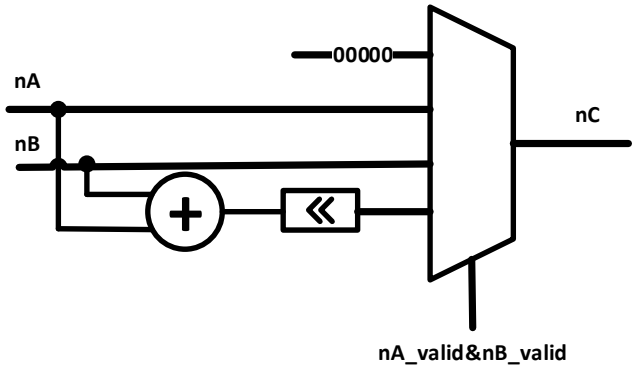


Fig. 6. Table selector architecture

B. *Encoding CAVLC stage architecture*

The encoding CAVLC architecture is illustrated in Fig.7. The CAVLC hardware design has the outputs of the CAVLC pre-encoder design as inputs. It is composed of seven main modules: five modules in charge of encoding the different syntax elements, one module for the main controller, and another one for the output packet. These various modules and the optimized techniques used at this stage are detailed in the following subsections.

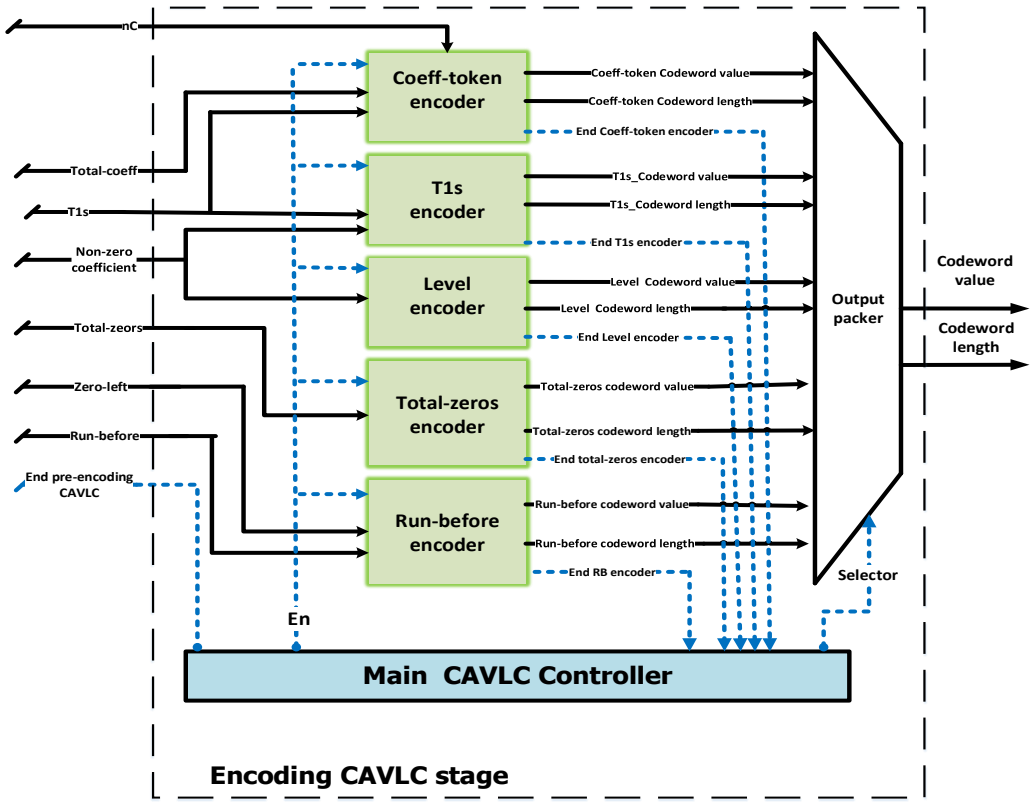


Fig. 7. Encoding CAVLC architecture

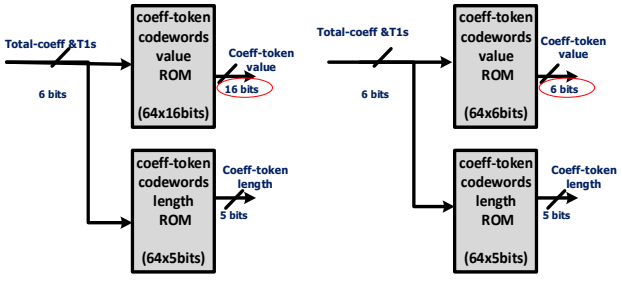
1) Optimized VLC LUTs for coeff-token and total-zero encoders:

The coeff-token and total-zero encoders are conventionally coded by different VLC LUTs in the ITU-T Recommendations [1]. However, large memory size is required to store the whole codewords' values and lengths, as presented in these traditional VLC LUTs. In the light of these details, we suggest a new representation of the codeword length and codeword value into small size. For instance, the length of the original codewords in conventional VLC coeff-token LUTs is in the range of 1 to 16, and their values are in the range of 0 to 63. Therefore, 5 bits are enough to represent the length information into the "coeff-token codeword value ROM" memory, and 6 bits are enough to represent the value information into the "coeff-token codeword length ROM" memory. An example of the new representation of codewords is given in Table I.

This method is applied for all VLC LUTS needed for coeff-token and total-zero sub-module encoders .It enables optimizing the VLC LUTs for both coeff-tokens and total-zeros. An example of an optimized VLC LUT is depicted in Fig.8.

TABLE. I. AN EXAMPLE OF A NEW REPRESENTATION OF CODEWORD IN VLC LUT

Original codeword		**Proposed codeword**	
length	***Value***	***Length***	***Value***
10000	00000000000000010	10000	000010
5 bits	16 bits	5 bits	6 bits



(a) (b)

Fig. 8. Block-diagram representation of an example of (a) traditional VLC coeff-token LUT (b) optimized VLC coeff-token LUT

2) Arithmetic table elimination technique for level encoder:

Levels are encoded using the arithmetic table elimination technique to replace seven level VLC LUTs represented in the ITU-T recommendations [1]. This technique reported from [6] permitted the reduction in the memory cost area. Table II reports the pseudo-code describing the elimination procedure, which presents the advantage of a very simple implementation circuitry.

The format of the level code is arranged as follows. The maximum width of codewords' length is 28 bits.

$$\text{Code} = \underbrace{0\ldots0}_{\text{Prefix length}}\ 1\ \underbrace{x\ldots x}_{\text{Suffix length}}\ s \qquad (5)$$

Note where s is the level sign, 1 for negative, 0 for positive, and the sequence of zeros on the left of 1 and the sequence of bits on its right are respectively the level prefix and the level suffix, whose lengths, prefix length and suffix length, distinguish the codewords.

This step also illustrates the context-adaptive characteristic such that suffix length N (ranging from 0 to 6), used for encoding the actual level, must be the same one to encode the previous level. Otherwise, it will be eventually incremented if its magnitude satisfies ($3x2^{(N-1)}$). The pseudo-code of the adaptive context is shown in Fig.9.

```
If | Level | > 3x2^(N-1) then
        N <= N +1;
Else
        N <= N;
End if;
```

Fig. 9. Pseudo-code of adaptive-context at level coding step

TABLE. II. CODING ALGORITHM FOR LEVEL SYMBOL

N	Range	Coding algorithm
N=0	|level|≤7	Code=0…0 1 Prefix length= (|level|≪1)-2+s Suffix length=0 Size=prefix length+1
	8≤|level|≤15	Code=0…0 1 s Prefix length=14 Suffix length=3 Size=19 Level suffix=binary value(|level|)
	|level|≥16	Code=0…0 1 x…x s Prefix length=15 Prefix length=11 Size=28 Level suffix=|level|-1-[15≫ (N-1)]

N=1 to 6	All	Code=0…0 1 x…x s
		If (|level|-1< [15≪ (N-1)]) then
		Prefix length= (|level|-1) ≫ (N-1)
		Suffix length=N-1
		Size =prefix length + suffix length
		Level suffix=|level-1|%2(N-1)
		Else
		as case |level|≥16 for N=0
		End if

3) Arithmetic table elimination technique for run-before encoder:

The seven VLC LUTs required for run-before encoding are eliminated and substituted by a circuitry implementing the pseudo-code in Table III. With this approach, we achieve a reduction in the memory cost as well.

TABLE. III. CODING ALGORITHM FOR RUN-BEFORE SYMBOL

Zeroleft	Coding algorithm

<3	If Runbefore(i)=0 then
	Code=1
	Size=1
	Else
	Code=Zeroleft(i)-Runbefore(i)
	Size =Zeroleft(i)
	End

≥3 and <6	If RunBefore(i)≤6-Zeroleft(i) then
	Code=3-RunBefore(i)
	Size=2
	Else
	Code=Zeroleft(i)-RunBefore(i)
	Size=3
	End

=6	If RunBefore(i)=0 then
	Code=3
	Size=2
	Elsif RunBefore(i)=1 then
	Code=0
	Size=3
	Elsif RunBefore(i)=6 then
	Code=4
	Size=3
	Else
	If LSB[RunBefore(i)]=0then
	Code =RunBefore(i) >> 1
	Else
	Code =RunBefore(i)
	End if
	Size=3
	End

>6	If RunBefore(i)<6 then
	Code =7-RunBefore(i)
	Size=3
	Else
	Code=1
	Size=RunBefore(i)-3
	End

4) Main CAVLC controller:

The proposed CALVC controller is presented in Fig.10. The "idlestate " represents the initial state. When the pre-encoding stage is finished (indicated by the signal "end pre-encoding CAVLC"), the finite state machine will go to the "coeff-token state". When the coeff-token encoder process is finished (indicated by the signal "end coeff-token encoding"), the finite state machine will affect the appropriate value of the signal "mux-selector" to select the output of the coeff-token encoder as final outputs. Afterwards, the finite state machine will go to the "T1s state" . When the T1s encoder process is completed, the finite state machine will produce an appropriate value for the signal "mux-selector" to select the outputs of the T1s encoder as final ones.

This process, which is produced in the coeff-token and T1s states, will be replicated at level, total-zero and run-before states. At the end of the run-before encoding process, the signal "end run-before encoding" is set high, informing that the CAVLC completely encodes the 4x4 block, and a new block can be encoded.

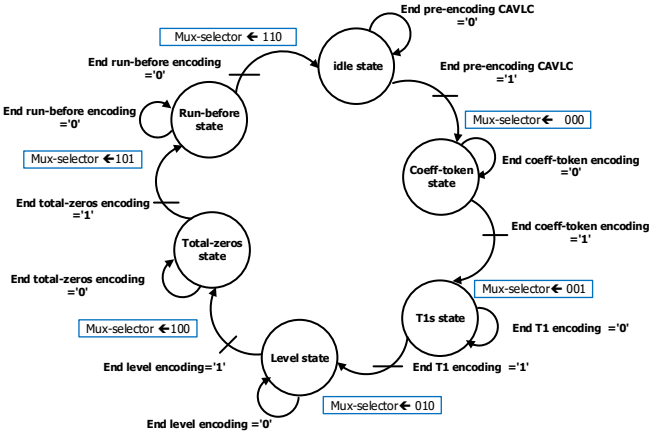


Fig. 10. Main CAVLC controller

5) Output packer:

The output packet receives as an input the signal "mux-selector" from the main controller and all the outputs of the encoder modules (codeword values and codeword lengths). Two-word multiplexers compose this module: one to select the appropriate codeword value and the other to select the appropriate codeword length. The codeword value and codeword length serve as final outputs of a CAVLC coder.

IV. Proposed Exp-Golomb Architecture

The proposed Exp-Golomb design is presented in the form of modules in Fig.11. Every module represents the functioning way of each stage of the Exp-Golomb algorithm already explained in section II.

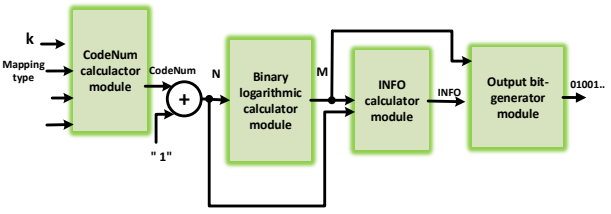


Fig. 11. Block diagram of Exp-Golomb

Firstly, in every k entry, the"codeNum generator" module generates the corresponding codeNum value according to a mapping type (ue, te, se or me). When the mapping type is ue, te or se, the codeNum value will be generated from the"codeNum generator1" Module. This block produces the codeNum according to various mathematical operations described in section 2, which only involves shifting, complementation, and increasing by 1. Otherwise (mapping type =me), the codeNum will be generated by a second generator module, called "codeNum generator2", based on four ROMs according to two mode types (intra or inter) and to the prediction mode. The detailed architecture of these two generators of codeNum values is shown, respectively, in Fig.12 and Fig.13.

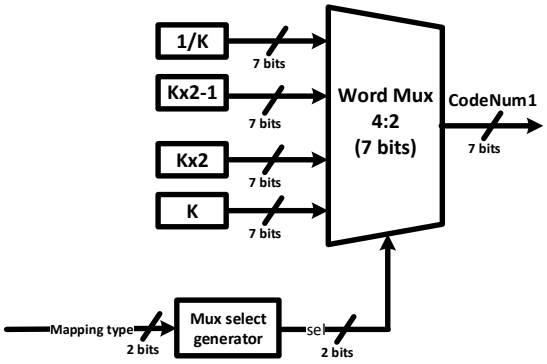


Fig. 12. Diagram of codeNum generator1

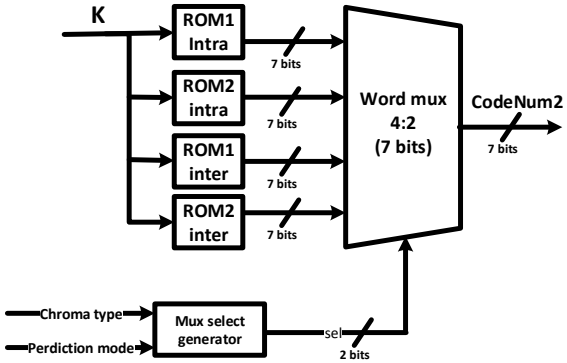


Fig. 13. Diagram of codeNum generator2

The logarithm operation is required to produce the value of M, which is utilized for the calculation of the codeword length (equivalent to 2M+1). However, its implementation requires an expensive circuit that constitutes the hardware challenge of implementing an Exp-Golomb encoder. This problem can be solved in the following way: Consider that log2(N) is equivalent to the number of M times divided by 2 until the output reaches the zero value as in equation 3. Thus, we acquire an approach to get the value of M by computing the shift operation number.

$$M = \log2(N) \leftrightarrow N = 2^{M} = \underbrace{2x2x2x2\ldots.x2}_{M \text{ times}} \quad (6)$$

The suggested architecture of the logarithm operation is given in Fig.14. The output of the barrel shifter is loaded in the register FF. The output Q of this register is connected to the inputs of the multiplexer and the combinatorial circuit of the OR gates. This circuit is responsible for checking whether the output Q reaches the value 0 or not by producing a one-bit value, noted C, as an output.

Initially, the counter is set at 0. If the value of Q is different from zero; the value C is equal to 1. Consequently, the AND gate will be an ascending counting; the counter will count up by a single step. In this case, the multiplexer is going to assign the value Q to the input K of the barrel shifter. When Q reaches the value 0, the value of the output C is set to 0. Therefore, the

logic 0 generated on the output of the AND gate stops the counter. In this case, the output of the counter corresponds to the output value M such as M =log2 (N).

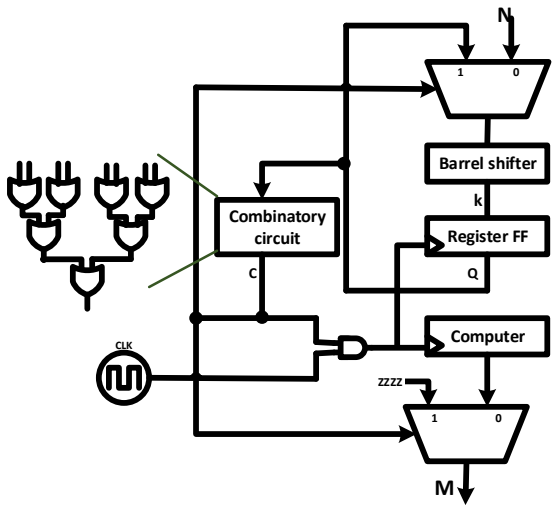


Fig.14. Architecture of binary logarithm

After the logarithm operation, the INFO value should follow formula (3), which involves shifter and subtraction operations.

The last module (Exp-Golomb bit-generator) is in charge of producing the output code word considering the value of M and INFO. It is designed by the implementation of the finite state machine that contains two states, as shown in Fig.15. The i^{th} counter is initialized to state 1. The second state corresponds to the generation of the output codewords bit by bit, following the structure presented in formula (1). Each bit is generated in one clock.

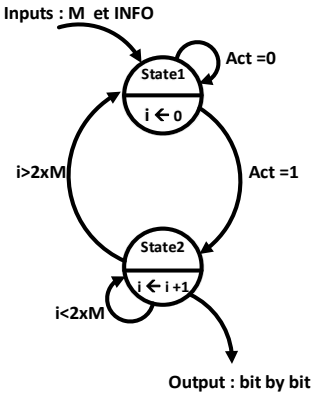


Fig. 14. Finite state machine of bit Exp-Golomb bit-generator

V. Performance Analysis and Comparison

A. Performance analysis

The proposed CAVLC and Exp Golomb architectures are modeled in VHDL, simulated, and synthesized by Modalism 6.4 and Xilinx ISE development tools 14.1, respectively. The synthesis results of physical resource utilization on Virtex VI for CAVLC and Exp-Golomb modules are reported respectively in Table VI and Table V.

TABLE. IV. Physical Resources Utilization of CAVLC Modules on Virtex VI

Module		Slice LUT	Slice register
Pre-encoding CAVLC		298	467
nC calculator		10	0
Encoding CAVLC	Coeff-token encoder	143	0
	T1s encoder	42	21
	Level encoder	316	77
	Total-zero encoder	62	0
	Run-before encoder	156	103
CAVLC controller		6	3
Output Packet		61	0
Total CAVLC		589	557

TABLE. V. Synthesis Results of Exp-Golomb on Virtex VI

Module	Slice LUT	Slice register
CodeNum generator	24	0
Binary logarithm	36	38
INFO-calculator	10	30
Exp_golomb bit-generator	44	26
Exp-Golomb Controller	26	13
Total Exp-Golomb	126	109

Through the obtained results, it is possible to verify that the CAVLC coder achieves an operation frequency of 234.14 MHz and requires an area occupancy of 847 LUTs. The maximum frequency of the Exp-Golomb architecture is 234.14 MHz, and the memory cost is 847 in terms of LUTs.

It is worth mentioning that no external or embedded memory is used to give a platform independent estimation of memory cost reduction, suitable for ASICs and FPGAs of different generations and families.

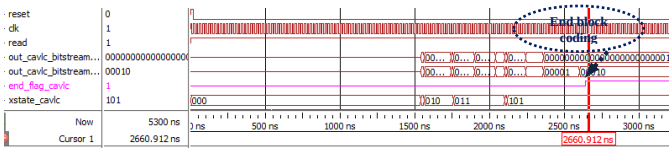


Fig. 15. Simulation results of CAVLC

The simulation results provided in Fig.16 show that the processing time per block exhibit a large variety. We take an average of 131 cycles per block. The performance of our proposed architecture is calculated as follows:

The number of clock cycles needed for 4CIF (704 x 576) video with 30-fps=

The number of clock cycles per block

x the number of blocks per macroblock

x the number of macroblock per 704

x 576 video frame x the number of frames per second=

131 x 27 x 1665 x 30 clock cycles=

176,673,150 clock cycles

The number of required clock cycles is calculated. We assume the worst case without the skip mode. This means that our suggested architecture can meet the real-time processing of 4CIF @30fps when running at 234 MHz.

B. State-of-the-art comparison

To give a reasonable comparison, both designed CAVLC and Exp-Golomb are synthesized with different FPGA platforms, as presented in Table VI.

TABLE. VI. COMPARISONS TO HIGH-PERFORMANCE DESIGNS CAVLC

	Tech	**Frequency** (MHz)	**Area**	
			Gates (ASIC)	**LUTs** (FPGA)
[2]	0.13	250	32K	-
[3]	Virtex 5	180	-	1079
[8]	Stratix IV	200	-	6549
[9]	Spartan 3	62.5	-	3447
[10]	Virtex 5	204.3	-	2563
[11]	0.18	100	73.5k	-
[12]	0.18	125	15K	-
Proposed	Spartan 3	91.43	-	754
	Virtex 5	234.14	-	847
	Spartan 3	313.87		1176

Concerning speed performance, the proposed CAVLC design exhibits a maximum operating frequency, which is mostly superior compared to other CAVLC design solutions. The memory cost of our design is also very promising, thanks to the optimized VLC LUTs and the arithmetic table elimination techniques.

TABLE. VII. COMPARISONS TO HIGH-PERFORMANCE DESIGNS EXP-GOLOMB

	Tech (um)	**Logic (LUTs)**	**Frequency (MHz)**
[10]	Virtex VI	134	309.98
[13]	Stratix II	199	191.8
Our design	Virtex VI	126	254.4

Table VII summarizes the specification of the suggested Exp-Golomb encoder and gives a comparison with the work presented in [10] and [13]. The operating frequency of the proposed architecture is lower than those presented in [10], but the suggested design has a lower area demand. Compared to the design shown in [13], the proposed architecture employs a higher area demand and a higher operating frequency.

VI. CONCLUSION

In this paper, a full hardware design entropy encoder for the H.264/AVC baseline profile has been put forward. Different techniques have been employed to improve the performance of the entropy encoder. Parallel modules are applied to speed up the coding efficiency. Meanwhile, the employment of the optimized VLC LUTs and Arithmetic method have been used to reduce the area cost. The synthesis results on Virtex IV have shown that the design occupies about 847 LUTs and can be targeted for a real-time 4CIF video format when operating at 234 MHz.

REFERENCES

[1] Draft ITU-T Recommendation and Final Draft International Standard of Joint Video Specification (ITU-T Rec. H.264 ISO/IEC 14496-10 AVC)," in Joint Video Team (JVT) of ISO/IEC MPEG and ITU-T VCEG, JVT-G050rl, May 2003.

[2] HuiboZhong,Yibo FAN, Xiaoyang ZENG , A Parallel CAVLC Design For 4096x2160p Encoder, in Proc. 2012 IEEE International Symposium on Circuits and Systems

[3] F.L.L. Ramos, B. Zatt, T.L. Silva, A. Susin, and S. Bampi. A High Throughput CAVLC Hardware Architecture with Parallel Coefficients Processing for HDTV H.264/ AVC Enconding. In Proceedings of the 17th IEEE International Conference on Electronics, Circuits, and Systems (ICECS), 2010, Dec 2010, pages 587 – 590

[4] C. D. Chien, K. P. Lu, Y. H. Shih, and J. I. Guo, "A high performance CAVLC encoder design for MPEG-4 AVC/H.264 video coding applications," in Proc. IEEE ISCAS'06, 2006, pp. 3838–3841."

[5] Nguyen, N. M., Tran, X. T., Vivet, P., & Lesecq, S. (2012, October). An efficient Context Adaptive Variable Length coding architecture for H. 264/AVC video encoders. In Advanced Technologies for Communications (ATC), 2012 International Conference on (pp. 158-164). IEEE."

[6] Albanese, L. F., & Licciardo, G. D. (2010, September). An area reduced design of the Context-Adaptive Variable-Length encoder suitable for embedded systems. In I/V Communications and Mobile Network (ISVC), 2010 5th International Symposium on (pp. 1-4). IEEE.

[7] Tsai, C. Y., Chen, T. C., & Chen, L. G. (2006, July). Low power entropy coding hardware design for H. 264/AVC baseline profile encoder. In 2006 IEEE International Conference on Multimedia and Expo (pp. 1941-1944). IEEE.

[8] M. P. Hoffman, E. J. Balster, and W. F. Turri, "High-throughput CAVLC architecture for real-time H. 264 coding using reconfigurable devices," Journal of Real-Time Image Processing, vol. 11, pp. 75-82, 2016.

[9] Albanese, L.F.; Licciardo, G.: High-speed CAVLC encoder suitable for field programmable platforms. In: Proceedings of 2010 international conference on signals and electronic systems (ICSES), pp. 327–330 (2010)

[10] Thiele, C. C., Vizzotto, B. B., Martins, A. L., da Rosa, V. S., & Bampi, S. (2012, October). A low-cost and high efficiency entropy encoder architecture for H. 264/AVC. In VLSI and System-on-Chip, 2012 (VLSI-SoC), IEEE/IFIP 20th International Conference on (pp. 117-122). IEEE.

[11] N.-M. Nguyen, E. Beigne, S. Lesecq, P. Vivet, D.-H. Bui, and X.-T. Tran, "Hardware implementation for entropy coding and byte stream packing engine in H.264/AVC," in Proceedings of the International Conference on Advanced Technologies for Communications (ATC), Ho Chi Minh City, October 2013, pp. 360–365.

[12] Licciardo, G.D., Albanese, L.F."Design of a context-adaptive variable length encoder for real-time video compression on reconfigurable platforms." Image Process. IET 6(4), 301–308 (2012)

[13] Silva, T., Vortmann, J., Agostini, L., Bampi, S., & Susin, A. (2007, February). FPGA based design of CAVLC and exp-golomb coders for H. 264/AVC baseline entropy coding. In Programmable Logic, 2007. SPL'07. 2007 third Southern Conference on (pp. 161-166). IEEE.

A Comparison of Collaborative Access Control Models

Ahmad Kamran Malik
Department of Computer Science
COMSATS Institute of Information Technology (CIIT)
Islamabad, Pakistan

Abdul Mateen
Department of Computer Science
Federal Urdu University of Arts, Science & Technology (FUUAST)
Islamabad, Pakistan

Muhammad Anwar Abbasi
Department of Computer Science
Federal Urdu University of Arts, Science & Technology (FUUAST)
Islamabad, Pakistan

Basit Raza
Department of Computer Science
COMSATS Institute of Information Technology (CIIT)
Islamabad, Pakistan

Malik Ahsan Ali
Department of Computer Science
Federal Urdu University of Arts, Science & Technology (FUUAST)
Islamabad, Pakistan

Wajeeha Naeem
Department of Computer Science
COMSATS Institute of Information Technology (CIIT)
Islamabad, Pakistan

Yousra Asim
Department of Computer Science
COMSATS Institute of Information Technology (CIIT)
Islamabad, Pakistan

Majid Iqbal Khan
Department of Computer Science
COMSATS Institute of Information Technology (CIIT)
Islamabad, Pakistan

***Abstract*—Collaborative environments need access control to data and resources to increase working cooperation efficiently yet effectively. Several approaches are proposed and multiple access control models are recommended in this domain. In this paper, four Role-Based Access Control (RBAC) based collaborative models are selected for analysis and comparison. The standard RBAC model, Team-based Access Control (TMAC) model, Privacy-aware Role-Based Access Control (P-RBAC) model and Dynamic Sharing and Privacy-aware RBAC (DySP-RBAC) model are used for experiments. A prototype is developed for each of these models and pros and cons of these models are discussed. Performance and sharing parameters are used to compare these collaborative models. The standard RBAC model is found better by having a quick response time for queries as compared to other RBAC models. The DySP-RBAC model outperforms other models by providing enhanced sharing capabilities.**

Keywords*—RBAC; Collaboration; Privacy; Access control; Security; Information sharing

I. INTRODUCTION

User's act of accessing data, information, and resources is controlled to keep check on authorized users and to avoid unauthorized users. Access control is considered as one of the most challenging and complex issues that dynamic collaborative environments face during security administration. The Role-based Access Control (RBAC) model is an approach to control the access of authorized users whenever roles and privileges are involved in a scenario. National Institute of Standards and Technology (NIST) has provided the standard model for RBAC [1]. It has been extended by many researchers to incorporate requirements posed by different applications and scenarios. Collaborative applications are an important research area for access control which tries to control the access of collaborating users. Many different RBAC based models have been proposed for collaborative environments. As such, it appeared that the RBAC model was a good candidate to provide access control. However, a closer examination revealed that although the RBAC model was a good start, additional notions were necessary to effectively apply the RBAC model in a collaborative setting. The first observation was a need for a hybrid access control model that incorporated the advantages of having broad, role-based permissions across object types, yet required fine-grained control on individual users in certain roles and on individual object instances. A second requirement was a need to recognize context associated with collaborative tasks and to apply this context for permission activation. This can be better understood by drawing a distinction between active and passive security models. A passive security model is the one that primarily serves the function of maintaining permission assignments, like RBAC where permissions are assigned to roles. The standard RBAC model is not suitable for collaborative environments because it does not include many data elements that are fundamental for a collaborative

environment, such as team, task, user relationships, purpose of access and many more.

The Team-based Access Control (TMAC) model grants more permission based on the team as compared to the Standard RBAC model and works better in a teamwork environment, as the team is the key element in TMAC model [4].

The Privacy-aware RBAC model (P-RBAC) is good in privacy and sharing at the same time because this model implements the privacy policies and uses more data elements to enhance their privacy and sharing due to which this model is better than standard RBAC and TMAC model [3]. This model is more suitable for collaborative environments as compared to the RBAC and TMAC models.

The Dynamic Sharing and Privacy-aware RBAC (DySP-RBAC) model [2] is the best model that works in the most collaborative scenarios, as it introduces more elements (Task, Collaborative Relationships, and Access Level) which are more helpful in maintaining privacy and sharing, so this model is more suitable as compare to other models. This paper selects the DySP-RBAC model which is a collaborative model, evaluate and compare it with the other collaborative RBAC models.

There is always a trade-off between information sharing and privacy. This increases twofold in collaborative scenarios. It is much difficult to quantify who should share how much information with whom in a collaborative system. Access control is normally used to control the access to information. Simply RBAC model does not work in collaborative scenarios where users have collaborative relationships among them which are more granular than roles. For this purpose, RBAC model needs to be extended according to collaboration requirements. This paper focused on aforesaid RBAC models; Standard RBAC model, TMAC model, P-RBAC model, and DySP-RBAC model**.**

The main problem is to identify which model is suitable in the specific scenario by comparing collaborative access control models. The objective of this research is to compare and find the pros and cons of collaborative access control models. This will be very helpful for the researchers who want to use, extend or compare standard RBAC model with their own extensions. Using the comparison of collaborative RBAC models, users will be able to select the best matching model for their application requirements.

This research is carried out to distinguish which RBAC model is better to use for which purpose and in which collaborative environment. It also shows the limitation of standard RBAC model in handling collaboration. Four collaboration based RBAC models are selected, a prototype is implemented for each model and compared the models using metrics selected for performance, access control, and information sharing.

In this paper, Section II describes literature review that explains the background of collaborative RBAC models. Section III is an overview of four RBAC models. Section IV explains the methodology of implementation of four RBAC models. Section V includes results and discussion. Section VI concludes the paper and presents future work.

II. Related Work

In the 1970s, fundamental forms of RBAC were implemented in a variety of ad-hoc forms on many systems. Today's RBAC derives from the model proposed in [15] and the RBAC model proposed by Sandhu [20]. Ferraiolo and Kuhn also define a formal definition of roles as the set of permissions, hierarchies, subject-role activation, subject-object mediation, as well as constraints on user/role membership and role activation [15]. In 1994, a role graph model for RBAC was developed, by giving efficient algorithms for analyzing role relationships [16]. Ferraiolo, Cugini, Kuhn presented the concept of the separation of duty forms [17]. The family of RBAC models was introduced by Sandhu Coyne, Feinstein, and Youman in 1996 [20] and the method for implementing MAC on RBAC system was also proposed in 1996 [18]. From 1997-1998, Sybase, Secure Computing, Siemens announce RBAC products described as based directly on Ferraiolo-Kuhn RBAC model. The RBAC ANSI standard model was proposed in 2000 [1]. Further, in 2004, American National Standards Institute, International Committee for Information Technology Standards (ANSI/INCITS) adopts RBAC offer as an industry agreement standard.

The concept of the RBAC model is used in different software application and organizations. The purpose of the RBAC models is for management, security and operating system products. This concept is first time introduced in the market as a standard by NIST. This standard is not applicable in every scenario and situation, so it has been extended by many researchers [5, 9, 10, 12, 14]. The RBAC model is very useful in large scale authorization, widely used in many organizations. This model is widely accepted, still, RBAC has some uncertainty and some problems. There are several RBAC based extended models that are used in different scenarios and situations. There are some models like privacy-aware role-based model, team-based access control model and some other models for handling collaborations. Still these RBAC models are not applicable in every scenario and situation.

This research provides a comparison between four RBAC models including the standard RBAC model. Collaborative information sharing environment requires better information sharing among users while privacy laws require for the protection of user's information from unauthorized access and usage [2]. The DySP-RBAC model is true representative in both domains. A privacy-aware role-based access control (P-RBAC) model is presented in [3]. This model is to force organizations to set privacy policies, privacy framework and enforce the management ideas within organizations. In an organization, there are different kinds of entities like tasks, purposes, relations, and interactions. It can be noticed that in privacy-aware models these kinds of entities are not handled. The P-RBAC model extends the standard RBAC model to express highly complex privacy-related policies, that's why full-fledged P-RBAC solution is easy to deploy in systems already adopting RBAC, thus allowing seamless integration of

access control and privacy policies [3]. There are many more extensions of the RBAC model for handling privacy [11, 13]. The comparison of privacy languages is given in [8]. There is another RBAC extended model TMAC [4] which revolves around teams, where a "team" is an abstraction that encapsulates a collection of users in specific roles and collaborating with the objective of accomplishing a specific task or goal. Users who belong to a team are given access to resources used by a team. Moreover, Collaborative Task Role-Based Access Control (CTRBAC) model [19] and MT-RBAC [7] for the multi-tenants environment to control access to shared resources are available for latest scenarios like Cloud environment. A semantic access control model is also suggested to provide more flexible RBAC for inter and intra-organization environments [6].

III. RBAC Models Selected for Comparison

This section briefly explains four RBAC models that are selected for implementation and further comparison.

A. The Standard Role-Based Access Control (RBAC) Model

This is the NIST standard RBAC model to address the core access control issues. This model is organized into four different components Core RBAC, Hierarchical RBAC, Static separation of duty and Dynamic separation of duty.

Core RBAC describes the main aspects of the RBAC Standard model as shown in the Fig. 1. The concept of this model is to assign the roles to the users. Role is a group of permissions; one role can have many permissions. A user can be assigned to many roles and one role can be assigned to many users. The Core RBAC model also included the concept of session in the model. One user can have many sessions but one session is related to one user only. A user can activate one or more roles (that are assigned to her) in a session. A user session tells about the active and inactive roles of that user.

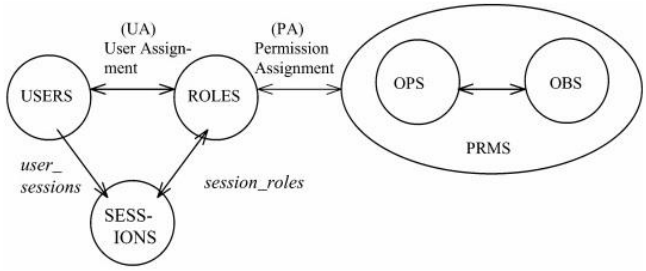


Fig. 1. Core RBAC model [1]

The core RBAC model includes set of elements and their relations. In this model, there are five basic elements that are called the user, roles, objects, operations and their permission. In the RBAC model, users are assigned to the roles and the roles are assigned to permissions. There are many to many relations between user and role, and role and permission. This model also has different kinds of sessions between the user and active roles. A user is assumed to be a human being or any machines or intelligent agents. A role is a job like an employee is assigned to the manager role in the organization and user is fully responsible for the role. Permission assignment is the permission that are assigned to roles for performing an operation on objects. Operations are the set of instructions that execute for the user, for example, in the database system read, write, insert, delete or update.

B. Team-Based Access Control (TMAC) Model

This model introduces the concept of team in a collaborative environment by applying the RBAC model. The team consists of a group of users with their assigned role. The team must perform their assigned activity or task. It is a more efficient model because it can assign permission to user in time in a group fashion and support higher degree security. This model plays a very important role in context information related to collaborative activity and can apply this context to decide on permission access. According to TMAC model, the team has two context elements, first one is user context and the second one is object context. The user context is the current user of the team and the object context is the groups of objects that are needed by the team to complete the activities and goals. There are two key directions of the team based access control model. Fig. 2 represents the C-TMAC model components.

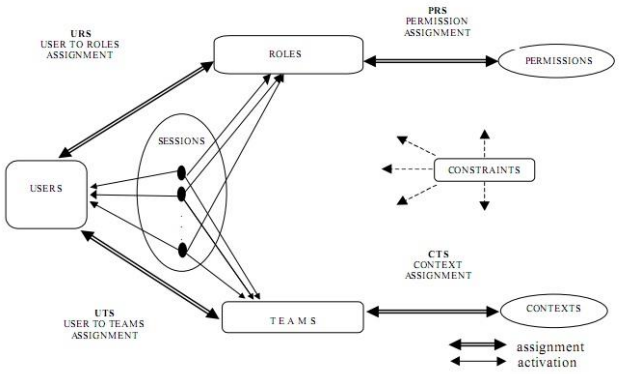


Fig. 2. C-TMAC [4]

Context aware-TMAC (C-TMAC) is an extended version of the TMAC model has five sets of elements that are users, roles, permissions, teams, and contexts. This model also has a set of sessions. It assigns users to the roles and permission to the roles. A team is a group of users, and every user can be a member of one or more teams. There are many to many relations between the role and team through user sessions. This model also has different kinds of sessions between user teams and active roles. Permission assignment is the permission that is given to roles for performing an operation on objects. Permissions are compliance of a particular mode of access to one or more objects. User assignment (UA) and permission assignment (PA) have many-to-many relationships between user-roles and role-permissions respectively. A user can have many roles, and a role can be assigned to many users. Similarly, a role may have many permissions and the same permission can be assigned to many roles. Contextual information examples such as locations and time intervals can be used while granting and denying access. The team theory is used as a system that connects users with contexts.

C. Privacy-aware Role-Based Access Control (P-RBAC) Model

Privacy is one of the important issues in software technology and has received increasing attention from users, companies, and researchers. The privacy protection can only be achieved by forcing privacy policies within an organization. The conventional access models; Mandatory Access Control (MAC), Discretionary Access Control (DAC), and the RBAC model are not made to force the privacy policies and almost not meet privacy and safety requirements. The data collected for one purpose should not be used for another purpose without the approval of the owner of data. The importance of purposes, conditions, and obligations originates from the protection of privacy and personal information. Obligations are the operations to be performed after an operation has been executed on data objects, are essential for some cases.

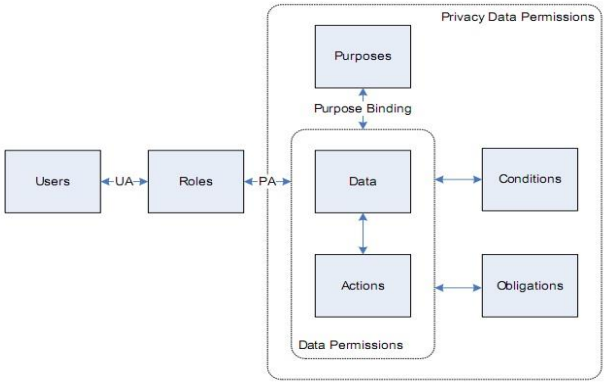


Fig. 3. PRBAC [3]

The Core P-RBAC model has the following set of elements as shown in the Fig. 3: Users, Roles, Data, Actions, Purposes, Obligations, and Conditions. Data are like object. An action is the set of instructions that executes data object for the user. The type of actions depends on the type of system that to be implemented. This model also introduced three new notions purposes, conditions and obligations. In the Core P-RBAC model, permissions are allocated to roles and users get the permissions by being allocated to roles. Conditional access is granted to users using the Conditions data element. Obligations are the conditions that need to be fulfilled after the data access is granted.

D. The Dynamic Sharing and Privacy-Aware RBAC (DySP-RBAC) Model

Collaborative information sharing environment requires better information sharing among users while privacy laws require the protection of user's information from unauthorized access and usage. Keeping this trade-off in view, there is a need for a flexible and better information sharing model that preserves the privacy of user's information.

The DySP-RBAC model extends the RBAC model to integrate sharing and privacy related requirements as shown in the Fig. 4. This model defines the following set of elements: Team, Task, Object, User, Role, Session, Permissions Collaborative relationships, Access level, and three privacy elements; Purpose, Condition, and Obligations. A team is a group of users that performs a specific task. For enhanced sharing, this model defines the sharing elements Collaborative Relationship and Access Level.

In this model action is an executable image of a program that can be used to execute to perform some activity. Permission is an operation allowable on an object. The elements in this model that control the level of data object sharing among collaborating users are Access Level and Collaborative Relationship. Collaborative Relationship element limits the sharing of data objects to only those users who are in a collaborative relationship with each other and Access Level element is used to share only a specific level of information.

The DySP-RBAC model helps in enhanced sharing and is applicable in most collaborative scenarios.

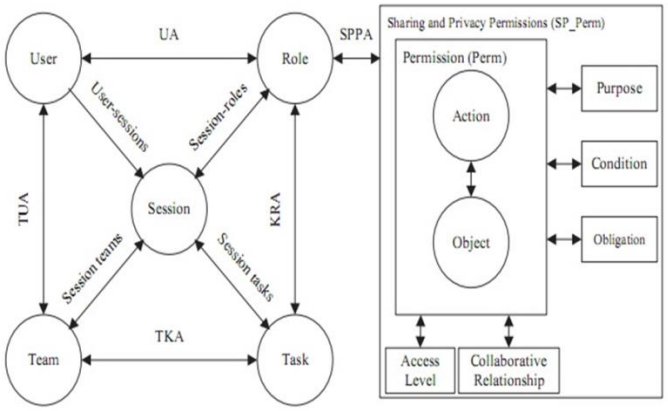


Fig. 4. DySP-RBAC MODEL [2]

IV. Methodology and Implementation

As mentioned earlier, four collaboration based RBAC models are selected for this research and a prototype of each model is implemented using PHP and XAMPP database. Further, these models have been evaluated and compared based on the performance and information sharing metrics. The postman application is used to find out the response time and permissions grants.

Standard metrics for comparison are selected from a list of metrics which are provided by the NIST standard. The metrics are response time, permissions, grants and denial based on several queries. The experiments used the prototype implementations of the RBAC models by comparing the rules and policies for the access control systems. The following parameters are used for the comparison of these collaborative RBAC models.

One of the metrics is performance, which is calculated based on the response time of every model. For this purpose, scenarios are created for each model. Only three data elements, that are common in all models, are selected for comparison of all models using performance metric. These three data elements include role, object, and operation. Using the access control rules based on these three elements, the models are evaluated and compared based on permissions and

response time metrics. The permission metric is related to the number of permissions (access rules) relevant to the query and response time metric measures the query response time.

In the scenario for the standard RBAC model, users are assigned to roles. Every role has permissions and user can request for the permission. Permission assignment is the permission that is given to roles for performing an operation on objects. Users, roles, objects, operations, and permissions are defined. For the experiment, 25000 permissions are generated for this model and 25 queries are executed to find out its performance and relevant permissions. For the TMAC model, 25 teams are created in this scenario. Each team has a group of users with their assigned roles. For P-RBAC model, a few more elements are used, those are purposes, conditions, and obligations. Whereas in DySP-RBAC model scenario, there are numerous elements including users, task, team, role, obligation, access level, collaborative relationship condition object, and operations.

Another metric is sharing which is used as a comparison parameter. Using this metric, permission grants of every model are found to check the sharing of collaborative RBAC models. There are 25 queries executed for each model. For sharing metric comparison, the data elements used for each model are listed here. For standard RBAC, only three data elements role, object, and operation are considered in sharing scenario. In TMAC model four parameters role, object, team, and operation are examined. The P-RBAC model uses six data elements including role, object, purpose, condition, obligation and operation. Moreover, role, object, team, task, purpose, condition, obligation and operation are used in the DySP-RBAC model.

V. Results and Discussion

The performance of the four RBAC based models is calculated based on response time, as shown in Table 1. The response time of each model is calculated for an equal number of permissions.

TABLE. I. Queries Vs Response Time

No. of Queries	Response Time of RBAC Models			
	RBAC	TMAC	PRBAC	DySP-RBAC
1-5	1419	1531	1666	1751
6-10	1354	1424	1600	1712
11-15	1418	1512	1605	1764
16-20	1369	1491	1571	1799
21-25	1376	1487	1575	1802

The Fig. 5 shows the response time of all collaborative RBAC models for the sum of five queries. Response time is increasing with the increase in queries. The point to be noted here is that the standard RBAC model has the minimum response time than all other models whereas the DySP-RBAC model has maximum response time than other collaborative RBAC models.

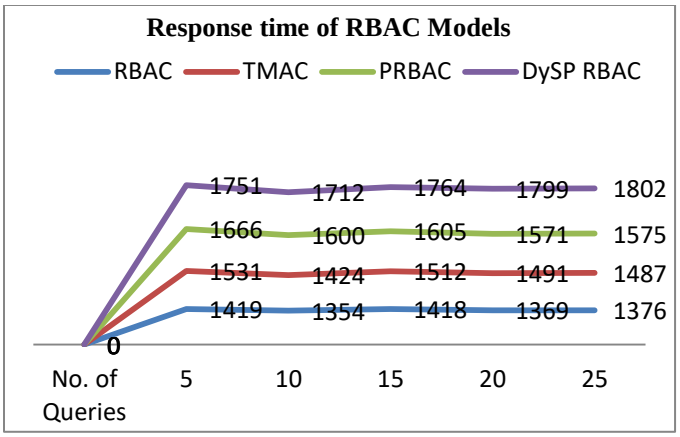


Fig. 5. Response Time Comparison of RBAC Models

Table 2 shows the total running time of 25 queries for all RBAC models.

TABLE. II. Running Time Comparison

RBAC MODELS	RBAC	TMAC	PRBAC	DySP-RBAC
Runtime	6,936	7,445	8,017	8,828

The Fig. 6 shows the response time of all collaborative RBAC models graphically for 25 queries. Standard RBAC model has the minimum running time for queries than other models.

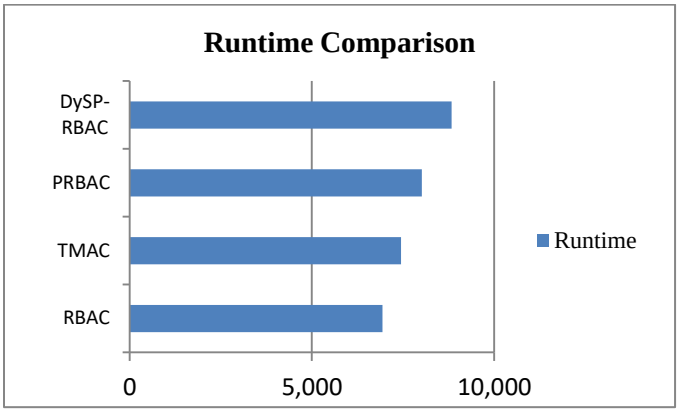


Fig. 6. Running Time Comparison of RBAC Models

The sharing of four RBAC based models is calculated using permissions grant as shown in Table 3. The DySP-RBAC model has the maximum number of permissions grant while executing different sets of queries than all other models, so it can be said that this model is best data sharing model.

TABLE. III. QUERIES VS PERMISSIONS GRANTED

No. of Queries	Permissions Grant of RBAC Models			
	RBAC	TMAC	PRBAC	DySP-RBAC
1-5	835	869	1228	2058
6-10	893	928	1334	2433
11-15	846	894	1284	2338
16-20	821	866	1284	2317
21-25	843	886	1276	2370

The standard RBAC model has the minimum number of permissions grants than other models and DySP-RBAC has granted the most number of permissions for given set of queries as shown in Fig. 7.

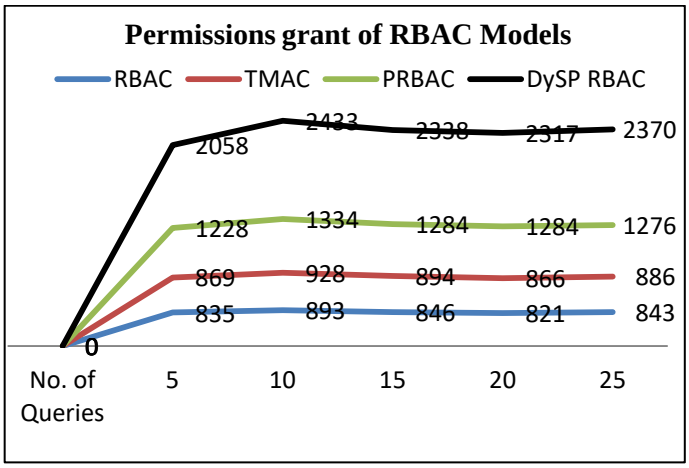


Fig. 7. Permission Grant Comparison

Further, P-RBAC is also good in flexibility for granting permissions than RBAC and TMAC. Even TMAC outperforms standard RBAC in this case. Table 4 and Fig. 8 both represent the permissions granted for all 25 queries for all RBAC models.

TABLE. IV. PERMISSION GRANT COMPARISON

RBAC MODELS	RBAC	TMAC	PRBAC	DySP RBAC
Permission Grant	4,238	4,443	6,406	11,516

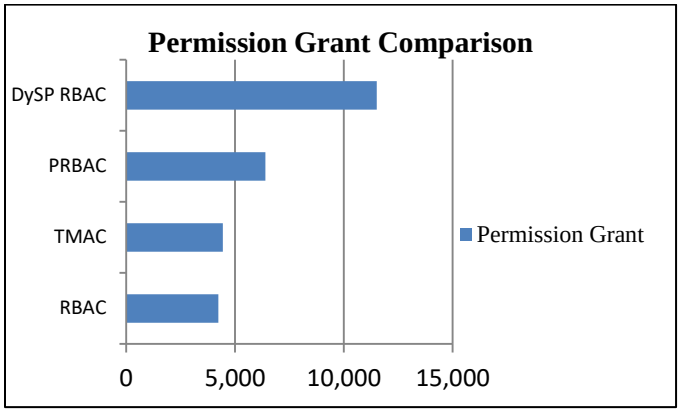


Fig. 8. Permission Grant Comparison of RBAC Models

After the implementation and comparison of four RBAC based collaborative models, it is found out that this research is helpful to explain which RBAC model is better to use for which purposes and in which collaborative environment.

The evaluation of the models predicts that the standard RBAC model is better in performance as compare to other models but less suitable for sharing. The standard RBAC model does not work well in collaborative scenarios. The TMAC model is suitable in the environment where teamwork is involved and can give better performance in sharing as compared to standard RBAC but with more response time. The organizations where the privacy is the key point in sharing data, the most applicable model is P-RBAC model which outperforms standard RBAC and TMAC in privacy and sharing scenarios. The DySP-RBAC model is more suitable in collaborative scenarios and sharing while having maximum response time due to the use of many sharing and privacy data elements. If somebody emphasizes on sharing whatever the response time is, she may opt the DySP-RBAC model. It depends on user requirements and their environment to consider which RBAC model is to be selected.

VI. CONCLUSION

In this paper, four collaborative RBAC based models have been selected for comparison. A prototype for each of these models is implemented and compared based on performance and sharing metrics. After comparison and analysis, it is found that which RBAC model is better to use for which purposes and in which collaborative environment. The performance and sharing of all the models is calculated based on response time and permissions grant. In the end, results are discussed in the form of graphs.

In future, we would like to further work on the RBAC models and try to elaborate their implementation and significance in inter and intra-organizational structures. It would be interesting to provide a complete picture of privacy-aware RBAC models which are more suitable for different collaborative environments. It is also intended to implement an extended version of the DySP-RBAC model for Cloud systems.

ACKNOWLEDGEMENT

This research is funded by *Higher Education Commission (HEC),* Pakistan, through the project *"Information Privacy in Collaborative Environments" approved for CIIT, Islamabad.*

REFERENCES

[1] D. F. Ferraiolo, R. Sandhu, S. Gavrila, D. R. Kuhn, and R. Chandramouli, "Proposed NIST standard for role-based access control", ACM Trans. Inf. Syst. Secur, vol. 4, no. 3, pp. 224-274, 2001.

[2] A. K. Malik, and S. Dustdar,"Enhanced Sharing and Privacy in Distributed Information Sharing Environments", in Proceedings of Information Assurance and Security (IAS), Malacca, Malaysia, pp. 286-291, 2011.

[3] Q. Ni, A. Trombetta, E. Bertino, and J. Lobo, "Privacy-aware role based access control", In Proceedings of symposium on Access control models and technologies (SACMAT), New York, NY, USA, pp. 41-50, 2007.

[4] C. K. Georgiadis, I. Mavridis, G. Pangalos, and R. K. Thomas, "Flexible team-based access control using contexts", in Proceedings of symposium on Access control models and technologies (SACMAT), New York, NY, USA, pp. 21-27, 2001.

[5] M.L. Damiani, H. Martin, Y. Saygin, M.R. Spada, and C. Ulmer. "Spatio-Temporal Access Control: Challenges and Application", in Proceedings of symposium on Access control models and technologies (SACMAT), New York, NY, USA, pp. 175-176, 2009.

[6] A. Kamoun, and S. Tazi. "A semantic role-based access control for intra and inter-organization collaboration", in Proceedings of IEEE Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE), Parma, Italy,pp. 86-91, 2014.

[7] B. Tang, Q. Li, and R. Sandhu, "A Multi-Tenant RBAC Model for Collaborative Cloud Services", in Proceedings of Privacy, Security and Trust (PST), , Tarragona, Catalunya, pp. 229-238, 2013.

[8] A. H. Anderson, "A comparison of two privacy policy languages: Epal and xacml", In Proceedings of Secure web services (SWS), New York, NY, USA, pp. 53–60, 2006.

[9] G. Cabri, L. Ferrari and L. Leonardi, "Agent Role-based Collaboration and Coordination: a Survey About Existing Approaches", in Proceedings of IEEE Systems, Man and Cybernetics, Hague, Netherlands, pp. 5473-5478, 2004.

[10] M. N. Kamel Boulos and S. Wheeler, "The emerging web 2.0 social software: an enabling suite of sociable technologies in health and healthcare education," Health Info Libr J, vol. 24, no. 1, pp. 2–23, 2007.

[11] A. K. Malik and S. Dustdar, "A hybrid sharing control model for context sharing and privacy in collaborative systems", in Proceedings of Information Networking and Applications (WAINA), Biopolis, Singapore, pp. 879–884, 2011.

[12] G. Ahn and R. Sandhu, "Role-based authorization constraints specification", ACMTrans. Infosys. Sec., vol. 3,no. 4, pp. 207-226, 2000.

[13] R.W.Baldwin, "Naming and grouping privileges to simplify security management in large databases", in Proceedings of IEEE Research on Security and Privacy, Los Alamitos, Calif, pp. 116–132, 1990.

[14] D. E. Bell, and L.J.L.Padula, "Secure computer systems: Unified exposition and MULTICS Interpretation", Tech.Rep. ESD-TR-75-306, the MITRE Corporation, Bedford, pp. 1-129, March,1976.

[15] D. Ferraiolo and R. Kuhn, "Role Based Access Control" in Proceedings of National Computer Security (NCSC), Baltimore, Maryland, pp. 554-563, 1992.

[16] M. Nyanchama and S.L. Osborn. "Access rights administration in role-based security systems", in Proceedings of IFIP WG11.3 working conference on database security, Amsterdam, The Netherlands, pp. 37-56, 1994.

[17] D.F. Ferraiolo, J. A. Cugini, D.R. Kuhn, "Role Based Access Control: Features and Motivations", in proceedings of Computer Security Applications, New Orleans, LA, pp. 241–248, 1995.

[18] R. S. Sandhu, "Role Hierarchies and Constraints for Lattice Based Access Controls", in Proceedings of European Symposiums on Research in Computer Security, Rome, Italy, pp. 65-79, 1996.

[19] M.A. Madani, M. Erradi, and Y. Benkaouz, "A Collaborative Task Role Based Access Control Model", Journal of Information Assurance & Security, vol. 11, no. 6, pp. 348-358, 2016.

[20] R. S. Sandhu, E. J. Coyne, H. L. Feinstein, and C. E. Youman. "Role-Based Access Control Models", Computer, vol. 29, no. 2, pp. 38-47, 1996.

Measuring the Impact of the Blackboard System on Blended Learning Students

Thamer Alhussain
E-Commerce Department
Saudi Electronic University
Riyadh, Saudi Arabia

***Abstract*—With the advantages of using learning management systems (LMS) such as Blackboard in the educational process, assessing the impact of such systems has become increasingly important. This study measures the impact of the Blackboard system on students at Saudi Electronic University (SEU) in order to help improve the quality of existing learning environment. For this assessment and measurement, the IS-Impact Measurement Model is used, since it is the most comprehensive model that is valid in the context of this study. The results of this paper indicate how Blackboard is influencing individual performance. It concludes that the use of the Blackboard system has a positive impact on individuals.**

***Keywords*—learning management system; IS impact/measurement; Blackboard; blended learning**

I. Introduction

The rapid growth of information and communication technologies provides unique opportunities for e-learning to improve the educational process. Leaning management systems (LMS), such as WebCT and Blackboard, are used by many universities and educational institutions to provide and improve learning. LMS take advantage of new technologies to make learning available and accessible anywhere in the world and at any time. According to [2], the success of e-learning can be attributed to the availability of LMS. LMS, which are also known as virtual learning environments (VLE) or learning platforms, enable educational institutions "to develop electronic learning materials for students, to offer these courses electronically to students, to test and evaluate the students electronically, and to generate electronically student databases in which student results and progress can be charted" [2, p. 2]. The use of LMS helps learners to easily keep track of their courses, and instructors to simply evaluate and track each learner.

Given the value of LMS, assessing the success and impact of LMS has become increasingly important to improve the quality of the educational process, especially in view of the fact that many studies, such as Aceto et al. [3], Wang et al. [4], and Alkhalaf [5], highlight the need for measuring and evaluating e-learning systems.

This research is situated within the field of information systems, and measures the impact of the Blackboard system adopted at Saudi Electronic University (SEU). More specifically, this study uses the IS-Impact Measurement Model [1] to assess the impact of the Blackboard system on students in order to help improve the quality of the existing learning environment at SEU as the only university in Saudi Arabia that has adopted a blended learning style. This paper is structured as follows. The relevant literature is reviewed followed by the description of the empirical research that involved a descriptive survey of the students in SEU. Finally, the results and conclusions are presented.

II. Blackboard as a Learning Management System

LMS have been defined differently by various scholars. Ayub and others [6] gave a precise definition and termed LMS as a Web-based technology that aids in the design, distribution and assessment of a certain process of learning. An LMS is basically software that has been designed to guide the entire learning process as well as provide learning resources to the learners. It can also be described as a set of tools and framework that enable easy creation of Web content while guiding learning [7]. Wahlstedt and Honkaranta [8] affirm that LMS are an advancement of traditional learning, since they comprise instructional devices, learning content, and evaluation devices. What is unique about LMS is that they can be used to plan, convey, and manage learning, thus combining various tasks earlier distributed to different stakeholders. Management tasks of LMS include delivery, examinations, statistical analysis, and virtual classes [7]. According to Paulsen [2], "Learning management systems manage the log-in of registered users, manage course catalogs, record data from learners and provide reports to the management." It is, therefore, a crucial tool in educational institution management because it basically brings all these facets on board.

An LMS is a crucial platform where learners and their instructors can interact and simultaneously share learning materials. An LMS can, therefore, be regarded as an advanced Internet-based technology solution for both the learners and the instructors because it allows the two parties to connect with the help of interactivity features such as forums, file-sharing platforms, and thread discussions [7]. An LMS can be used by an instructor to distribute course material while aiding in instructor–learner interaction [9]. The management function of LMS is particularly of great importance because it requires less effort and saves time that would otherwise have been wasted by the instructor without changing the entire instructional process. Threaded discussions, video conferencing, and discussion forums are key characteristics of LMS [7]. These features allow for an interactive learning environment.

LMS have a tremendous effect on e-learning. According to Paulsen [2], the presence of an LMS determines how e-

learning will succeed. With an LMS in place, an institution can easily develop Web content, teach electronically, evaluate learners electronically, and generate learners' databases through which the learners can access their results [2].

Despite being helpful in aiding e-learning, there has been a gap between reality and other advanced instructional tools, such as the multimedia type that are believed to be of help in instruction [9]. On many occasions, these multimedia tools are not normally used, or if they are used, instructors do not exploit them fully. For example, many institutions are currently using LMS to facilitate e-learning but instructors limit themselves to uploading course materials and barely use the other features, such as discussion forums [9]. Other users have been discouraged by the fact that they do not receive immediate feedback from features such as email [9]. Although these interactive features have been included in LMS, their use may still be restricted by the commitment of both parties. LMS can be used to bridge the gap that exists between reality and advanced instructional tools. This can only be possible if the LMS is built to be more adaptive and customizable [9]. Building an adaptive and customizable LMS will help in ensuring that learners and instructors with different levels of computer literacy are accommodated.

The Blackboard system is Web-based software that features a customizable open architecture for course management that permits amalgamation with student information systems and authentication protocols. This system may be installed on local servers or hosted by Blackboard ASP solutions, and its core purposes are to develop completely online courses with a few or no face-to-face meetings and to add online elements to courses that are delivered conventionally face to face. The Blackboard Learning System provides users with a platform for sharing content and communication [9]. With regard to communication, the Blackboard system enhances announcements, that is, instructors can post items for learners to read. Such announcements may be created as pop-up messages or via the announcement available in the Blackboard system. A discussion feature makes it possible for professors and students to create discussion threads and offer feedback. The chat function in the Blackboard system allows learners to converse and share ideas. Lastly, the Blackboard mail allows students and teachers to send mail to each other or to groups. The learning modules feature allows professors to post various lessons for students to access. Instructors can also post assignments and receive assignments via the assessment tab. Teachers and professors can use the grade book feature to post grades for students to view. Lastly, videos and other media can be posted under the media library function.

III. Saudi Electronic University (SEU)

A royal decree was issued by King Abdullah Bin Abdul-Aziz, the custodian of the Two Holy Mosques, on October 8, 2011 to launch the Saudi Electronic University (SEU) as a government educational institution. The SEU is the only specialized university in blended learning in the Kingdom of Saudi Arabia, and it offers both graduate and undergraduate degree programs along with lifelong education. The goals of the university are to represent the nation and to compete with other international universities, to present a flexible and distinguished example of higher education, to support self-learning skills and offer knowledge, to offer higher education based on the best applications and technologies of e-learning, to transfer and localize knowledge, and to support the mission and the concept of lifelong e-learning and blended education for all members of Saudi society.

SEU adopted a blended learning pattern, which is the latest style of learning used in universities around the world. It is based on the combination of 25% direct traditional education with 75% e-learning using virtual classrooms, educational forums, and interactive activities. This style of learning is based on self-discipline and leadership for self-learning. Moreover, it considers the student to be the focus of the educational process, and he or she is the initiator and the leader. The teacher's role is to motivate and direct the educational process.

The adopted blended learning pattern in SEU combines the features of both traditional education and e-learning in an integrated model that obtains the maximum benefit from the technology and the means available to each of them in order to achieve the desired optimal learning objectives.

IV. The IS-Impact Measurement Model

The IS-Impact Measurement Model proposed by Gable, Sedera, and Chan in 2008 has always been regarded as a comprehensive and valid IS success measurement model [10]. Gable et al. [1] proposed a definition of the IS-impact of an information system (IS) and they defined it as "a measure at a point in time, of the stream of net benefits from the IS, to date and anticipated, as perceived by all key-user groups" [11]. This model was designed based on the work of DeLone and McLean [12], and it corrects the setbacks of the DeLone and McLean IS success model.

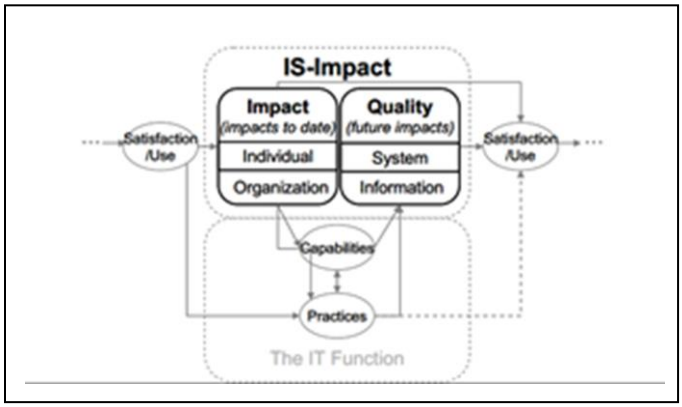


Fig. 1. IS-Impact Measurement Model [1].

The IS-Impact Measurement Model differs from the old DeLone and McLean IS Measurement Model in five ways: 1) it reflects a true measurement model rather than the causal/process model depicted by the D&M model; 2) the use of dimensions has been omitted; 3) the aspect of satisfaction is seen as a measure of success rather than a dimension of success; 4) the modern IS context has been taken into consideration through the inclusion of new measures; and 5) additional measures have been added to deeply examine organizational dimension [11].

As illustrated in Figure 1, within the IS-Success/Impact framework, the success and impact of an IS system can be measured in terms of the quality of the information produced (information quality), the performance of the system from a technical perspective (system quality), the impact on individual users (individual impact), and the impact on the relevant organization (organizational impact).

The IS-Impact Measurement Model was selected because it comprehensively takes into account the evaluation of information systems through comprising 37 measures in four important dimensions of the system: "System Quality," "Information Quality," "Individual Impact," and "Organizational Impact." As such, it is a more comprehensive and valid model for use. According to Rabaa'i [11], this model has been tested statistically though surveys; it has proven to be valid and it employs perceptual measures. Such tests depict the validity and reliability of this model. Despite borrowing heavily from the DeLone and McLean model by adopting its constructs, it has succeeded in employing them for a different purpose [1]. The model and approach employ perpetual measures, aiming to offer a common instrument answerable by all relevant stakeholder groups, thereby enabling a combination or a comparison of stakeholder perspectives [10].

Moreover, a study conducted by Alotaibi [13] validated the IS-Impact Measurement Model and emphasized the completeness and validity of IS-Impact Measurement Model as a hierarchical multidimensional formative measurement model in the Saudi Arabian context. Accordingly, this model has been adopted in this research owing to its strengths in comparison to other models. It is quite clear that this model has eliminated all the weaknesses of the other models by including and reviewing their constructs.

V. Individual Impact

This paper will focus on measuring the impact of the Blackboard system on blended learning students as individuals. As stated by Gable et al. [1; p. 289], "The 'individual impact' is a measure of the extent to which [the IS] has influenced the capabilities and effectiveness, on behalf of the organization, of key-users." Based on the IS-Impact Measurement Model [1], the variables for the construct of "individual impact" are the following:

- I have learned much through the presence of Blackboard.
- Blackboard enhances my awareness and recall of job-related information.
- Blackboard enhances my effectiveness in the educational process.
- Blackboard increases my productivity.

Accordingly, individual impacts are concerned with how the Blackboard system influences individual performance. The hypothesis of this construct is that the Blackboard system used in SEU has a positive impact on the individual.

VI. Methodology

This study adopts a positivist paradigm of research that seeks to test theories, verify hypotheses, and investigate the real world as it exists [14, 15]. This paper is a part of the research that will evaluate and measure the use of the Blackboard system adopted at SEU by testing the IS-Impact Measurement Model developed by Gable, Sedera, and Chan in 2008 (Figure 1). In particular, this paper measures the impact of the Blackboard system on blended learning students by using a single case design. This is because the use of a single case design is more suitable for research that aims to test a theory, anomaly, or special case [16]. This research will use a single case design to delve more deeply into the phenomena in order to insure that a rich description and understanding is provided. It will use a case study to help achieve the aim of the research, which is to evaluate and measure the impact of the Blackboard system adopted at SEU for the purpose of improving the quality of the existing learning environment. As mentioned by Benbasat, Goldstein, and Mead [17, p. 370], "A case study examines a phenomenon in its natural setting, employing multiple methods of data collection to gather information from one or a few entities (people, groups, or organizations)."

For the purpose of this research, a questionnaire was used for data collection. The questionnaire was designed based on the IS-Impact Measurement Model [1]. It was distributed to blended learning students at SEU, including males and females from all branches of the university. The questionnaire included two main sections. The first intended to collect demographic information on the respondents, while the second included the 37 measures of the IS-Impact Measurement Model in four important dimensions: System Quality, Information Quality, Individual Impact, and Educational Impact. As discussed earlier, this paper is focused on the individual impact, which includes four variables to test construct validity. The researcher distributed 2256 questionnaires (to 203 master's students and 2053 bachelor's students). Of these, 447 were returned by the participants. The participants answered the questions on a scale of 1 to 5, where 1 represented "Strongly disagree"; 2, "Disagree"; 3, "Neutral"; 4, "Agree"; 5, "Strongly agree."

VII. Results

All questionnaire responses were stored in the SPSS (Statistical Package for the Social Science) software, which was used for the analyses. Statistical analysis included the frequency and the percentage of each variable, the chi-square value, and its level of significance. As mentioned earlier, only the survey questions that measure the impact of the Blackboard system on blended learning students were included.

It is noteworthy that the total sample of the survey consists of 447 participants, comprising 48 male students and 399 female students. The highest percentage (89%) of the participants were studying for their bachelor's degree, while the lowest percentage (11%) were postgraduate students. All of the participants had at least one year of experience with the Blackboard system

TABLE I. RELATIVE NUMERICAL DISTRIBUTION AND BASIC STANDARDS, INCLUDING THE CHI-SQUARE VALUES OF VARIABLES RELATED TO INDIVIDUAL IMPACT

Item	Strongly disagree		Disagree		Neutral		Agree		Strongly Agree		M	SD	X^2	Relative weight	Order
	f	%	f	%	f	%	F	%	f	%					
1	10	2.2	32	7.2	78	17.4	237	53.0	90	20.1	3.82	.912	352.519**	3.58	1
2	8	1.8	36	8.1	74	16.6	246	55.0	83	18.6	3.81	.894	383.436**	3.57	2
3	11	2.5	46	10.3	92	20.6	217	48.5	81	18.1	3.70	.964	272.810**	3.46	3
4	15	3.4	55	12.3	125	28.0	185	41.4	67	15.0	3.52	.999	197.172**	3.3	4

[a.] * denotes significance at 0.01

[b.] ** denotes significance at 0.05

[c.] Items: 1. I have learned much through the presence of Blackboard; 2. Blackboard enhances my awareness and recall of job-related information; 3. Blackboard enhances my effectiveness in the educational process; 4. Blackboard increases my productivity.

The survey results clearly indicate that only 2% responded that they strongly disagree that the presence of the Blackboard system has helped them to learn much, while 7.2% disagree on the same, and 17% remained neutral. On the other hand, the majority (53%) of the students agree that the presence of the Blackboard system has helped them to learn much and 20% of the respondents strongly agree that the presence of the Blackboard system has helped them to learn much. It is evident from the data collected that 73% of the students say that the presence of the Blackboard system has helped them to learn much, whereas 9.2% say that the presence of the Blackboard system has not helped them to learn much, and 17% remain neutral on the subject.

As evidenced from the data analysis, only 9.1% of the students perceive that the Blackboard system has not enhanced their awareness and recall of relative information, while the majority (73%) agree that the Blackboard system has enhanced their awareness and recall of relative information. However, 16 % are neutral on the same question. With regard to item number 3, a low percentage (12%) of the respondents indicate that the Blackboard system has not enhanced their effectiveness in the educational process, while a high percentage (66%) of the students either agree or strongly agree that the Blackboard system has enhanced their effectiveness in their educational process, and 20% of the respondents remain neutral on this matter. On the last question, only 15% of the students disagree that the Blackboard system has increased their productivity, while 28% of the students neither agree nor disagree, and a high percentage (56%) of the students believe that the Blackboard system has increased their productivity.

All this can be generalized to the whole population of the students, since the standard deviations are very small and the chi-square statistic on all the answers given by the respondents are significant.

VIII. CONCLUSION

This paper reports on measurements of the impact of the Blackboard system on blended learning students. This measurement and evaluation was based on the IS-Impact Measurement Model developed by Gable, Sedera, and Chan in 2008. Results of this research support a number of findings reported in literature regarding the impact of LMS on individuals. Analysis of the results shows that the use of the Blackboard system has a positive impact on the individual at SEU. The analysis of the results indicates that using the Blackboard system has helped students to learn much and increased their ability to interpret and recall relative information. The findings also highlight that the Blackboard system has enhanced students' effectiveness in the educational process and has increased the overall productivity of students in the learning process.

REFERENCES

[1] Gable, G., Sedera, D. & Chan, T., "Reconceptualizing information system success: the IS-impact measurement model". *Journal of the Association for Information Systems*, *9*(7), 377-408, 2008.

[2] Paulsen, M. F., "Experiences with learning management systems in 113 European institutions". *Educational Technology & Society*, *6*(4) 134-148. Retrieved from http://ifets.ieee.org/periodical/6_4/13.pdf, 2003

[3] Aceto, S., Delrio, C., Dondi, C., Fischer, T., Kastis, N., Klein, R., "e-Learning for Innovation: Executive Summary of the Helios Yearly Report 2007". Brussels: MENON Network EEIG, 2007.

[4] Wang, Y., Wang, H., & Shee, D., "Measuring e-learning systems success in an organizational context: Scale development and validation". *Computers in Human Behavior*, *23*(4), 1792-1808, 2007.

[5] Alkhalaf, S., "Creating effective e-learning systems for higher education in Saudi Arabia", PhD thesis, Griffith University, Australia, 2013.

[6] Ayub, A. F. M., Rohani, A. T, Wan, M. W. J., Wan, Z. W. A., & Luan, W. S., "Factors influencing students' use [of] a learning management system portal: perspective from higher education students", *International Journal of Education and Information Technologies*, *4*(2), 100-108, 2010.

[7] Adzharuddin, A. & Ling, H. L., "Learning management systems among university students: does it work?", *International Journal for e-Education, e-Business, e-Management, and e-Learning*, *3*(3), 248-251, 2013.

[8] Wahlstedt, A. & Honkaranta, A., "*Bridging the gap between advanced distributed teaching and use of learning management systems in the university context"*, Seventh IEEE International Conference on Advanced Learning Technologies (ICALT), 2007.

[9] Almarashdeh, A., Sahari, N., Zin, M. & Alsmadi, M., "The success of learning management system among distance learners in Malaysia Universities", *Journal of Theoretical and Applied Information Technology*, *21*(2), 80-91, 2010.

[10] Elias, N. F. & Cao, L., "Validating the IS-impact model: two exploratory case studies in China and Malaysia", *Pacific Asia Conference on Information Systems (PACIS) 2009 Proceedings*.

[11] Rabaai, A. A. & Gable, G., "*Extending the IS-impact model into the higher education sector"*, research in progress, Queensland University of Technology, Brisbane, 2009.

[12] DeLone, W. & McLean, E., "Information systems success: the quest for the dependent variable", *Information Systems Research*, *3*(1), 60-95, 1992.

[13] Alotaibi, N., "Extending and validating the IS-impact model in Saudi Arabia: accounting for computer network quality", PhD thesis, Queensland University of Technology, 2012.

[14] Guba, G. & Lincoln, Y. S., "Competing paradigms in qualitative research", In N. K. Denzin and Y. S. Lincoln (Eds.), *Handbook of qualitative research.* Thousand Oaks, CA: SAGE Publications, 1994.

[15] Walliman, N., "*Social research methods",* London: SAGE Publications, 2006.

[16] Yin, R. K., "*Case study research: design and methods",* (4th ed.). Thousand Oaks, CA: SAGE Publications, 2009.

[17] Benbasat, I., Goldstein, D. K. & Mead, M., "The case research strategy in studies of information systems", *MIS Quarterly*, *11*(3), 369-386, 1987.

Design and Architecture of a Location and Time-based Mobile-Learning System: A Case-Study for Interactive Islamic Content

Omar Tayan[1, 3], Moulay Ibrahim El-Khalil Ghembaza[2, 3], Khalid Al-Oufi[1]
Dept. of Computer Engineering[1], Dept. of Computer Science[2],
College of Computer Science and Engineering,
IT Research Center for the Holy Quran and Its Sciences (NOOR) [3],
Taibah University, Madinah
Kingdom of Saudi Arabia

***Abstract*—This paper describes a software design, architecture and process of a novel mobile-learning (m-Learning) approach based on smart-phone devices for retrieving relevant content in real-time based on the user's-location and the current-time and presenting the content to the user in a manner that support portable learning on-the-move. Mobile-learning using Islamic content is used as a case-study of the proposed system, which can easily be adapted for other learning-content. The proposed system is highly interactive and frequently purges the host-device for details of the current user-location and current-time (e.g. the time, day and month in the solar and the lunar calendars) before such details are used to retrieve the most relevant content. In this study, Quranic-verses, corresponding interpretations (Tafseer) and Hadith (Prophetic words or actions) relates to the online content being fetched in this application. For example, a user may be performing some travel/pilgrimage during Ramadan, in which the relevant content/teachings (based on the user's location and the current time) are presented to the user in a timely manner in order to learn the rituals of that day or location. The information fetched is then presented and displayed in an interactive user-friendly manner. A summary and comparative analysis of some related applications is presented, showing the limitations of other m-Learning applications and demonstrating the new contribution of our architecture design. Finally, the described system allows authorized scholars to upload and report Islamic-decrees made in real-time based on findings/experience at a particular time and/or location-of-interest (e.g. the new rulings/decrees are then published online in real-time). It is anticipated that millions of end-users shall benefit from the proposed system through the benefits of fast, highly-accessible, user-friendly and relevant information retrieved online in real-time. Advantageously, the potential application and large impact of the proposed m-Learning approach for use with other learning-content/courses is notable.**

Keywords*—Mobile-Learning; Context-Aware Notifications; Time and Location Based Reporting; Interactive-Knowledge based User-Events and Activities; Design and Architecture

I. INTRODUCTION & BACKGROUND

Islamic applications for smart user-devices have become an integral part of our daily activities, particularly in times of worship and good deeds. For example, during the month of Ramadan, or during the months of Hajj, and other rituals, it would be beneficial if such applications could be employed in an effective way for the user. Today, such Islamic applications are numerous and can be found in online stores for Apple iOS devices (App Store), Google Android devices (Play Store), Microsoft Windows 10 Mobile devices (Windows Store), and BlackBerry 10 devices (BlackBerry World). Many of those applications now include a rich set of functions such as; tools, services, and useful Islamic content, and are free of charge in most cases.

With the arrival of the blessed month of Ramadan or during other Holy events, it would be useful to utilize smart-phones as support tools to assist in the remembrance of God, and assist users to perform their worship and invocations efficiently with an improved level of learning and understanding by providing relevant-knowledge based on particular user-events encountered. To achieve this goal, this paper considers the characteristics of each ritual/act-of-worship in order to retrieve the most relevant verses, supplications and invocations according to the current user-environment. Significantly, this paper proposes an interactive Islamic application which helps to perform acts of contemplation/worship and remembrance. Moreover, the proposed application helps the user to understand the association between events, times and places, as well as the corresponding Islamic knowledge/rulings by retrieving the relevant Quranic verses, interpretations, and related Prophetic Hadiths (Prophetic Sayings) based on the user's current time and geographical location.

The proposed application supports automatic notifications as periodic reminders of supplications on a daily basis, so the user remembers to mention God at any time or place, in compliance with the verse: "And remind, for indeed, the reminder benefits the believers" [Chapter51, Verse55], and "So remember Me; I will remember you. And be grateful to Me and do not deny Me" [Chapter2, Verse152].

Essentially, this paper presents and focuses on the link between Quranic verses and Prophetic Hadiths on one hand, and a range of events and religious rituals on the other hand, which in turn are associated with a particular time or place, or with both time and place simultaneously. This paper develops

an architecture in which the linkage between the user-activities and events is processed with all the relevant knowledge/information being presented using a smart-phone platform. Topics of events and rituals are numerous and include: holidays and worship, travel, decency, ethics, reciting Dhikr (devotional acts such as the remembrance of God), invocations, supplications, prayers and going to mosques, ablutions, fasting the month of Ramadan, acts of charity, Zakat (obligatory charity or alms-giving), Hajj and Umrah (major and minor pilgrimage), and reading/reciting the Holy Quran. Notably, such events and Islamic rituals are linked by specific times and/or geographic locations, or with both time and location conditions. Both factors play a vital role in the lifetime of Muslims.

II. Related Work

There are hundreds of Islamic applications that are compatible with smart-phones and tablets of various kinds, some of which can be downloaded for free, while others can be purchased. Today, the most popular applications are commonly selected according to the range of features and services provided for the end-user, which also includes; built-in reminders and alerts such as relevant supplications and virtues. The degree of content-driven notifications provided by other applications is the main relationship used for comparison with this study. Hence, the following discussion provides a general description of relevant applications as well as the key highlights of any time-based or location-based notification features supported in order to compare this work with the related state-of-the-art in related m-Learning/content-driven applications.

Classification and summary of relevant and popular applications

"**Fathkrony**" denoted in Arabic, which means "Remember Me" as denoted in Arabic [1], is an a smart-phone application providing remembrance-supplications (Dhikr), prayer times and a direction indicator of the Holy Kaabah (Qiblah) with an Arabic-only interface. This application contains the Holy Quran as an e-book, prayer times, a Qiblah direction indicator, daily remembrance supplications, invocations, an Islamic calendar, Ramadan calendar and a guide for pilgrims. Moreover, the application displays the prayer times and the times remaining before the next prayer commences. The user can also read the Quran and its interpretation, and provides the ability to search for specific verses and share it with friends via Facebook and Twitter. Supplication reminders are also provided after each prayer. What distinguishes "Remember-Me" from other applications is the possibility of scheduling the recitation of the Holy Quran, where you can choose the number of days to complete the recitation of the Quran. The application also calculates the daily Quran recitation progress. Some of the relevant notification features in this application includes: notifications of prayer Athans (call-to-prayer) and notifications for morning and evening supplications. However, some limitations in the application include; an Arabic-only interface, inaccurate Qiblah direction indicator, and no means for listening to the recitation of the Quran or supplications.

"**AlMosally**" [2] is a smart-phone application that searches for nearby mosques and provides remembrance-supplications. This application is freely available for iOS and Android devices and provides a tool that searches for nearby mosques. A map-based interface is also used to find the most accurate routes. The user can also access a number of other functions, including: prayer time alerts, Quranic supplications, fasting supplications, the direction of the Qiblah from any location, and finally, the Hijri (Islamic) calendar. One of the main advantages that is unique to this application is the "silent mode during prayer" times feature, and the feature to "show the nearest mosque" using the GPS functionality. Similar to [1], this application had provided notifications for Athan times and morning/evening supplications. The drawbacks found in this application include; lack of support for the supplications feature (limited for some functionality), and that there is no possibility of listening to the supplications or the recitation of the Quran.

"**Mutawef**" [3] is an Arabic-only smart-phone application dedicated for assisting pilgrims with the rituals of Hajj and Umrah (e.g. the major and minor pilgrimage to the Kaabah/House-of-God). This application provides a complete guide for pilgrims visiting the House of God. Essentially, it informs users of important places to visit as part of the rituals using live pictures and maps for ease and convenience. It provides the user with supplications used during Hajj and displays descriptions of Hajj and Umrah, with details of common mistakes to avoid.

Some of the main features of this application include:

- "Guide me" service: this feature protects the pilgrim from straying away from his residence in Makkah or Mina/Arafat through the use of maps and guidelines.
- "Al-Haram" guide: is an intelligent guide for all services and locations within the perimeter of Al-Kaabah, and includes; the gates of Al-Kaabah, elevators, escalators, places of ablutions, toilets, security-lockers, health centers, the Red Crescent, places of prayer for disabled people, etc.
- "Audio guide": a free voice service through which users listen to an explanation of all the rituals of Hajj in various languages provided by the Ministry of Islamic Affairs in Saudi Arabia.
- "Advise me": a free voice service through which users can inquire with Islamic scholars concerning Islamic rulings provided by the General Secretariat of the Islamic awareness in Saudi Arabia.

Only very limited support for content notifications was found in this application, and is only available in Arabic. In addition, the settings options are only provided for one notification feature.

"**Athan Pro Prayer Times**" [4] is a smart-phone application that provides prayer timings, Athan and a Qiblah-direction indicator. This application is used by thousands of users worldwide. It provides user-location prayer times and many other useful features such as the Holy Quran recitation mode (through a service provided by the Quran Pro Application [8] by the same developer), remembrance-supplications, and a calendar of Islamic holidays. Additionally, the application calculates the timings of prayers with good accuracy. It provides multiple calculation methods

for worldwide prayer-times that include: 1. Umm Al Qura University, 2. Muslim World League, 3. University of Islamic Sciences of Karachi, 4. Egyptian General Authority of Survey, 5. Islamic Union of North America. Some notification functionality was provided in terms of Athan alerts and messages that relate the Hadith of the day. The main drawback of this application is that many advertisements appear within every service, and removing the advertisements is only available with a service cost.

"**Muslim Pro – Prayer Times, Azan, Quran & Qibla**" [5] is a smart-phone application providing Athan, Qiblah-direction, prayer timings and Holy Quran recitation. This free application is used by more than 25 million users throughout the world. It uses an accurate timing calculation method for the prayer timings and Athan based on the user-location. Muslim Pro also contains the Holy Quran text with audio recitations and translations. Additionally, it provides a Qiblah direction indicator and an Islamic Hijri calendar. The map-service identifies locations of Halal restaurants, and mosques. Finally, the application and Quran-text are fully interpreted in the following languages: Bahasa Indonesia, Bahasa Melayu, Deutsch, English, Spanish, French, Italian, Hollands, Portugal, Turkish, Arabic, Urdu, Russian, Simplified Chinese, Japanese, and Thai Language. Notification features were evident in the form of prayer-time alerts and verse-of-the day alerts. On the other hand, the Qiblah direction indicator is inaccurate and not straightforward to use. Additionally, many advertisements appear within the application, and removing them is only available with a service cost. Additional voices for the audio Quran and the Athan are only available with a service cost.

"**AlSalam**" [6] is a free and comprehensive Islamic application that provides a library of Islamic content and rituals. AlSalam provides prayer times, Quran, Qiblah direction and Athan functionality. Some of the more notable features of this application includes: a categorized library of knowledge, user-location prayer-times, Qiblah direction, a calendar convertor, a search-engine and multiple-language Quran translations. In this application, only one notification feature was found in the case of prayer-time alarms. It was noted, however, that the Qiblah direction indicator was inaccurate and its use was not straightforward. Many advertisements appear in the application, and removing them is only available with a service cost. Finally, the settings options in this application are very limited.

"**Athan – Prayer times, Qibla and Mosques finder**" [7], is a smart-phone application featuring Athan, prayer-timings and Qiblah direction. This free application is used by over one million users according to the latest download statistics [7]. This well-designed application is available for Android and iOS devices. However, this application has many interrupting advertisements, and removing them is only available with a service cost. Notifications were only found for prayer alarms. The Qiblah direction indicator is inaccurate and its use is not straightforward. Finally, the application is a limited version and extra features, such as; complete Quran translations and Athan voices are only available with an extra cost.

"Quran Pro" [8], is a smart-phone application with an English-only interface that features; Quran-text with Othmanic font and 20 language translations, Arabic Tafseer, bookmarks, supplications, daily verse of the day notifications and Friday's reminder. However, the Athan function invokes another application provided by the same developer.

"**Golden Quran**" [9], provides an Othmanic font for the Holy Quran and provides audio recitations in an interactive way. The application interacts with the verse and provides many features about interpretations, Tajweed (rules governing pronunciation of recitation), Morphology, and other Quranic sciences. In addition, other features are provided that includes: prayer times, Qiblah direction indicator and a search engine for Quranic verses and chapters. The application provides advanced Quranic services and a settings option. Notifications are provided for invocations, reminder to read the Quran, and others, but some notifications do not work. Finally, the Qiblah direction indicator is inaccurate and not straightforward to use.

"**Muslim Mate**" [10], is a comprehensive Islamic mobile application that provides; prayer-times, a Qiblah direction indicator, a Quran search engine, multilingual phonetic transliterations, translations & recitations, and a locator for nearby places of interest. However, this application is only available for iOS devices. Finally, notifications are mainly provided for Athan/prayer times and calendar/events only. The application provides advanced setting options, however, the settings options are scattered in the application under each service. The application only displays the Quran verse-by-verse rather than page-by-page. Many adverts appear during the use of the application, with can only be removed with a service cost.

The next section explains the analysis used for comparing the other related efforts with this work. Section IV describes the proposed design and architecture.

III. Discussion and Analysis

Related studies focusing on mobile applications had proposed various feature-analysis criterion; and based on those criteria, had presented comparisons, evaluations and assessments concerning the use of such applications, as well proposing requirements guidelines in order to help users select between a large range of applications. For instance, the work in [11] was based on the user friendliness, completeness, and quality of use as the main assessment criteria. On the other hand, the work in [12] had considered the perceived usefulness, content quality, layout/graphic design, and application's ease of use as their main criteria.

The criterion used in this study consists of four categories and fourteen subcategories used in the requirements analysis when comparing between the related applications. Table 1 shows the categories and subcategories used and presents a summary of the main features of the most relevant smart-phone applications [1 – 10]. The criteria selection was largely tied to the nature of the application, its content and it main goals. The first category (*Content*) is related to the content, which is a very sensitive point when it comes to the Holy Quran. For instance, the Othmanic text font for the Quran is widely used for print back copies and could be a requirement of many users of smart-phone applications. Secondly, the complete Quran text in such applications is desirable since the

Quran is considered as the companion of its users in all times and places. According to Table 1, we note the absence of those two criteria in two of the ten selected applications as samples for the study, while [4] uses an external application [8] (from the same developer) to fulfill this requirement. Requirements 3 and 4 are considered as important in terms of access to information within the application; and in two of the ten selected applications there is no search feature for Quranic chapters that allow the user to navigate to the chapters and verses directly. It is noted that in five of the ten examined applications, there is no search feature for the Quranic chapter names or the Quranic verses, while in [10] this feature is available with a service cost. Hence, this offers researchers and developers an opportunity for further development in this area.

The second category of requirements (*Multimedia and Notifications*) is considered essential for educational purposes, and allows users of the application to improve recitation skills by listening to audio recitations of the Quranic verses. This criteria is only present in four out of ten of the selected applications. As one of the objectives of this study was to compare alerts/notifications with related applications, it was noted that notification features were found in some of those other applications to various extents and with their limitations. For instance, some applications had used voice and text for alerts, while others used text only as in [3 and 4]. Concerning the offline access to the text and audio content, it was found that five of ten applications had provided support for this feature.

In the third category of requirements (*Translation and Interpretation*), we find that many of the applications being examined had no support for providing the interpretation of the Quranic verses in any language, while in [10] this feature is available with a service cost. The interpretation of the Quran is considered necessary for users to understand the meanings and purposes of the verses. Additionally, some applications do not support multiple languages in the user-interface, which is considered as a major limitation since the proportion of non-Arab users is much higher than the proportion of Arab users.

The last category of requirements (*Implementation and Distinguished Features*) concerns the applications themselves in terms of usage costs, periodic updating, technical support, follow-up development updates and the availability of the application in different mobile operating systems. Most of the applications considered here had addressed those requirements [2 - 8, and 10].

A number of relevant studies were found that describe efforts related to the latest developments in Quran applications using mobile technology [13], whilst the quality-assurance factors and security concerns for such applications were highlighted in [14] and [15 - 16], respectively. Readers interested in solutions for dealing with sensitive digital image content found in many content-based applications are referred to the discussions made in [17 - 19].

IV. System Design and Architecture

The proposed system design and architecture comprises of two main parts:

1) The monitoring and display at the user-end, and;

2) The content-search and retrieval algorithm at the backend server.

The monitoring and display part at the user's side is where the system gathers local-time and current-location data, which is sent to the backend servers using online communications before displaying the retrieved content. Figure 1 illustrates the concept of monitoring and content-display at the user-end.

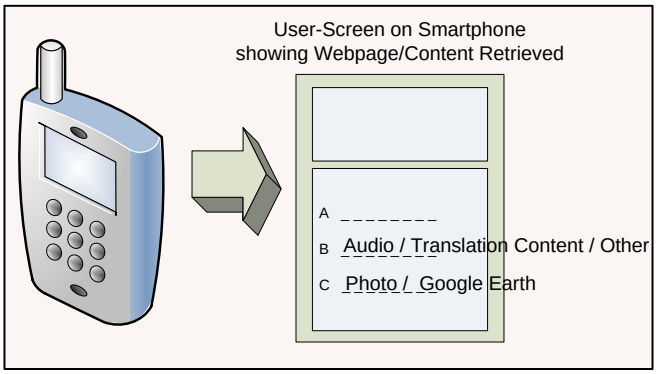


Fig. 1. Conceptualization of User-End Content Monitoring and Display

TABLE I. COMPARATIVE ANALYSIS OF RELEVANT CURRENT RELATED APPLICATIONS

Legends		[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]
Feature is Supported	✓										
Feature is Supported with a Limitation	L										
Feature is Available with Service Cost	C										
Feature is not Supported	×										
Category & Sub-category											
Content	Complete Quran	✓	✓	×	Provided through Quran Pro App.	✓	✓	×	✓	✓	✓
	Othmanic Text	✓	✓	×	Provided through Quran Pro App.	✓	✓	×	✓	✓	✓
	Chapter/Verse Indexing	✓	✓	×	Provided through Quran Pro App.	✓	✓	×	✓	✓	✓
	Search Functionality	✓	×	×	×	✓	✓	×	×	✓	C
Multimedia and Notifications	Audio Recitations	×	×	×	Provided through Quran Pro App.	C	✓	×	✓	✓	C
	Text and Audio Notifications	✓	✓	L	L	✓	✓	✓	✓	✓	✓
	Offline Accessibility	✓	×	×	L	✓	×	✓	✓	✓	L
Translations and Interpretations	Number of Supported Languages in the U.I.	01	06	01	11	16	02	02	01	02	15
	Quran Translation	×	×	×	×	36	36	×	20	15	01/ C
	Verse Interpretation	✓	×	×	×	×	✓	×	✓	✓	C
Implementation and Distinguished Features	Regular Updates/Bug Fixes	×	✓	✓	✓	✓	✓	✓	✓	✓	✓
	Free Download	✓	✓	✓	L	L	✓	L	L	✓	L
	Online Support	✓	✓	✓	✓	✓	✓	✓	✓	×	✓
	Mobile OS Platform	iOS & Android & BlackBerry 10	iOS & Android	iOS & Android	iOS & Android & Windows 10 Mobile	iOS & Android	iOS & Android	iOS & Android	iOS & Android & Windows 10 Mobile	iOS & Android	iOS Only

At the backend server side, the retrieved time and location data for each user is then processed by searching through a database of Quranic verses and Hadith, and retrieving the most appropriate verses and Hadith (if any) for that particular user time and location. Data retrieval and processing at the backend can be decomposed into three essential components, including: converting between Gregorian and Islamic calendars as per notification requirements, determining prayer-time calculations based on the user-location, and determining the user's immediate environment. In the first instance, calendar conversions are typically achieved in the implementation by employing integrated API functionality within the library functions. The user's immediate location and surroundings are obtained using the Google Map Web Service (GMWS) [20], while the location-based prayer-time calculations were determined using the GMWS data as input to the Islamic-Finder service for selecting the desired calculation algorithm provided through their web-service [21]. Finally, the retrieved data is transmitted to the user's device for display and/or playback. The proposed system framework also supports uploading Fatwas into the database by

authorized Islamic scholars in real-time, and consequent retrieval and display of Fatwas at the user-end (where applicable) along with the other content retrieved. The principle system architecture diagram, showing the main functional components at the user-end and backend server is illustrated in Figure 2.

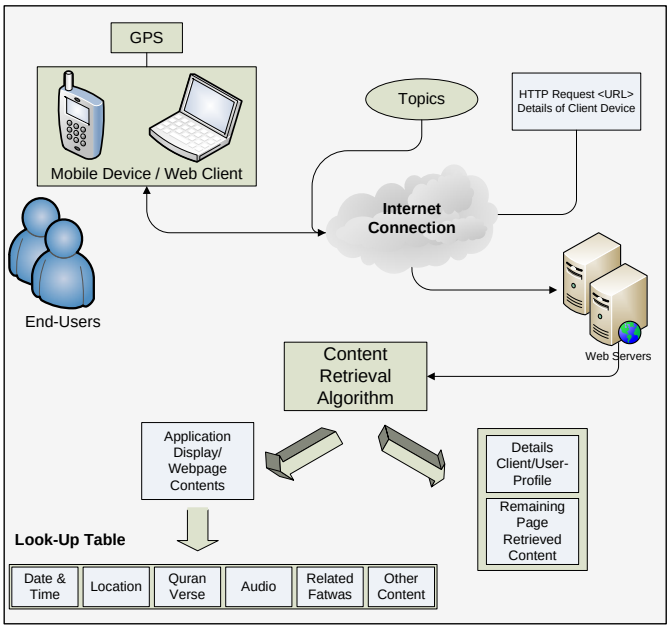


Fig. 2. System Architecture Diagram

V. Initial Results

This section presents some results that report on initial tests performed on the developed application. Table 2 presents the results related to two test cases. The first case relates to a time-based condition only (e.g. at Asr prayer time), while the second case relates to a time and location based condition (e.g. at Adhan times and at a transport station).

TABLE II. Sample Results of Notifications Retrieved

User-Events		Corresponding Notification Verse(s)
Time	**Place/ Location**	
Asr prayer Time	Anywhere (prayer times depends on user-location)	Maintain with care the [obligatory] prayers and [in particular] the middle prayer and stand before Allah, devoutly obedient [Al-Baqarah (The Cow), 283].
At Every Adhan	At any transportation station (train, bus, airport) OR while travelling	And when you travel throughout the land, there is no blame upon you for shortening the prayer, [especially] if you fear that those who disbelieve may disrupt [or attack] you. Indeed, the disbelievers are ever to you a clear enemy [An-Nisa' (The Women), 101].

VI. Conclusions

The location and time based content retrieval system presented in this work provided an effective and practical approach for users to learn about relevant content at any time, while on the move. In this case-study, time-based and location-based Quran verses and Hadiths related to user's daily lifestyles (including events and rituals) were presented in an interactive way using mobile technology. The main novelty in this work was found in the idea, its application and system framework for use with mobile technology, which was not found in the current literature or in existing mobile applications. The proposed framework executes a number of functional stages at the user-side and backend server-side, and could be easily adapted and employed for use in many other related m-Learning application domains. Finally, some initial results were presented to demonstrate the functionality of the proposed architecture using several examples as a proof-of-concept of the described design and architecture.

References

[1] "Fadhkroni". Available at http://moubadarah.bankalbilad.com/fathkrony. Last Accessed on 10th December 2016.

[2] "AlMosally". Available at https://www.facebook.com/almosaly. Last Accessed on 10th December 2016.

[3] "Mutawef". Available at http://www.mutawef.com. Last Accessed on 10th December 2016.

[4] "Athan Pro Prayer times". Available at http://islam.quanticapps.com. Last Accessed on 1st February 2017.

[5] "Muslim Pro – Prayer Times, Azan, Quran & Qibla", Available at http://www.muslimpro.com. Last Accessed on 10th December 2016.

[6] "AlSalam". Available at https://www.facebook.com/Rommanapps. Last Accessed on 1st February 2017.

[7] "Athan – Prayer times, Qibla and Mosques finder". Available at https://www.islamicfinder.org. Last Accessed on 1st February 2017.

[8] "Quran Pro". Available at http://islam.quanticapps.com. Last Accessed on 1st February 2017.

[9] "Golden Quran". Available at http://www.blogofapps.com/best-islamic-apps-for-iphone. Last Accessed on 1st February 2017.

[10] "Muslim Mate". Available at http://www.muslimmateapp.com. Last Accessed on 1st February 2017.

[11] K. Z. Zamli and al., "Feature Analysis of Android based Holy Quran Applications", 2013 Taibah University International Conference on Advances in Information Technology for the Holy Quran and Its Sciences.

[12] M. Alqahtani, A. Fayyoumi, "Mobile Application Development for Quran Verse Recognition and Interpretations", International Journal of Interactive Mobile Technologies, Vol. 9, No 1, 2015.

[13] M. Zakariah, M.K. Khan, O. Tayan, K. Saleh, "Digital Quran Computing: Review, Classification, and Trend Analysis", Arabian Journal for Science and Engineering · February 2017, DOI: 10.1007/s13369-017-2415-4.

[14] Hafidh Alsamarrai; Omar Tayan; Maysam Abbod; Yasser M. Alginahi, "Requirements Assessment for Organizations, Users, Software Developers, and Funders for the Propagation of the Holy Quran and Its Sciences", Taibah University International Conference on Advances in Information Technology for the Holy Quran and Its Sciences, Madinah, KSA, 22-25 Dec 2013.

[15] O.Tayan, "Concepts and Tools for Protecting Sensitive Data in the IT Industry: A Review of Trends, Challenges and Mechanisms for Data-Protection", International Journal of Advanced Computer Science and Applications 8(2) · March 2017.

[16] O. Tayan, Y. Alginahi, "A review of recent advances on multimedia watermarking security and design implications for digital Quran computing", International Symposium on Biometrics and Security Technologies (ISBAST), 2014, At Kualal Lumpur, Malaysia, DOI: 10.1109/ISBAST.2014.7013139.

[17] L. Laouamer, O. Tayan, "An Efficient and Robust Hybrid Watermarking Scheme for Text-Images", International Journal of Network Security 18(6):1152-1158 · November 2016.

[18] L. Laouamer, O. Tayan, "A Semi-Blind Robust DCT Watermarking Approach for Sensitive Text Images", Arabian Journal for Science and Engineering 40(4):1097-1109 · April 2015, DOI: 10.1007/s13369-015-1596-y.

[19] L. Laouamer, O. Tayan, "An enhanced SVD technique for authentication and protection of text-images using a case study on digital Quran content with sensitivity constraints", Life Science Journal 2013;10(2).

[20] Google Places Web-API [Online] Available at https://developers.google.com/places/web-service/supported_types. Last Accessed on 10th December 2016.

[21] Islamic-Finder Prayer Times Web-Service [Online] Available at https://www.islamicfinder.org/world. Last Accessed on 10th December 2016.

Computation of QoS While Composing Web Services

Khozema Ali Shabbar
Research Scholar
School of Engineering & Technology
Career Point University
Kota, India

Dr. Tarun Shrimali
Principal
Faculty of Engineering
Sunrise Group of Institutions
Udaipur, India

Dr. Mohemmed Sha
Assistant Professor
Dept. of Computer Science
Prince Sattam bin Abdulaziz University
Al Kharj, KSA

***Abstract*—Composition of web services has emerged as a fast growing field of research since an atomic service in its entirety is not capable to perform a specific task. Composition of web services is a process where a set of web services, heterogeneous in nature, are clubbed together in order to perform a specific task. Individually, Component web services may be performing well as far as Quality of Service (QoS) is concerned but the core issue is that while composing, do they satisfy Users requirements in terms of QoS? Computation of QoS while composing web services appears to be a big challenge. A lot of research work in this regard, has already been undertaken to come out with new, innovative and credible solutions for the same.**

This Paper presents a thorough review-study of different frameworks, architectures, methodologies and algorithms suggested by different researchers in their efforts to compute the overall QoS while composing web services. Moreover, Effectiveness of different methods in terms of QoS while composing is also presented.

***Keywords*—Web Services; Web Services Selection; Web Services Composition; Composite QoS (CQoS); Quality of Service (QoS)**

I. Introduction

Web services are getting more popularity as far as distributed computing and e-business/ commerce are concerned. Since Web services are loosely coupled, allowing developers to construct, produce and compose them even at execution time.

The concept of Web services composition has emerged as an effective method for integration of business related applications where numerous features of web services are combined together in order to satisfy complex requirements that an individual web service could not do. In other words, Composition is the process where prevailing composite web services or atomic web Services are clubbed together to perform a specific business operation. This composition/ integration can be done manually or automatically. Since area of scope of this paper is confined to automatic/ dynamic composition of web services, we are presenting here various processes involved in composition, clearly depicting them in the following diagram (Fig 1).

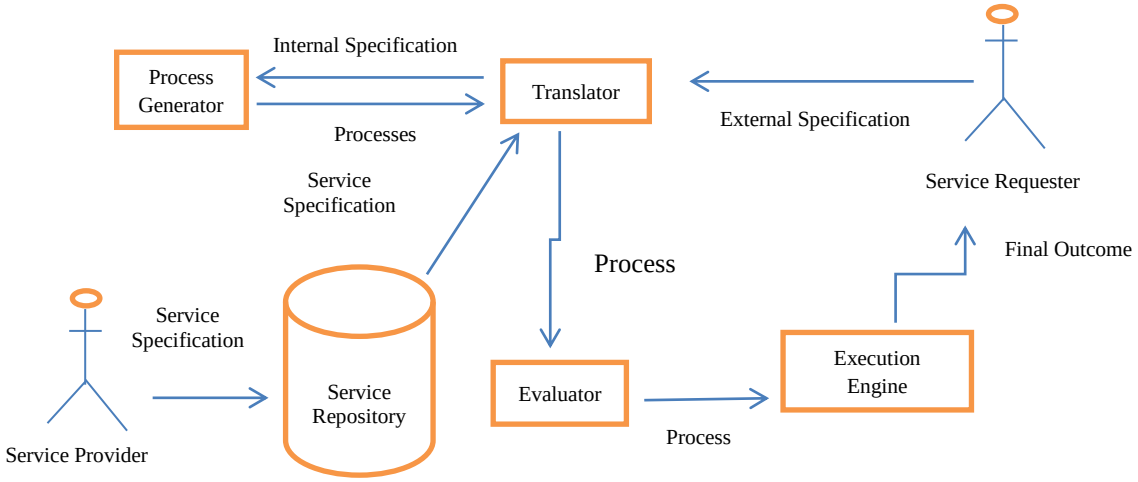


Fig. 1. A Generalized Composition Mechanism

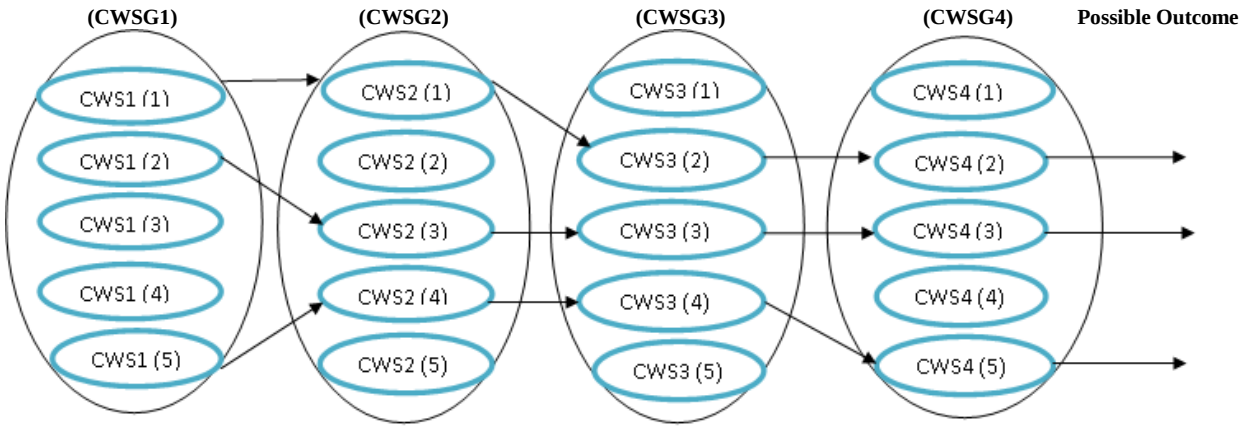


Fig. 2. Three Possible Outcomes of Composite web services

Services published by the service providers lie there in the service repository. Service requester initiates the process by making a service request specifying functional as well as non-functional constraints. This request first goes to the Translator whose job is to translate the request from its external form into a standard form acceptable to the system. Services meeting user's requirements are selected from the repository. Now Process Generator composes theses services. As shown in the (Fig 2) above, when there exists more than one possible outcome of Composite Services, all of them meeting user's criteria then it is the job of the Evaluator to evaluate all those composite services and returns the best one.

Here CWSG1, CWSG2…. Stand for Candidate Web Services Group1, Candidate Web Services Group2, and so on. Similarly CWS1, CWS2 …stand for Candidate Web Service1, Candidate Web Service2, and so on. Candidate services are group based on their functional similarities. Each group is having candidate web services with similar functionality attributes but differing in their QoS. For a particular request let say there are three possible outcomes of composite web services. Then which one is to be selected by the evaluator solely depends upon the QoS of composite web service as a whole and constituent web services individually.

In this paper, we are studying various approaches and techniques adopted to perform composition of Web Services based on QoS, computing the overall QoS of the Composite web services called CQoS. The paper is divided into different sections. Section 2- Related Work. Section 3- QoS based Web Service Composition Methodologies – an overview. Section 4- Comparison of Effectiveness of Composition Methods. Section 5- Conclusion.

II. Related Works

Composition of web services can be performed manually or automatically. In manual approach, each of the web services gets executed one after another in an ordered way to attain the specified goal, seems to be a complicated method. It requires a lot of time and efforts. With the high number of web services offered over the Internet, automatic composition appears to be more feasible. Automatic composition can be performed using any of the three methods - static, semi-static or dynamic. Static composition is done at design time and requesters have to form an abstract process model prior to actual composition.

Dynamic composition is considered to be a complicated one as is done at run time upon the user's request. But the main advantage lies in discovering and invoking of web services dynamically on request. Some of the examples of dynamic composition are Negotiation, Semantic Web, intelligence algorithm etc. etc. Various compositions related Research works are discussed here under.

Fatma Siala et al. [1] proposed a Multi-Agents architecture to discover the optimized Composite QOS (CQoS). Composite Web services were selected having negotiation with Multiple Agents. For web service composition, user's preferences were taken into account first then agents were used to negotiate the QoS value and finally, Service Providers, providing different services, were dynamically selected in the composition. The main focus was lying in negotiating with only those Web services providers who are available, resulting in an improved CPU time. Experimental results have even proved it.

Lou Yuan-sheng et al [2] designed a 'QoS and workflow' based framework for Web service selection & dynamic composition and executed a sample system (prototype). It used visual interface to perform customization of Web service composition process. Also, it implemented global optimization algorithms to attain the dynamic selection of the suitable service, meeting composite web service QoS and the existing services QoS.

Dong Rang-sheng et al [3] presented a QoS mechanism to dynamically compose the Web Services addressing the problem of selection & coordination. They also developed a WS_TSC algorithm for the selection of services (considering the constraints like – success rate, composition rate and response time). Lastly, they carried out simulation experiments to measure the performance of dynamic composition of Web services and found better results w.r.t. success rate, composition rate and response time.

Lu Li et al [4] proposed a web service composition selection model built on the concept of Multi-dimension QoS. This concept was introduced to express the QoS attributes of web service composition. Also, it computed the QoS of web service composition depending upon the nature of web service

composition. Finally, Web service composition that optimally satisfies the non-functional requirements (besides functional requirements) was chosen among the web service compositions.

Farhan Hassan Khan et al [5] proposed an innovative technique for auto-dynamic composition of web services combining both of the techniques based on interface and functionality. They focused on issues related to QoS, data distribution and execution problems etc. To resolve the problems of decentralized dataflow, they have come up with a framework which resulted in highest throughput, better response time, and minimal latency. Also, they have presented a solution for the difficulties faced during the composition of web services because of the continuous variations in parametric values (both input/output), independent nature and network related issues of the web services.

Ming-Wei Zhang et al [6] proposed an entirely different technique for the composition of web services on the basis of 'Production QoS rule' and adopted 'black box' technique of analysis to optimize composite services. Execution information pertaining to composite service is first stored and then used as a base for the ensuing statistical analysis and QoS knowledge mining. Web services periodic QoS values are computed and the production QoS rules are mined. (Basically, these rules are applied in order to specify various performances of 'Web service QoS' that take place in diversified environments). Finally, resultant QoS knowledge about Web services is used to discover optimized composite service.

Zhi Zhong Liu et al [7] introduced a reliable "Web Service Composition" method on the basis of the decomposition of global QoS attributes and dynamic prediction of QoS. Proposed method has two critical stages: (1) prior to the composition of web services, decomposition of global QoS attributes into local attributes takes place and the issue related to the dynamic composition of Web services becomes a localised optimization issue. (2) While execution, predicted QoS values are considered to be the base for the selection of the best Web service for the present abstract service.

Rajesh Karunamurthy et al [8] presented an innovative composition mechanism by extending the existing business model of Web service to explicitly carry out composition of Web services. Proposed method supported four characteristics of Web services viz - functional, non-functional, behavioral & semantic enabling identification, selection and clubbing of different Web services being part of composition procedure. The proposed mechanism comprises different components namely – description framework, composition framework and business model.

Sabrina Mehdi et al [9] introduced a model of web service composition built on auto multi-agent based planning with high availability of web services. The said model integrates communities based substitution process, swapping an unsuccessful web service (service agent) with new one. The new one adheres to the community, presenting functionality similar to that of failure one in order to guarantee its availability.

Olfa Hammas et al [10] proposed composition architecture incorporating two aspects: 1) Dynamic Selection i.e. binding of constituent services at runtime and 2) Adaptive Composition i.e. presuming to have updated/ latest knowledge about the status of constituent services at runtime, ensuring global QoS Optimization as well as taking care of failed services by replacing them. An algorithm for the QoS aware service selection based on Ant Colony Optimization was also proposed.

Freddy L´ecu´ et al [11] introduced a framework to perform service composition dynamically by performing semantic matchmaking between outputs and inputs service parameters enabling interaction and interconnection. This semantic matchmaking paves the way in discovering semantic compatibilities between service-descriptions defined autonomously. Furthermore, they introduced an algorithm for composition which follows a semantic graph based approach, where a graph denotes service compositions and nodes denote semantic relations among services. Additionally, both non-functional and functional attributes of services are taken into account to compute the most suitable and relevant service composition.

Wang Denghui et al [12] presented a novel recommendation method for the composition of web services calculating QoS credibility of every web service unit. Authors used user's comment/ weight of user's preference to calculate the reliability of different dimensions of QoS.

Pooya Shahrokh et al [13] have proposed a semi-heuristic genetic algorithm (a combination of both a heuristic method and the genetic algorithm). This heuristic method changes chromosomes based on unsatisfied constraints. As per authors, Research outcomes have endorsed the fact that the proposed method satisfies user's requirements more efficiently than other methods.

Namrata Kashyap et al [14] introduced 'QoS based service composition' model incorporating a Membership function. The said function does prioritize functionally similar candidate web services to a further higher level based on response time in order to be included in the composition, enhancing user's satisfaction.

Wei Zhang et al [15] presented a QoS based technique for composition of web services dynamically and used Ant Colony Optimization (ACO) algorithm for optimization. An updated version of ACO algorithm was presented to resolve the multi objective optimization problem aiming to find better settlements between multiple objectives.

Alexandre Sawczuk da Silva et al [16] presented a graph based PSO approach for service composition and selection which ascertains an optimum workflow & near-optimum Web services to be clubbed together in QoS based composition. Authors have addressed successfully various limitations of prevailing PSO-based techniques as the approach adopted here neither requires any selection of an initial configuration nor depends upon the users, having domain expertise. Experimental results of proposed graph based PSO technique were found to be better than greedy based PSO technique.

III. QoS based Web Service Composition Methodologies – An Overview

There exists many ways to combine web services. Following is a simple diagram (Fig 3) showcasing different methodologies for the composition of Web services.

A. Static Composition

Static composition is performed in design-phase where the design of the software system and the architecture are planned. The components to be used are selected, connected together, compiled and deployed finally. It may work well as far as the web service environment – service components and business partners change rarely. Examples of static composition engines include Bea WebLogic and Microsoft Biztalk etc. It requires lot of time & efforts.

Static Composition approaches are almost obsolete ones. Rapid growth of Web services and their well-defined interfaces prompted researchers towards auto-dynamic approaches for the Composition of Web service which are found to be more practical & feasible. In the following section it has been discussed in detail.

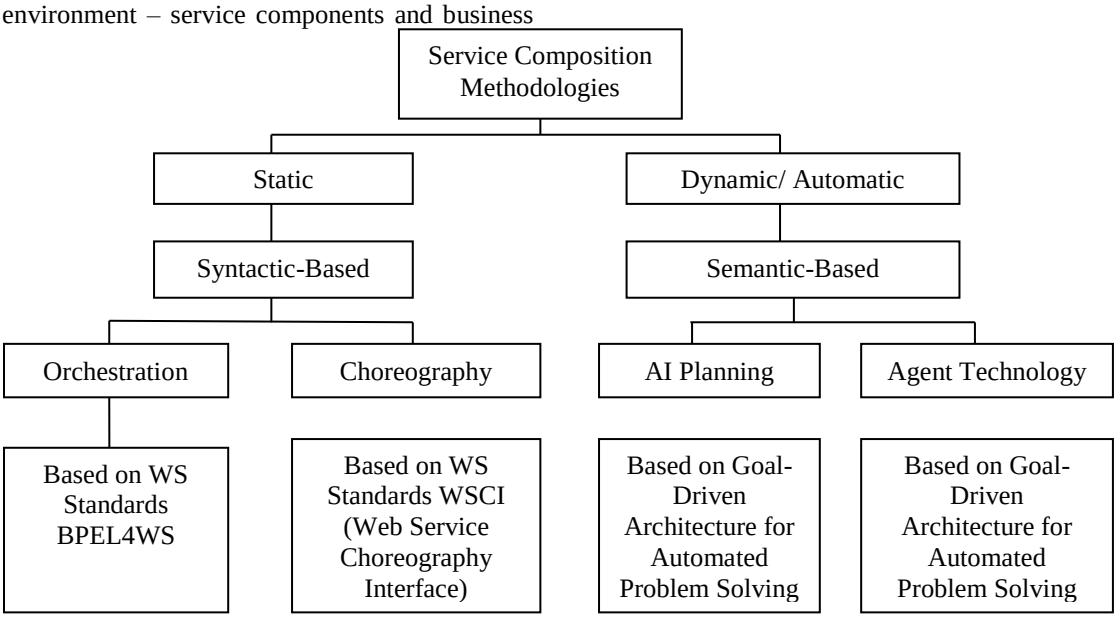


Fig. 3. Service Composition Methodologies

B. Dynamic/ Automated Composition

Area of automatic/ dynamic composition of web services is getting more attention as researchers are seeking innovative ways to perform compositions more proficiently finding better quality of results. The job of performing automatic Composition of web services is complicated one as it searches for candidate web services from a huge web service repository and handles service descriptions, not following a common standard. Most of the auto-dynamic compositions are based on workflow and/or Artificial Intelligence techniques.

1) Based on Multi-agent

In [1] [9] authors are incorporating multi- agents whose role is to talk about the QoS values and make selection of service providers dynamically for different services in the composition to discover the optimal Composite QoS. Both are focusing to ensure high availability of the required web services - [1] by talking to available providers (of Web services) only, resulting in improved CPU time and [9] incorporating a process of substitution, swapping an unsuccessful web service with a new one.

2) Based on Workflow

In [2] authors have proposed a framework for the composition of web services built on work flow incorporating dual features - dynamic composition & QoS scalability. This 'QoS scalability' facilitates adding/modifying existing QoS attributes (or say an extension to user defined properties) which can deliver distinct services to satisfy some distinguished QoS constraints. For QoS optimization, a hybrid version of algorithm depending on ant colony algorithm (CA) and genetic algorithm (GA) was also presented.

In [6] authors have proposed an innovative composition mechanism for web services on the basis of 'Production QoS rules'. The novelty of the mechanism lies in taking into account the relationships of Web service QoS to environments, which is generally neglected. First, execution repository is constructed by recording composite service performance information in a tabular form. Web services QoS point dataset is extracted and timely QoS values are computed. Then the production QoS rules are mined. Finally, estimated TQoS (True QoS) and Web services production QoS rules are utilized to discover optimized composite service.

In [3] authors have proposed a QoS model for dynamically composing Web services addressing problems of service components selection and services composition coordination. Authors have developed a WS _ TSC algorithm for the selection of services based upon the constraints - success rate, composition rate and response time. At the end, they performed simulation experiments in order to measure the performance of dynamic composition of Web services and

found better results w.r.t. success rate, composition rate and response time.

3) Based on Multi-Dimension QoS

In [4] authors presented a web service composition selection mechanism built on Multi-dimension (time, spatial, reliable and cost) QoS. Proposed concept of Multi-dimension describes the QoS attributes of web service composition. It then, computed the average QoS of each dimension of the constituent web service to measure the QoS of each dimension of composite web service. At the end, they applied improved Euclidean Distance algorithm to compute the QoS of composite web services in order to find the best composite web service, meeting User's requirements.

In [12] authors presented a reliable QoS-aware recommendation method for the composition of web services wherein at first, all global QoS dimensions of each of service units were added to find the global QoS value of composite service and computed the credibility global QoS dimensions of composite service w.r.t. advertised QoS value and execution result. Next, they selected the lowest QoS value in all of the service units as the local QoS dimension. The credibility local QoS dimensions was computed based on the reliable user's experience. User's preference weight was attached to the credible QoS information in order to compute evaluation result and the service with the highest value was finally chosen.

4) Based on Improved Genetic Algorithm (IGA)

In [13] authors have proposed a semi-heuristic genetic algorithm, a combination of both a heuristic method and the genetic algorithm. This heuristic method changes chromosomes based on unsatisfied constraints. Authors have added several heuristic methods to control the randomness of GA in such a way that mutation is carried out based on the quality parameters of the issue so that the algorithm does not move away from the optimal space of the optimal combination without being trapped by the local optimality. The dynamic weighting was adopted too to determine the required quality parameters.

5) Based on Fuzzy Logic

In [14] authors proposed Web-service composition System introducing a new formula (member function) that does prioritize the atomic/aspirant web services based on their response times to be included in the composition using fuzzy logic. The main idea here is that a composite service is assumed to be a fuzzy set where best candidate services having highest priority are included and priority is determined based on the response time.

6) Based on Ant Colony Optimization

In [15] authors have proposed a technique of performing decomposition of composite services into parallel execution paths, having a general flow structure. The problem of dynamic composition of web services for each execution path was considered to be a multi-objective optimization problem and hence presented MO¬_ACO (Multi-Objective Ant Colony Optimization) algorithm to handle it. MO¬_ACO is an improved version of the existing Ant Colony Optimization (ACO) algorithm. Experimental outcomes confirmed that proposed new algorithm was able to discover near-optimum results for multi-objective problems in a very efficient way as well as was scalable to perform composition of web services, highly complex in nature.

In [10] authors have proposed dynamic composition architecture for web services incorporating mechanism for adaptive composition, ensuring global QoS optimization. The problem of QoS based service selection was mapped to the multi-dimensional multi-choice knapsack problem, to find an atomic service from each service class to be put in the knapsack to build the composite service, ensuring that the aggregated QoS values should satisfy users QoS constraints.

7) Based on Intelligent Algorithm

In [7] authors introduced an innovative mechanism for QoS based composition of web services comprising two important steps. At first, global QoS attributes are decomposed into local attributes in order to find the combination of optimum local QoS attributes for each of the service classes and a novel CGA optimization algorithm was constructed combining Genetic Algorithm with that of Culture Algorithm. Secondly, they have designed QoS prediction method on the basis of improved case-based reasoning to predict the candidate services QoS prior to the selection a specific Web service for a particular service.

8) Based on Graph

In [11] authors have introduced a model for the composition of web services on the basis of functional attributes wherein services are bound depending on their functional description - Input, Output, Preconditions and Effects (IOPEs). The proposed model used the Causal Link Matrix (CLM) formalism to enable the final service composition computation in the form of semantic graph. Semantic connections between constituent web services are represented by the nodes of the semantic graph. At last, selection of the set of valid service compositions is made depending upon the non-functional attributes of the constituent web services. If a service composition fails to meet the non-functional attributes of the requested service, it is ignored.

In [16] authors have presented a PSO method based on graph for the QoS-aware composition of Web service. Instead of preselecting an abstract workflow for optimization, proposed method relied on the creation of a master graph of candidate services. They presented a greedy-based PSO method that too preselects an abstract workflow. A Set of experiments were conducted on both of the models to compare their efficiency and results of the experiments showed that the graph-based method was proved to be more efficient in producing solutions whose fitness values couldn't be matched by the solutions found from the greedy-based model.

9) Based on Decentralized Dataflow

In [5] authors have introduced an algorithm for dynamic composition of web services to resolve composition problems relating to availability, reliability, QoS and data distribution. In order to ensure data availability and make the system more reliable, authors have proposed a mechanism, introducing numerous repositories and Web Service Databases. In other words, data availability and retrieval of up to date information have been guaranteed by means of multiple registries and

aging factor respectively. The proposed system based on Quality of service, is a fault tolerant and reliable one, performing fast data retrieval.

10)Based on Semantic Web

In [8] authors have come up with a novel architecture by extending the existing business model of Web service to explicitly carry out composition of Web services. Proposed architecture supports four characteristics of Web services - non-functional, functional, semantic and behavioral enabling the identification, selection, and clubbing together of various component Web services involved in the composition.

There are two important aspects of the proposed method that distinguishes it from rest of the methods - 1) standard business model of Web service has been extended to carry out composition of existing services in order to create unavailable services that consumers search for and 2) described and composed Web services integrating all four characteristics altogether in a common semantic domain (other existing description techniques do not allow all of them altogether).

IV. Comparison of Effectiveness of Composition Methods

In this section, we will be presenting a comprehensive analysis of the various methods employed to perform Web Service Composition (WSC) based on QoS. All the methods discussed in this paper are Semantic based and automatic/ dynamic in nature. Below find here are the important factors solely based on the related research works. Each of the related works included here, showcased the effectiveness of the technique/ algorithm employed for the composition of web services.

TABLE I. Comparison of Methodologies

Related Article	**Approach/ Method**	**Technique/ Algorithm**	**Parameters**	**Effectiveness**
Fatma Siala et al. [1]	Based on Multi-Agent Negotiation	Multi-Agents Negotiation to find out the best CQoS. Used – 'Extended FIPA Protocol' and 'Contract Net Protocol'	Speed, Accuracy Availability, Price, Reliability, Reputation	Experimental results of the Proposed Multi-Agents Model demonstrated better results in terms of Execution Time (better CPU Time) compared to other existing Models. It can be applied to other distributed Computing Paradigms too
Lou Yuan-sheng et al [2]	Based on Work-flow	A hybrid version of $Algo^{m}$ (Ant Colony Optimization + Genetic $Algo^{m}$) for QoS OptimizationFlowchart (established according to User's QoS requirements) is taken as XML documents to compute Score	Reliability	Proposed Service Composition framework based on QoS supported dynamic Web service selection and the visualization modeling
Dong Rang-sheng et al [3]	Based on Work-flow	Objective function F as Performance Metric Criteria. Developed Algorithm, WS_TSC for the Selection of Services	Rate of Composition, Rate of Success and Response Time	Proposed architecture was found performing better w.r.t. the rate of Composition, rate of the Success and Response time i.e. higher rate of Composition & Success and Less Response time
Lu Li et al [4]	Based on Multi-Dimension QoS	Improved Euclidean Distance Algorithm	Four Dimensions – 1. Time: (Execution & Communication Time) 2. Spatial: (Message length & Storage Capacity) 3. Reliable: (Reliability & Availability) 4.Cost: Service Cost	Proposed Model presented Multi-Dimension QoS which can more exactly describe the QoS Attributes of Composite Web Service as well as help conduct a better study of the QoS of Composite Web Service.

Farhan Hassan Khan et al [5]	Based on Decentralized Dataflow	Auto-Dynamic Composition Mechanism by combining Approaches based on Interface and Functionality.	Reliability, Availability, Latency, Throughput, Response Time	Proposed Mechanism is Reliable, Fault Tolerant, Efficient in Data Retrieval and based on QoS. Proposed Algorithm for Dynamic Web Services Composition resolves the composition issues related to Availability, Reliability, QoS and Data Distribution
Ming-Wei Zhang et al [6]	Based on Work-flow	Black-Box technique of Analysis to Optimize Composite Services and Mining Algorithm - QoSR based on Production QoS Rule	Cost, Response Time, Robustness, Security	Proposed Service Composition Approach based on the production QoS rule does – • Improve the System Efficiency • Increase the System Stability and • Reduce the possibility of service system abnormality in Dynamic Service Execution Environment
Zhi Zhong Liu et al [7]	Based on Intelligent Algorithm	Novel CGA Optimization Algorithm (Integrating GA into the framework of CA) and Improved Cased-Based Reasoning (CBR)	Cost, Response Time, Availability, Reliability	Proposed Approach not only reduces the complexity but also increases the flexibility and the efficiency of Service Composition besides greatly enhancing the credibility of Composite Web service
Rajesh Karunamurthy et al [8]	Based on Semantic Web	Pellet Reasoner and Isabelle Theorem Prover	Non-functional, Functional, Semantic and Behavioral characteristics	Proposed Architecture executes faster Composition of Web Services resulting in higher Response Time and taking into account all of the four characteristics altogether – Non-functional, Functional, Semantic and Behavioral
Sabrina Mehdi et al [9]	Based on Multi-Agent	Multi-Agent Automatic Planning Architecture (using Java/ JADE), Contract Net Protocol	Availability Cost, Response Time, Reliability	Proposed Model ensures higher availability of web services and treats the failure of web service or fault tolerance (Auto- Substitution mechanism based on Communities)
Olfa Hammas et al [10]	Based on Ant Colony Optimization	Ant Knapsack Algorithm	Availability, Reliability, Response Time, Execution Price, Latency, Throughput,	Proposed Composition Model incorporated two Concepts – 1. Dynamic Selection (binding of services at runtime) and 2. Adaptive Composition (assuming to have updated knowledge about constituent web services status)
Freddy L´ecu´ et al [11]	Based on Semantic Graph	Causal Link Matrix (CLM) formalism, Algorithms for the Composition of Web Services & Ranking of Composition Results	Cost, Response Time	Proposed general Model for the Composition is quite suitable for Web Services which are described using WSMO (capability model), SA-WSDL, or OWL-S (service profile) specification
Wang Denghui et al [12]	Based on Multi Dimension QoS	QoS Aware Web Service Composition Recommendation Method	Global QoS Dimensions are - Price, Execution Time, Local QoS Dimensions are - Successful Rate, Availability, Reliability, Security	Proposed Approach considered the user's experience to describe all QoS dimensions and calculated the credibility of end user's experience to ensure the objectivity of user commend

Pooya Shahrokh et al [13]	Based on Improved Genetic Algorithm (IGA)	Semi-heuristic Genetic Algorithm (combination of both heuristic method and the genetic algorithm),	Availability, Execution cost, Response Time, Successful Execution Rate, Reputation	Proposed Method can be applied to discover a Composition technique which satisfies user's requirements more efficiently than other methods. Applying heuristic genetic algorithm resulted in improved Execution Time
Namrata Kashyap et al [14]	Based on Fuzzy Logic	Fuzzy logic & Membership functions	Reputation, Availability, Execution Price, Successful Execution rate, Execution Duration	Performance of the new improved Afive formula was found to be better than the existing Atri formula as it successfully prioritized Web Services at more precise and finer level on the basis of their response times Atri and Afive stand for the tri-value and the five-value QOS attribute of A of the Web
Wei Zhang et al [15]	Based on Ant Colony Optimization	Multi Objective Optimum Path Selection Technique and MO_ACO Algorithm (an improved Version of existing Ant Colony Optimization Algorithm)	Response Time Cost, Availability, Reliability	Experimental Outcomes demonstrated that Proposed new Algorithm (MO_ACO) was able to discover near-optimum results for Multi Objective Problems in a very efficient way as well as was scalable to perform Composition of web services, very complex in nature
Alexandre Sawczuk da Silva et al [16]	Based on Graph-based Particle Swarm Optimization - PSO	Graph-based Particle Swarm Optimization (PSO) Technique and Service Composition Algorithm for Particle as well as Fitness Function	Availability, Response Time Execution Cost, Reliability	Graph-based technique produced far better solutions surpassing to the solutions achieved from the Greedy-based technique.

After thoroughly reviewing the effectiveness of the above listed QoS based Composition mechanisms, techniques employed by Lu Li et al [4], Farhan Hassan Khan et al [5] and Wei Zhang et al [15] were found to be more practical & effective in terms of quality. Techniques employed by Pooya Shahrokh et al [13] and Rajesh Karunamurthy et al [8] were found to satisfy user's requirements (functional aspects) more efficiently than other methods.

V. CONCLUSIONS

The above mentioned composition methods give us a better insight while composing web services to reach the functionality of a specific task. To accomplish a specific task where an atomic web service is not able to do it lonely, composition of web services appears to be the best solution. Component web services are those best web services selected based on the individual performance in terms of quality (QoS). While composing, computation of overall QoS still remains a big question to be answered to. Composite QoS (CQoS) plays vital role in finding the best Composite Web Service. Though, Researchers have done a great job & have come up with innovative techniques to compute Composite QoS more accurately in order to retrieve the most relevant & desirable composite web service but still, they are more or less depending on either Providers published information or users provided feedback. As discussed earlier, Providers published information may not always be true, credible & up-to-date. Similarly users provided feedback may also be biased sometimes. Hence we reach to the conclusion that an independent & consistent QoS evaluation system is still needed to measure more accurate value of QoS. Researchers have to focus their research on developing such a novel system which can work independently taking into consideration more number of non-functional attributes. As a future scope, based on the various frameworks reviewed here in this paper, an innovative framework will be proposed incorporating more number of non-functional attributes, so that it can compose web services more efficiently in terms of QoS.

REFERENCES

[1] Fatma Siala, Khaled Ghedira, "A Multi-Agent Selection of Web Service Providers Driven by Composite QoS",IEEE, pp. 55-60, 2011.

[2] Lou Yuan-sheng, Tao Zhen-hong, Yue Lu-lu, Xu Hong-tao, Xi Zhi-hong,Wu Zhi-feng, "A QoS-based Web Service Dynamic Composition Framework", IEEE, Ninth International Symposium on Distributed Computing and Applications to Business, Engineering and Science, pp. 188-192, 2010.

[3] DONG Rang-sheng, WANG Fei-ming, LUO Xiang-yu, "Dynamic Web Services Composition Based on QoS Model", IEEE, pp. 823-826, 2010.

[4] Lu Li, Mei Rong, Guangquan Zhang, "A Web Service Composition Selection Approach based on Multi-Dimension QoS", IEEE, The 8th International Conference on Computer Science & Education (ICCSE 2013) Colombo, Sri Lanka, pp. 1463-1468, April, 2013.

[5] Farhan Hassan Khan, M.Younus Javed, Saba Bashir, Aihab Khan, Malik Sikandar Hayat Khiyal, "QoS Based Dynamic Web Services Composition & Execution", International Journal of Computer Science and Information Security, vol. 7, no. 2, pp. 147-152, February, 2010.

[6] Ming-Wei Zhang, Bin Zhang, Ying Liu, Jun Na, Zhi-Liang Zhu, "Web Service Composition Based on QoS Rules", JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY, vol. 25, no. 6, pp. 1143–1156, November, 2010.

[7] Zhi Zhong Liu, Zong Pu Jia, Xiao Xue, Ji Yu An, "Reliable Web service composition based on QoS dynamic prediction", Springer, pp. 1409-1425, July, 2014.

[8] Rajesh Karunamurthy n, FerhatKhendek,RochH.Glitho, "A novel architecture for Web service composition", Journal of Network and Computer Applications, Elsevier, vol. 35, no. 2012, pp. 787–802, November, 2011.

[9] Sabrina Mehdi, Nacer eddine Zarour, "Composition of web services using multi agent based planning with high availability of web services", IEEE, 2nd International Conference on Advanced Technologies for Signal and Image Processing - ATSIP'2016 Monastir, Tunisia, pp. 10-15, March, 2016.

[10] Olfa Hammas, Saloua Ben Yahia, Samir Ben Ahmed, "Adaptive Web Service Composition Insuring Global QoS Optimization", IEEE, 2015.

[11] Freddy L´ecu´e, Eduardo Silva, and Lu´ıs Ferreira Pires, "A Framework for Dynamic Web Services Composition", Emerging Web Services Technology, vol. 2, 2008.

[12] Wang Denghui, Huang Hao, Xie Changsheng, "A Novel Web Service Composition Recommendation Approach Based on Reliable QoS", IEEE, Eighth International Conference on Networking, Architecture and Storage, pp. 321-325, 2013.

[13] Pooya Shahrokh, Faramarz Safi-Esfahani, "QoS-based Web Service Composition Applying an Improved Genetic Algorithm (IGA) Method", International Journal of Enterprise Information Systems, vol. 12, no. 3, pp. 60-77, July-September 2016

[14] Namrata Kashyap, Kirti Tyagi, "Dynamic Composition of Web Services Based on Qos Parameters Using Fuzzy Logic", IEEE, International Conference on Advances in Computer Engineering and Applications (ICACEA), IMS Engineering College, Ghaziabad, India, pp. 778-782, 2015.

[15] Wei Zhang, Carl K. Chang, Taiming Feng, Hsin-yi Jiang, "QoS-based Dynamic Web Service Composition with Ant Colony Optimization", IEEE, 34th Annual Computer Software and Applications Conference, pp. 493-502, 2010.

[16] Alexandre Sawczuk da Silva, Hui Ma, Mengjie Zhang, "A Graph-Based Particle Swarm Optimisation Approach to QoS-Aware Web Service Composition and Selection", IEEE, Congress on Evolutionary Computation (CEC), Beijing, China, pp. 3127-3134, July, 2014.

Emotion Classification in Arousal Valence Model using MAHNOB-HCI Database

Mimoun Ben Henia Wiem
Université de Tunis El Manar,
Ecole Nationale d'Ingénieurs de Tunis,
LR-11-ES17, Signal, Images et Technologies de l'Information (LR-SITI-ENIT)
BP. 37 Belvédère, 1002, Tunis, Tunisie

Zied Lachiri
Université de Tunis El Manar,
Ecole Nationale d'Ingénieurs de Tunis,
LR-11-ES17, Signal, Images et Technologies de l'Information (LR-SITI-ENIT)
BP. 37 Belvédère, 1002, Tunis, Tunisie

***Abstract*—Emotion recognition from physiological signals attracted the attention of researchers from different disciplines, such as affective computing, cognitive science and psychology. This paper aims to classify emotional statements using peripheral physiological signals based on arousal-valence evaluation. These signals are the Electrocardiogram, Respiration Volume, Skin Temperature and Galvanic Skin Response. We explored the signals collected in the MAHNOB-HCI multimodal tagging database. We defined the emotion into three different ways: two and three classes using 1-9 discrete self-rating scales and another model using 9 emotional keywords to establish the three defined areas in arousal-valence dimensions. To perform the accuracies, we began by removing the artefacts and noise from the signals, and then we extracted 169 features. We finished by classifying the emotional states using the support vector machine. The obtained results showed that the electrocardiogram and respiration volume were the most relevant signals for human's feeling recognition task. Moreover, the obtained accuracies were promising comparing to recent related works for each of the three establishments of emotion modeling.**

Keywords*—Emotion Classification; MAHNOB-HCI; Peripheral Physiological Signals; Arousal-Valence Space; Support Vector Machine

I. INTRODUCTION

For affective and correct interaction between human and machine (HCI), recognizing human's emotion is a one of the key stage in affective computing field and especially in emotional intelligence for HCI issue. Thus, several researches could be targeted that will benefit from feeling assessment. We cite those done in medicine field and particularly for children with autism who are disable to clearly express their feelings [1]. Emotion recognition system can identify the critical states during driving by detecting the stress level assessments [2][3][4]. Moreover, there are applications that affect daily lives without stress[5] with more pleasing life[6].

The emotion can be noticeable from different modalities. The facial expression is the most popular way to recognize the affective states [7][8][9]. Also, the human speech [10][11] and motions or gestures are very used in emotion assessing problem. However, these channels cannot usually identify the real emotional states because it is easy to secret a facial expression or fake a tone of voice[12]. Moreover, they are not effective for people who cannot reveal their feeling verbally like autistic people [13]. Also, they aren't available unless the user is usually facing to the camera or microphone in an adequate environment with no dark or noise for data collection. To proceed these problems, Picard et al. [14] demonstrated that physiological signals are more pertinent than other modalities. In fact, they originate from the peripheral nervous system and central nervous system. Consequently, they cannot be falsified or hidden. Moreover, they are spontaneous and strongly correlated with human's emotion. The physiological signals are (1) Electroencephalograms (EEG), (2) Electrocardiogram (ECG), (3) Heart Rate Variability (HRV), (4) Galvanic Skin Response (GSR), (5) Muscle Activity or Electromyogram (EMG), (6) Skin Temperature (SKT), (7) Blood Volume Pulse (BVP) and (8) Respiratory Volume (RESP). Since, many studies became very enhanced in emotion recognition problem [15][16][17] and recent works combine physiological signals with two or more modalities to improve the results [18][19].

It is difficult to compare these investigated approaches because they are divergent in different ways. Indeed, the related works are dissimilar in the modality to recognize the affective states that can be natural or induced. Thus, emotion can be evoked by watching affective movies [20], video clips [21], since playing a video game [22], driving a car or listening to music [23][24]. Moreover, the emotion can be defined into different models: the first is Eckman's model that is based on universal emotional expressions to present out six discrete basic emotions: Happiness, Sadness, Surprise, Fear, Anger and Disgust [25]. The second is the Plutchik's model that presents out eight fundamental emotions: Joy, Trust, Fear, Surprise, Sadness, Disgust, Anger and Anticipation [26]. The third is based on Russel et al. model [27] who have focused on two-dimensional evaluation ,like the valence-arousal model [19]. Some other works merge the previous models to define the emotion in the continuous space using affective keywords [28][20].

Among recent and related researches, we cite the work based on MAHNOB dataset [20]. They classified the affective states into three defined classes and they achieved 46.2%, 45.5% for arousal and valence, respectively. Another similar work done by Koelstra et al. who created freely multimodal dataset DEAP, is detailed in [21]. They classified the emotional statements into two classes for valence, arousal and liking. In the previous contribution, they obtained 57%, 62.7%, 59.1% for arousal, valence and liking, respectively. We notice that we cannot directly compare these studies, because they used

different classes in arousal valence model. Moreover, the manner to define the emotion is dissimilar: In fact, in [20], they classified the affective statements using nine discrete emotional keywords tagging to define three classes in arousal valence model. However, in [21], they used discrete self-rating scales from 1 to 9 for arousal, valence and liking.

This paper aims to identify the human affective states into arousal-valence area using three ways of modeling and defining the emotion in this continuous space. The proposed approach is based on peripheral physiological signals (ECG, Resp, Temp and GSR) to use wearable and non-obtrusive sensors for future work. We began by defining the emotional states into two classes which are "positive" and "negative" for valence and "High" and "Low" in arousal. Then, we established three classes using the self-reported discrete scaling values (from 1to 9 scales in arousal and valence axis). The three classes are named calm, medium aroused, and excited, unpleasant, neutral valence and pleasant). Finally, we defined these three classes using nine emotional keywords (Happiness, amusement, neutral, anger, fear, surprise, anxiety, disgust and sadness). In the last emotion's definition, we combined the emotion's model done by Russel and Ekman [28][29][30]. Another purpose of this paper is to select the most relevant peripheral physiological signal for emotion sensing problem. Firstly, we began by classifying one signal and then, we fused their level features. We explored the recent multimodal MAHNOB-HCI database. A judiciously process was applied in preprocessing, features extraction and classification stages. For the last step we used the support vector machine (SVM).

The foregoing of this paper is organized as follows. Section II describes the proposed approach. The next section gives the details of the preprocessing data and feature extraction stages. In section IV, we present SVM classifier and how we modeled the emotion states. The section V summarizes the obtained results. Finally, we conclude this contribution and present future work in section VI.

II. Experimental Setup

A. Proposed approach

The emotion recognition system had several steps that ought to be carefully done to have promising classification rate. As a first step, we pre-processed the data to smooth the signals. Then, a bunch of selected features were extracted. After normalizing all features, an early level feature fusion (LFF) was applied to compare the proposed approach to related works. Finally, we classified the data corresponding to their labels using the support vector machine. All these steps will be detailed in the foregoing sections. Fig .1 presents the block diagram of this work.

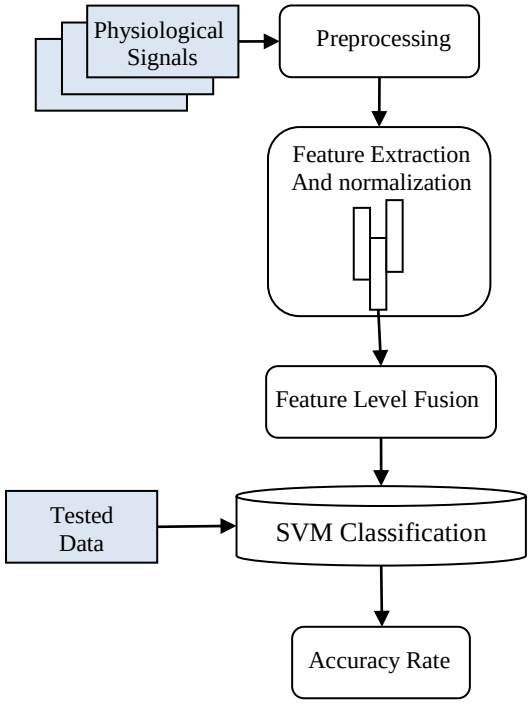


Fig. 1. Block diagram of the proposed approach

B. MAHNOB-HCI Multimodal Database

Investigations in emotion recognition field motivated the establishment of many databases to involve this issue. Some datasets contained speech and audio-visual signals as modalities to assess the human affective states [31][32]. Healey and Picard [33] collected one of the first affective physiological datasets at MIT. Their collected signals were the electrocardiogram (ECG), galvanic skin response (GSR), electromyogram (EMG) and the respiration pattern. This database of stress recognition is publicly available from Physionet[1]. Another novel dataset is the Database for Emotion Analysis using Physiological signals (DEAP) [21]. It contains the spontaneous bodily responses of 32 participants after inducing their emotional states by watching selected music videos clips. This dataset is freely available on the internet[2] for academic research. More recently, Soleymani et al.[20] created the MAHNOB-HCI multimodal database. They recorded the peripheral physiological signals from 24 participants after eliciting their emotion by 20 affective movies. These signals are the Electrocardiogram (ECG), Galvanic Skin Response (GSR), Skin Temperature (Temp) and Respiration Volume (RESP). They also recorded the EEG signal, eye gaze and the face videos.

1 https://physionet.org/pn3/drivedb/.
2 http://www.eecs.qmul.ac.uk/mmv/data sets/deap/.

In the proposed approach, we chose the latest database for several reasons. In fact, it had five modalities which were judiciously synchronized. Also, a comparative study between DEAP and MAHNOB datasets done by Godin et al. [34], demonstrated that the best accuracies were obtained after using the recorded signals in MAHNOB database. Table .I summarizes the content of the MAHNOB-HCI database.

TABLE I. DATABASE SUMMARY[20]

Number of participants	24
Recorded signals	Peripheral physiological signals 256Hz (ECG, GSR, Temp, Resp), Face and body video using 6 cameras (60f/s), 32 channels for EEG signal (256Hz), Eye gaze(60Hz), Audio (44.1kHz)
Number of videos	20
Self-report	Emotional keywords, arousal, valence, dominance and predictability
Rating values	Discrete scale 1-9

III. PREPROCESSING DATA AND FEATURES EXTRACTION

Obtaining promising accuracies required different stages mainly the pre-processing data, features extraction and finally the classification step.

According to the experimental setup of the MAHNOB-HCI database, each trial contained 30 seconds before the beginning of the affective stimuli experience and another 30 seconds after the end. So firstly, we eliminated these two 30 seconds to have the pertinent information. Next, Butterworth filters were applied to eliminate artefacts and baseline wandering for the GSR, Resp and ECG signals. The cut-off frequencies are 0.3 Hz, 0.45 Hz and 0.5 Hz, respectively.

Adding to characteristic features like the heart rate variability from the electrocardiogram (1) and the breathing rate from the respiratory volume, a bunch of statistical values were extracted from the data.

$$HRV = 60/t_{R\_R} \quad (1)$$

Whereas: HRV is the heart rate variability in beats per minute

t_{R_R} : The mean of RR intervals

To reduce the difference between the participants, we normalized features by mapping each one to the interval [0,1]. The preprocessing data and features extraction stages were based on the studies reported in [21] and [20].

IV. SVM CLASSIFICATION

Different machine learning algorithms were successfully applied to classify the human emotional states given a bunch of physiological features. We cite the artificial neural network (ANN), k-Nearest Neighbors (k-NN), Bayesian Network and Regression Tree (RT). In this approach, we employed the support vector machine which is the most popular and pertinent classifier in this issue [35]. Indeed, a comparative study described in [36], proved that the SVM gave the best accuracy rates rather than other machine learning techniques such as k-NN, regression tree and Bayesian network.

Basically, it is a supervised machine learning technique. Adding to linear classification, SVM resolves efficiently a non-linear problem with its several kernels to obtain the optimized classification rates. SVM performs the classification by finding the suitable hyper-plans that separate the classes very well by maximizing the distance between each class and the hyper-plans. For the implementation, we used the LibSVM library under MATLAB platform[3] [37].

Tables .II and .III present the two and three defined classes using 1-9 discrete scales in arousal-valence areas. The rated scales were reported by the participant after/during watching the affective video. On the other hand, we also defined the three classes in arousal valence model using the nine affective keywords, which are (1) Joy or Happiness, (2) Amusement, (3) Sadness, (4) Disgust, (5) Anxiety, (6) Fear, (7) Surprise, (8) Anger, and (9) Neutral. According to the table .IV, we assigned the labels "High" and "Low" for arousal, "Positive" and "Negative" for valence. The three classes were "Clam", "Medium", and "Activated" for arousal and "Unpleasant", "Neutral" and "Pleasant" for valence dimension.

TABLE II. TWO CLASSES IN AROUSAL-VALENCE MODEL

Categorization		**Rating Values "r"**
Arousal	***Valence***	
High	Negative	$r \leq 4.5$
Low	Positive	$4.5 \leq r$

TABLE III. THREE DEFINED CLASSES IN AROUSAL-VALENCE MODEL

Categorization		**Rating Values "r"**
Arousal	***Valence***	
Calm	Unpleasant	$1 \leq r \leq 3$
Medium	Neutral	$4 \leq r \leq 6$
Excited	Pleasant	$7 \leq r \leq 9$

TABLE IV. DEFINED CLASSES IN AROUSAL-VALENCE MODEL USING EMOTIONAL KEYWORDS

	Affective Classes	**Discrete Emotion Tagging**
Arousal	***Clam***	Sadness, Disgust, Neutral
	Medium	Happiness(Joy), Amusement
	Activated	Surprise, Fear, Anger, Anxiety
Valence	***Unpleasant***	Anger, Anxiety, Disgust, Sadness, Fear
	Neutral	Neutral, Surprise
	Pleasant	Happiness(Joy), Amusement

V. RESULTS AND DISCUSSIONS

In this section, we summarize and evaluate the obtained results for emotion classification in arousal valence dimension. We presented the emotional states in two and three defined classes, as explained earlier.

For the aim of this paper, we classified each peripheral physiological signal and then, we applied an early fusion for all descriptors to compare the proposed approach to related studies. The level feature fusion is to combine all the modalities before the training stage[38]. Thus, a simple concatenation was applied to all the extracted features. The table .V summarizes the classification accuracy after testing

[3] http://www.csie.ntu.edu.tw/_cjlin/libsvm/.

several SMV's Kernel functions for two defined classes. In this table, we can clearly note that the ECG and the RESP signals are the most relevant signals for the emotion assessing task, and precisely ECG for arousal and RESP for valence.

We achieved 64.23% in arousal and 68.75% in valence dimension and these accuracies were very promising compared to related works. In fact, Koelstra et .al [21] obtained 62.7% in valence 57% in arousal and Torres Valencia et al.[39] achieved 55% ± 3.9 and 57.50% ± 3.9 in arousal and valence, respectively. Both of these previous studies used the DEAP database. The achieved results prove the potential of the recorded data in MAHNOB-HCI database and their chosen videos were more powerful to evoke the emotion than videos clips used in DEAP. This explanation is well developed in[34]. Indeed, the authors proved that the heart rate variability calculated from the ECG (not available in DEAP), is a very relevant feature in emotion recognition task and it is more accurate than the HRV calculated from the PPG signal which is recorded in DEAP database.

TABLE V. CLASSIFICATION ACCURACIES AFTER USING SEVERAL SVM'S KERNEL (TWO CLASSES)

	Classification Accuracies							
	Linear		**Polynomial**		**Sigmoid**		**Gaussian**	
	Arousal	***Valence***	***Arousal***	***Valence***	***Arousal***	***Valence***	***Arousal***	***Valence***
ECG	65,03%	60,13%	65,73%	60,83%	62,10%	65,03%	66,4%	58,74%
GSR	55,94%	54,73%	53,14%	55,78%	54,73%	55,78%	62,23%	50,52%
RESP	61,05%	62,10%	60,83%	61 ,05 %	60,13%	53,84%	65,03%	62,10%
Temp	57,34%	53,84%	58,74%	53,84%	54,73%	57,34%	60,13%	57,34%
LFF	63,63%	**65,03%**	64,23%	60,83%	60,83%	57,34%	63,63%	**68.75%**

The table .VI presents the accuracies after classifying the emotion into three areas in arousal valence space using the self-reported scaling. On the other hand, Table .VII shows the results after defining the three classes basing on the nine self-reported affective keywords previously reported.

Both of these tables prove that the human's emotion is more noticeable from the respiration and electrocardiogram signals then other bodily responses (Temperature or Galvanic Skin Response). In addition, we can clearly see that the Gaussian kernel function is the best solution that could find the performed hyper-plans. Moreover, it is easier to recognize the emotion after fusing all the peripheral physiological signals, as shown in table V, VI and VII.

The table .VIII resumes the obtained results and three recent related works and it proves that the obtained accuracies are promising in the three ways for emotion's establishments in arousal-valence model.

The achieved accuracies are explained by the fact that we correctly pre-processed the signals to have the significant information. In addition, we warily selected features, which are relevant than chosen in the studies earlier mentioned [20][21][39].

TABLE VI. CLASSIFICATION ACCURACIES BY USING SEVERAL SVM'S KERNELS (3 CLASSES USING SCALING RATES)

	Classification Rates							
	Linear		**Polynomial**		**Sigmoid**		**Gaussian**	
	Arousal	**Valence**	**Arousal**	**Valence**	**Arousal**	**Valence**	**Arousal**	**Valence**
ECG	51,4%	44,05%	50%	52,12%	52,63%	46,80%	50,07%	46,8%
GSR	47,36%	48,93%	48,42%	48,93%	49,37%	45,74%	50,52%	47,87%
Resp	47,36%	53,19%	46,31%	50%	48,42%	48,93%	45,77%	52%
Temp	40,14%	43,3%	42,95%	45,45%	45,26%	51,05%	42,25%	45,45%
LFF	52,63%	48,93%	50,52%	52,12%	51,57%	46,36%	**54,73%**	**56,83%**

TABLE VII. CLASSIFICATION ACCURACIES BY USING SEVERAL SVM'S KERNEL (THREE DEFINED CLASSES USING SELF-REPORTED EFFECTIVE KEYWORDS)

	Classification Accuracies							
	Linear		**Polynomial**		**Sigmoid**		**Gaussian**	
	Arousal	***Valence***	***Arousal***	***Valence***	***Arousal***	***Valence***	***Arousal***	***Valence***
ECG	44,21%	48,42%	43,15%	49,47%	50,52%	48,42%	45,26%	51,57%
GSR	40%	41,05%	42,10%	43,15%	46,31%	44,4%	45,26%	42,10%
Resp	51,57%	49,47%	52,63%	46,31%	47,36%	48,42%	49,47%	44,21%
Temp	41,57%	43,15%	41,05%	44,2%	46,31%	45,26%	45,26%	45,26%
LFF	48,42%	48,42%	50,52%	54,73%	46,31%	49,47%	**59,57%**	**57.44%**

TABLE VIII. OBTAINED RESULTS COMPARED TO RELATED WORK FOR THE THREE WAYS IN MODELLING EMOTIONS

	Two Classes			**Three Classes**		
				Using 1-9 scaling values	***Using 9 emotional keywords***	
	Obtained Results	[21]	[39]	*Obtained Results*	*Obtained Results*	[20]
Arousal	**64.23%**	57%	55.00% ± 3.9	**54,73%**	**59,57%**	46.2%
Valence	**68.75%**	62.7%	57.50% ± 3.9	**56,83%**	**57.44%**	45.5%

VI. CONCLUSION

This paper presented a whole process in emotion recognition system from peripheral physiological signals. For this aim, we used the recent multimodal database MAHNOB-HCI. In this dataset, they collected the bodily responses from 24 participants after eliciting their feeling using 20 selected videos. Basing on the self-reported emotion from the participant, we proposed three ways to model the affective states in arousal valence space. In fact, we established two and three main classes using the discrete rating values (from 1 to 9 scales) and another model using 9 emotional keywords to define three areas in arousal valence dimension. We pre-processed the data to remove noise and artefacts from the data. Then, we extracted selected features. After normalizing them to minimize the difference between participants, an early level feature fusion was applied for further analysis. Finally, we classified for the first time each physiological signal and then the LFF data using the support vector machine. We used its different kernel's functions to perform the classification rates. Results showed the relevance of the electrocardiogram and respiration signals in emotion assessment task. Moreover, the RBF kernel is the most suitable algorithm. Results proved also, that detecting affective states is easier after fusing all the bodily responses. The obtained accuracies were promising compared to recent related works.

As future work, we aim to implement additional techniques such as the feature selection and reduction mechanisms (ANOVA, PCA, and Fisher) to eliminate the redundant information and select the most relevant features. Moreover, we would like to implement other classification algorithms that can lead for best results.

ACKNOWLEDGMENT

The authors of this paper would like to thank the MAHNOB-HCI's team for providing this freely multimodal database to develop this research[4].

REFERENCES

[1] K. G. Smitha and A. P. Vinod, "Hardware efficient FPGA implementation of emotion recognizer for autistic children," IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT), pp. 1–4, Jan. 2013.

[2] Benoit et al., "Multimodal focus attention and stress detection and feedback in an augmented driver simulator," Pers. Ubiquitous Comput., vol. 13, no. 1, pp. 33–41, Jan. 2009.

[3] D. Katsis, N. Katertsidis, G. Ganiatsas, and D. I. Fotiadis, "Toward Emotion Recognition in Car-Racing Drivers: A Biosignal Processing Approach," IEEE Trans. Syst. Man Cybern. - Part Syst. Hum., vol. 38, no. 3, pp. 502–512, May 2008.

[4] M. Paschero et al., "A real time classifier for emotion and stress recognition in a vehicle driver," IEEE International Symposium on Industrial Electronics , pp. 1690–1695, May 2012.

[5] Ghaderi, J. Frounchi, and A. Farnam, "Machine learning-based signal processing using physiological signals for stress detection," 22nd Iranian Conference on Biomedical Engineering (ICBME), pp. 93–98, November 2015.

[6] K. Rattanyu, M. Ohkura, and M. Mizukawa, "Emotion monitoring from physiological signals for service robots in the living space," in International Conference on Control Automation and Systems (ICCAS), pp. 580–583 , Octobre 2010

[7] Suchitra, Suja P., and S. Tripathi, "Real-time emotion recognition from facial images using Raspberry Pi II," 3rd International Conference on Signal Processing and Integrated Networks (SPIN), pp. 666–670, February 2016.

[8] N. Chanthaphan, K. Uchimura, T. Satonaka, and T. Makioka, "Facial Emotion Recognition Based on Facial Motion Stream Generated by Kinect," 11th International Conference on Signal-Image Technology Internet-Based Systems (SITIS), pp. 117–124, November 2015.

[9] Turan, K.-M. Lam, and X. He, "Facial expression recognition with emotion-based feature fusion," Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), pp. 1–6, December 2015.

[10] S. G. Koolagudi and K. S. Rao, "Emotion recognition from speech: a review," Int. J. Speech Technol., vol. 15, no. 2, pp. 99–117, Jun. 2012.

[11] F. Chenchah and Z. Lachiri, "Acoustic Emotion Recognition Using Linear and Nonlinear Cepstral Coefficients," Int. J. Adv. Comput. Sci. Appl., vol. 6, no. 11, 2015.

[12] S. Wioleta, "Using physiological signals for emotion recognition," 6th International Conference on Human System Interactions (HSI), pp. 556–561, June 2013.

[13] R. W. Picard, E. Vyzas, and J. Healey, "Toward machine emotional intelligence: analysis of affective physiological state," IEEE Trans. Pattern Anal. Mach. Intell., vol. 23, no. 10, pp. 1175–1191, October 2001.

[14] Y. Velchev, S. Radeva, S. Sokolov, and D. Radev, "Automated estimation of human emotion from EEG using statistical features and SVM," Digital Media Industry Academic Forum (DMIAF), pp. 40–42, July 2016.

[15] Y. Gu, K.-J. Wong, and S.-L. Tan, "Analysis of physiological responses from multiple subjects for emotion recognition," IEEE 14th International Conference on e-Health Networking, Applications and Services (Healthcom), pp. 178–183, October 2012.

[16] S. Basu et al., "Emotion recognition based on physiological signals using valence-arousal model," Third International Conference on Image Information Processing (ICIIP), pp. 50–55, December 2015.

[17] S. Thushara and S. Veni, "A multimodal emotion recognition system from video," International Conference on Circuit, Power and Computing Technologies (ICCPCT), pp. 1–5, March 2016.

[18] A. Torres, Á. A. Orozco, and M. A. Álvarez, "Feature selection for multimodal emotion recognition in the arousal-valence space," 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 4330–4333, July 2013.

[19] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, "A Multimodal Database for Affect Recognition and Implicit Tagging," IEEE Trans. Affect. Comput., vol. 3, no. 1, pp. 42–55, Jan. 2012.

[20] S. Koelstra et al., "DEAP: A Database for Emotion Analysis ;Using Physiological Signals," IEEE Trans. Affect. Comput., vol. 3, no. 1, pp. 18–31, Jan. 2012.

[4] https://mahnob-db.eu/hci-tagging

[21] G. N. Yannakakis, K. Isbister, A. Paiva, and K. Karpouzis, "Guest Editorial: Emotion in Games," IEEE Trans. Affect. Comput., vol. 5, no. 1, pp. 1–2, Jan. 2014.

[22] Jonghwa Kim and E. Andre, "Emotion recognition based on physiological changes in music listening," IEEE Trans. Pattern Anal. Mach. Intell., vol. 30, no. 12, pp. 2067–2083, Dec. 2008.

[23] Jonghwa Kim and E. Andre, "Emotion recognition based on physiological changes in music listening," IEEE Trans. Pattern Anal. Mach. Intell., vol. 30, no. 12, pp. 2067–2083, Dec. 2008.

[24] J. Fleureau, P. Guillotel, and Q. Huynh-Thu, "Physiological-Based Affect Event Detector for Entertainment Video Applications," IEEE Trans. Affect. Comput., vol. 3, no. 3, pp. 379–385, Jul. 2012.

[25] "Sentiment Symposium Tutorial: Language and cognition." [Online]. Available: http://sentiment.christopherpotts.net/lingcog.html. [Accessed: 19-Feb-2017].

[26] S. M, J. A. Russell, and L. F. Barrett, "Structure of self-reported current affect: Integration and beyond," J. Pers. Soc. Psychol., vol. 77, no. 3, pp. 600–619, 1999.

[27] H. Xu and K. N. Plataniotis, "Subject independent affective states classification using EEG signals," IEEE Global Conference on Signal and Information Processing (GlobalSIP), pp. 1312–1316, December 2015.

[28] "Emotion, core affect, and psychological construction." [Online]. Available: https://www.researchgate.net/publication/276947577_Emotion_core_affect_and_psychological_construction. [Accessed: 17-Nov-2016].

[29] P. Ekman and D. Cordaro, "What is Meant by Calling Emotions Basic," Emot. Rev., vol. 3, no. 4, pp. 364–370, October 2011.

[30] G. McKeown, M. F. Valstar, R. Cowie, and M. Pantic, "The SEMAINE corpus of emotionally coloured character interactions," IEEE International Conference on Multimedia and Expo, pp. 1079–1084, July 2010.

[31] M. Grimm, K. Kroschel, and S. Narayanan, "The Vera am Mittag German audio-visual emotional speech database," IEEE International Conference on Multimedia and Expo , pp. 865–868, June 2008.

[32] J. A. Healey and R. W. Picard, "Detecting Stress During Real-World Driving Tasks Using Physiological Sensors," IEEE Trans. Intell. Transp. Syst., vol. 6, no. 2, pp. 156–166, Jun 2005.

[33] C. Godin, F. Prost-Boucle, A. Campagne, S. Charbonnier, S. Bonnet, and A. Vidal, "Selection of the most relevant physiological features for classifying emotion," Proceedings in International conference on physiological computing systems (PhyCS), January 2015.

[34] C. J. C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition," Data Min. Knowl. Discov., vol. 2, no. 2, pp. 121–167.

[35] Changchun Liu, P. Rani, and N. Sarkar, "An empirical study of machine learning techniques for affect recognition in human-robot interaction," IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 2662–2667, August 2005.

[36] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," ACM Trans. Intell. Syst. Technol., vol. 2, no. 3, pp. 1–27, Apr. 2011.

[37] Z. Guendil, Z. Lachiri, C. Maaoui, and A. Pruski, "Emotion recognition from physiological signals using fusion of wavelet based features," in 2015 7th International Conference on Modelling, Identification and Control (ICMIC), pp. 1–6, December 2015.

[38] C. A. Torres-Valencia, H. F. Garcia-Arias, M. A. A. Lopez, and A. A. Orozco-Gutierrez, "Comparative analysis of physiological signals and electroencephalogram (EEG) for multimodal emotion recognition using generative models," Signal Processing and Artificial Vision 2014 XIX Symposium on Image, pp. 1–5, September 2014.

Performance Analysis of Route Redistribution among Diverse Dynamic Routing Protocols based on OPNET Simulation

Zeyad Mohammad[1]
Faculty of Science and Information Technology
Al Zaytoonah University of Jordan
Amman, 11733 Jordan

Adnan A. Hnaif[3]
Faculty of Science and Information Technology
Al Zaytoonah University of Jordan
Amman, 11733 Jordan

Ahmad Abusukhon[2]
Faculty of Science and Information Technology
Al Zaytoonah University of Jordan
Amman, 11733 Jordan

Issa S. Al-Otoum[4]
Faculty of Science and Information Technology
Al Zaytoonah University of Jordan
Amman, 11733 Jordan

***Abstract*—Routing protocols are the fundamental block of selecting the optimal path from a source node to a destination node in internetwork. Due to emerge the large networks in business aspect thus; they operate diverse routing protocols in their infrastructure. In order to keep a large network connected; the implementation of the route redistribution is needed in network routers. This paper creates the four scenarios on the same network topology by using Optimized Network Engineering Tools Modeler (OPNET 14.5) simulator in order to analyze the performance of the route redistribution among three routing protocols by configuring three protocols from a set of Routing Information Protocol (RIPv2), Enhanced Interior Gateway Routing Protocol (EIGRP), Open Shortest Path First (OSPF), and Intermediate System to Intermediate System (IS-IS) dynamic routing protocols on each scenario. The first scenario is EIGRP_OSPF_ISIS, the second scenario is EIGRP_OSPF_RIPv2, the third scenario is RIPv2_EIGRP_ISIS, and the fourth scenario is RIPv2_OSPF_ISIS. The simulation results showed that the RIPv2_EIGRP_ISIS scenario outperforms the other scenarios in terms of network convergence time, the hops number, jitter, packet delay variation, packet end to end delay; therefore, it fits real time applications such as voice and video conferencing. In contrast, the EIGRP_OSPF_ISIS scenario has better results compared with other scenarios in terms of response time in case of using web browsing, database query, and Email services.**

Keywords—route redistribution; dynamic routing protocols; EIGRP; IS-IS; OSPF; RIPv2

I. INTRODUCTION

Nowadays, Internet has transformed the life style from a classical environment to a technology based one. Due to the importance of routing protocols in Internet infrastructure, several routing issues and requirements must be considered in a network design phase. A Routing is a fundamental process for selecting optimal path from source to destination nodes. Routing protocols consist of interior and exterior gateway protocols. Border Gateway Protocol (BGP) is an exterior gateway protocol. BGP is designed to exchange routing information among autonomous system (AS) on the Internet. It is considered a distance vector routing protocol. An Interior Gateway Protocol is used to exchange routing information between gateways within an AS. It consists of distance vector and link state routing protocols. A distance vector algorithm builds a vector that contains costs to all other nodes and distributes a vector to its neighbors. A link state algorithm in which each node finds out the state of the link to its neighbors and the cost of each link. A distance vector routing protocol is a hop count metrics and the next hop presents a direction. It is based on Bellman Ford algorithm to calculate the optimal path. RIP is a distance vector routing protocol that measures its metrics by counting the number of hops between source and destination. It selects the minimum number of hops for reaching a destination. RIP has three versions; this study will consider RIPv2 only in a simulation. EIGRP is a distance vector routing protocol that it uses diffusion update algorithm to select the minimum cost between source and destination. A link state routing protocol is based on Dijkstra's algorithm to find a shortest path between source and destination. OSPF and IS-IS are a link state routing protocol. The enterprise networks are created from numerous routers that are running diverse routing protocols in order to exchange their route information; the configuration of the route redistribution in routers are needed. The route redistribution exchanges the route information between two different routing protocols that requires a common border router. A common border router runs routing processes in both routing protocols. The border router may be configured to redistribute routes from one routing protocol to the other, and vice versa. The route redistribution is needed in case of company mergers, multiple departments managed by multiple network administrators, multi−vendor environment, and split of two independent routing domains [1-2]. The route redistribution has two main goals. The first goal is to advertise routing information between different routing protocols for connectivity purposes. The second goal is route back up in case of a network failure, routing protocol should support alternate forwarding paths to each other. Moreover, most of existing solutions apply to

scenarios with only two routing protocols, but large operational networks usually include more than two routing protocols [3].

The route redistribution might raise issues during running multiple different routing protocols due to each routing protocol has its characteristics such as metrics, administrative distance, convergence rate, classful and classless capabilities. Each routing protocol uses different metrics in order to calculate the optimal path. RIPv2 uses a hop count in its metric, and its administrative distance is 120, but EIGRP uses bandwidth, delay, reliability, load, and maximum transmission unit (MTU) in its metric, where bandwidth and delay are default metric in EIGRP, and its administrative distance is 90 [4]. OSPF metric is based on bandwidth, and its administrative distance is 110 , but IS-IS metric is based on cost of link utilization, delay, expense and error, where Cisco implementation uses cost only, and its administrative distance is 115 [5-6].

Each routing protocol has a different network convergence time such as EIGRP convergence time is faster than RIP. A network convergence is the status of a group of routers that have the same topological information about network in which they work. When a link fails or recovers thereafter a set of routers needs to run their routing protocols in order to exchange their routing information with neighbors to form the same topological information about their network.

Many researchers have analyzed and compared the performance of the link state and distance vector dynamic routing protocols, and the route redistribution between two diverse routing protocols [7-29]. In an enterprise network might contain more than two diverse routing protocols in order to operate. This study focuses on analyzing and comparing the performance of the route redistribution among three different routing protocols that operate in an enterprise network. This paper creates the four scenarios on the OPNET 14.5 simulator in order to analyze the performance of diverse combinations of different routing protocols that operate in the same network. The first scenario is named by EIGRP_OSPF_RIPv2 that is configured from EIGRP, OSPF and RIPv2 routing protocols in the network topology. The RIPv2_OSPF_ISIS scenario is configured from RIPv2, OSPF and IS-IS routing protocols that is a second scenario. The third scenario is named by RIPv2_EIGRP_ISIS that is configured from RIPv2, EIGRP and IS-IS routing protocols. The fourth scenario is named by EIGRP_OSPF_ISIS that is configured from EIGRP, OSPF and IS-IS routing protocols. The goal of this paper is to analyze the performance of the four scenarios in terms of convergence duration time, number of hops, voice jitter, voice and video conferencing packet delay variations, voice and video conferencing packet end to end delays, remote login, database query, HTTP object, HTTP page, Email upload and Email download response times.

The rest of the paper is organized as follows: Section 2 presents a review briefly about the performance analysis of dynamic routing protocols and the route redistribution of different routing protocols. Section 3 describes the four scenarios of the designed network topology that have been created by the OPNET 14.5 simulator. A performance analysis of the four scenarios and their results discussion are presented in section 4. The conclusion and future works are presented in section 5.

II. Related Works

Abdulkadhim analyzed the performance of EIGRP, OSPF and RIP dynamic routing protocols in terms of the network convergence activity and time by using the OPNET simulator. He showed that OSPF has faster convergence time than RIP, and OSPF convergence activity is much more than RIP, therefore, OSPF can react more quickly in case of link failure [7]. Kodzo et al. simulated EIGRP, OSPF and their combination in OPNET. They analyzed the performance of EIGRP, OSPF and EIGRP_OSPF for real time application. They found that EIGRP_OSPF has less end to end delay, packet delay variation and packet loss for real application than both EIGRP and OSPF, and the combination of EIGRP and OSPF has maximum throughput than EIGRP and OSPF [8]. Mardedi and Rosidi presented the analysis and comparison of performance between EIGRP and OSPF based on Cisco Packet Tracer 6.0.1. They found that EIGRP is better than OSPF in terms of delay and convergence time [9]. Whitfield and Zhu compared the performance of OSPFv3 and EIGRPv6 by using real Cisco hardware in experiments. They noticed that EIGPRv6 outperforms OSPFv3 in terms of start-up and re-convergence speed but EIGRPv6 authentication mechanism negatively affected its performance, in contrast IP Security (IPSec) in OSPFv3 improved its performance [10]. Dey et al. presented a simulation based on Cisco Packet Tracer for dynamic routing protocols and redistribution among the protocols [11]. Patel et al analyzed the performance of OSPF and EIGRP routing protocols in terms of route summarization and route redistribution in Graphical Network Simulator (GNS3) [12]. Farhangi et al. presented the OPNET simulation based of a combination of EIGRP, OSPF and IS-IS routing protocols in a semi-mesh topology. A simulation showed that the performance of the mixed three protocols EIGRP, OSPF and IS-IS in terms of end to end delay, packet delay variation, Voice Jitter and Link throughput outperforms the other two combination of the same three routing protocols [13]. Jalali et al. evaluated the performance of RIP, OSPF, IGRP and EIGRP in terms of convergence, throughput, queuing delay, end to end delay and utilization by using the OPNET simulator. They found that EIGRP outperforms other routing protocols in their study [14]. Ashoor presented a survey in distance vector and link state dynamic routing protocols. She analyzed the performance of distance vector and link state algorithms in a mesh network [15]. Kuradusenge and Hanyuwimfura presented a comparative analysis of EIGRP configuration on IPv4 and IPv6 by modifying its metric of different values of composite metric to path selection [16]. Kaur and Mir demonstrated a comparative performance analysis of EIGRP, RIP and OSPF by using the OPNET simulator. They concluded that EIGRP is better than OSPF and RIP in terms of network convergence, throughput, utilization, queuing delay, HTTP page response and email upload response time [17]. Singh et al. configured EIGRP on IPv6 by using Cisco Packet Tracer simulator and evaluated the performance of EIGRP in IPv6 for small network [18]. Pavani et al. surveyed the performance of dynamic routing protocols in terms of router updates, link utilization and end to end delay [19]. Priyadhivya and Vanitha simulated RIP

and OSPF in IPv6 configuration by using GNS3 emulator. They analyzed the performance of OSPF and RIP in terms of convergence and packet loss, and their result showed that OSPF has faster convergence and less packet loss [20]. Shah and Rana analyzed the convergence time of OSPF and RIP by using the OPNET. They found that OSPF single area convergence time outperforms OSPF multi area and OSPF multi stub area and the convergence time of RIP is better inside network core than outside network core [21]. Vissicchio et al. presented a study in the route redistribution with safe router configuration, and they demonstrated the self-sustained routing loops and sub-optimal routing paths problems that might occur [22]. Ud Din et al. evaluated the performance of RIP, OSPF, IGRP, and EIGRP in terms of packets dropping, traffic received, end to end delay, and jitter. They used OPNET to simulate the network in their study and they found that IGRP outperforms the other routing protocols in their simulation [23]. Kaur and Singh presented a comparative performance analysis of IS-IS, OSPFv3 and the combination of IS-IS and OSPFv3 by using the OPNET simulator. They found that IS-IS protocol is better than others in terms of video end to end delay, OSPFv3 is better in jitter and IS-IS_OSPFv3 is better in voice end to end delay[24]. ShewayeSirika and SmitaMahajine studied RIP, EIGRP and OSPF in details and simulated these routing protocols on the OPNET and Cisco Packet Tracer simulators in order to compare their performance in terms of real time applications. They concluded that RIPv2 is suitable for small network and OSPF fits large network [25]. Gehlot and Barwar compared and evaluated the performance of EIGRP and OSPF by using best effort and quality of service method in OPNET simulator. They found that EIGRP outperforms OSPF performance in both quality of service and best effort [26]. NavaneethKrishnan et al. compared EIGRP and OSPF in terms of resource usage by using Cisco Packet Tracer. They found that EIGRP uses fewer resources than OSPF [27]. Al-Hadidi et al. presented a comparative performance evaluation between OSPF and EIGRP by using the OPNET and GNS3. They concluded that the performance of EIGRP is better than OSPF in real time application [28]. Kumar et al. implemented an experiment of route redistribution between EIGRP and OSPF routing protocol in computer network using GNS3 emulator [29].

III. The Proposed Network Topology

In order to analyze the performance of the route redistribution among three different routing protocols that is mixed from RIPv2, OSPF, EIGRP, and IS-IS. Four scenarios were created and implemented in the same network topology. The proposed network topology in this study consists of 15 routers, where R6 and R10 are the border routers that are used to exchange different routing information among the other routers, two switches, six servers, two work stations, four LAN 100BaseT local area networks, where 100BaseT_LAN object presents a fast Ethernet in a switched topology, Point to Point Digital Signal (PPP DS3) link is used to connect routers in which it supports 44.736 Mbps data rate, Ethernet 100BaseT is used to connect other components in our simulation, where 100BaseT duplex link presents Ethernet connection with 100 Mbps speed. The six servers in the proposed network topology consists of two servers that provide multimedia services, where voice Server provides voice with pulse code modulation (PCM) Quality and Silence Suppressed, and video server supports video conferencing with high resolution video, Email, remote login and database servers provide services with high load traffic, HTTP server supports web service with heavy browsing. In order to analyze the network convergence duration time of the proposed network, a Failure Recovery node is used in the proposed network in order to simulate of fails in links of the real communication networks. The link between R6 and R8 nodes is an important communication link in the proposed network due to the path between source and destination nodes is the shortest path (R10→R8→R6) as compared with the other path (R10→R9→R7→R6), so during the simulations, we apply failure recovery events as shown in Table 1, where the time is given in second. The total simulation time for each scenario is taken to be 15 minutes.

TABLE I. Link Failure and Recovery between R6 and R8

Failure	Recovery
120	300
420	480
540	570
630	640
700	705
765	766

Fig. 1 shows the EIGRP_OSPF_RIPv2 scenario that is a combination of EIGRP, OSPF, and RIPv2 routing protocols. The R6 and R10 are the border routers that are used to distribute different routing information among the other routers, where R6 is used to distribute EIGRP and OSPF, in the other side R10 distributes RIPv2 and OSPF routing information.

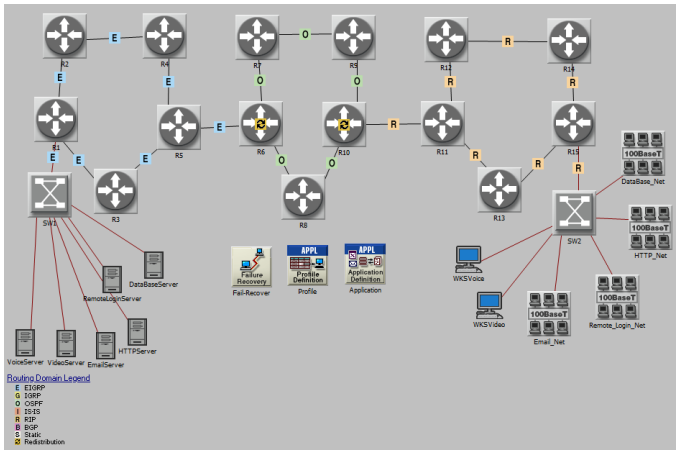


Fig. 1. The Route Redistribution among EIGRP, OSPF, and RIPv2

The RIPv2_OSPF_ISIS scenario is a combination of RIPv2, OSPF and IS-IS routing protocols that is shown in Fig. 2. In this scenario, R6 is used to advertise RIPv2 and OSPF in the proposed network, in the other side R10 is used to advertise OSPF and IS-IS routing information.

The RIPv2_EIGRP_ISIS scenario is depicted in Fig. 3, the route redistribution among RIPv2 and EIGRP is used by R6 router and R10 is used to distribute EIGRP and IS-IS routing information to the other side in the proposed network.

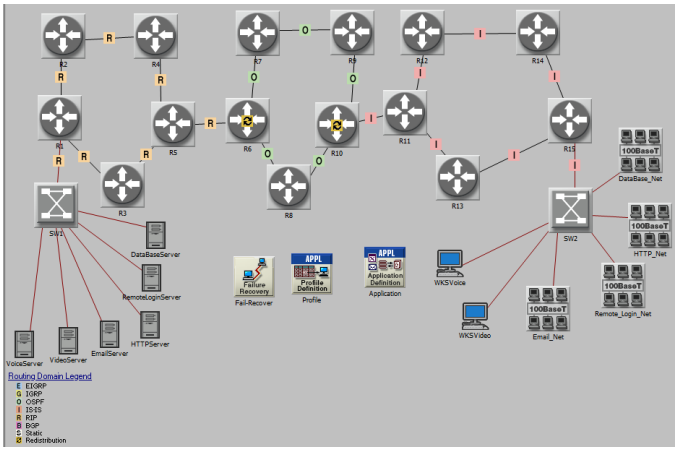


Fig. 2. The Route Redistribution among RIPv2, OSPF, and IS-IS

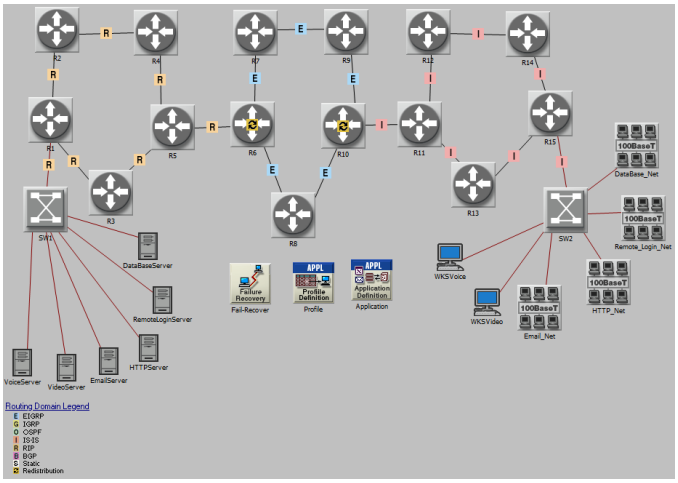


Fig. 3. The Route Redistribution among RIP, EIGRP, and IS-IS

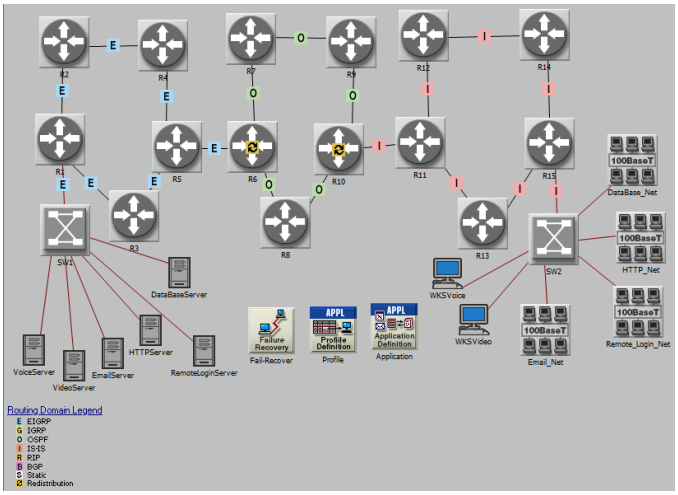


Fig. 4. The Route Redistribution among EIGRP, OSPF, and IS-IS

Fig. 4 shows the fourth scenario EIGRP_OSPF_ISIS, where the R6 is the border router that is used to advertise different routing information from EIGRP and OSPF, on the other side the R10 is the border router that is used to distribute OSPF and IS-IS routing information in the proposed network.

IV. Results and Discussion

This section presents the results that obtained from the simulations of the four scenarios in this study, therefore, the simulation results are analyzed and compared for the proposed scenarios then a decision is made about the scenarios in terms of the fitting applications for each scenario.

A. Network Convergence Time

A Failure Recovery node is applied in the proposed network as shown in Table 1 in order to analyze an average convergence duration time of the simulated network topology in this paper, where a convergence time is a measure of a time that a set of routers need to converge the network to a stable status, and a convergence duration time demonstrates how fast the convergence to reach a stable state in the network. Fig. 5 shows the RIPv2_EIGRP_ISIS scenario that has less convergence time compared with the other scenarios in the sense of the RIPv2_EIGRP_ISIS scenario has the smallest value of a convergence duration time before a failure to be occurred in the network and after network recovery among the other scenarios. Therefore, a convergence duration time in the route redistribution among three protocols RIPv2, EIGRP, and IS-IS is the fastest one in network convergence time.

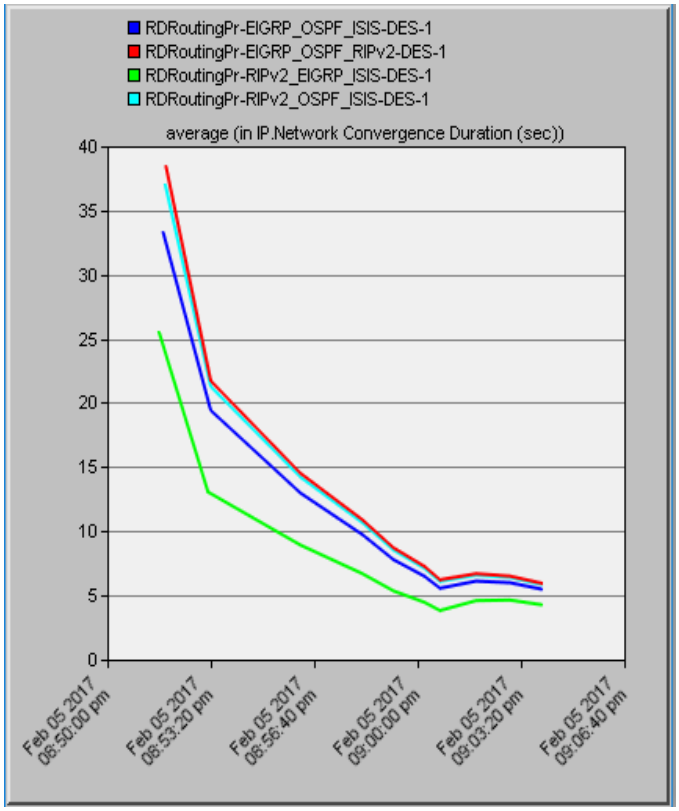


Fig. 5. An Average Convergence Duration Time in the Four Scenarios

B. Hops Number

Fig. 6 shows the RIPv2_EIGRP_ISIS scenario that has the optimal path compared with the other scenarios, because the hops number in the RIPv2_EIGRP_ISIS scenario are 10 hops before a failure to be occurred and it has the same number of hops after network recovery, but the other scenarios have 10

hops number before a failure to be happened and they have 11 hops after network recovery. Therefore, the shortest path of the route redistribution among three protocols RIPv2, EIGRP, and IS-IS is the best one.

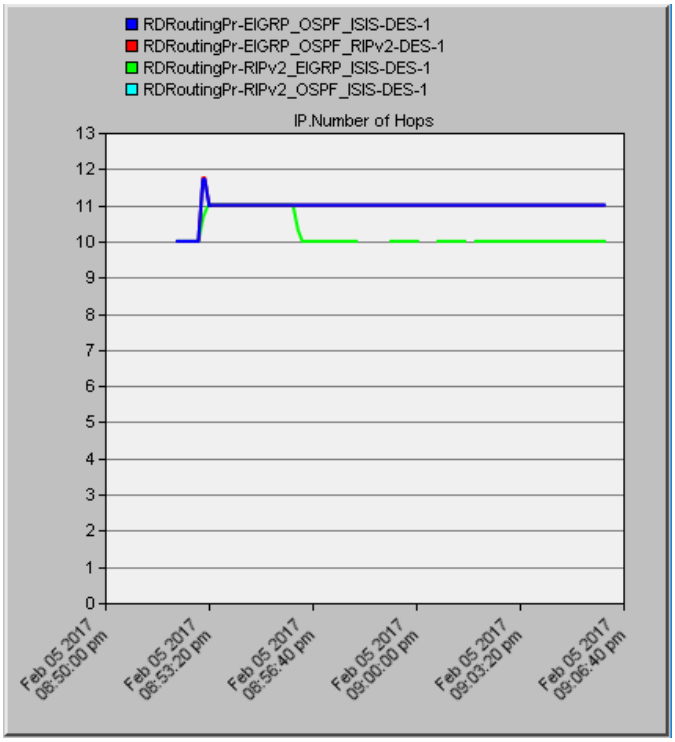


Fig. 6. The Hops Number in the Four Scenarios

C. Response Time

The remote login service in the RIPv2_EIGRP_ISIS scenario has the worst response time compared with the others three scenarios. In contrast, the EIGRP_OSPF_RIPv2 scenario is the fastest response time in the case of the remote login service; therefore, it is the best one as shown in Fig. 7.

Fig. 8 shows the EIGRP_OSPF_ISIS scenario that has less response time as compared with the three other scenarios, therefore, the route redistribution among three protocols EIGRP, OSPF, and IS-IS is the best in terms of using a service of data base query.

Fig. 9 shows the four scenarios in term of the HTTP object response time, where the route redistribution among three protocols RIPv2, OSPF, and IS-IS in the scenario RIPv2_OSPF_ISIS has the worst object response time among the three other scenarios. In contrast, The EIGRP_OSPF_ISIS scenario has less response time in terms of the HTTP object service, therefore, the scenario EIGRP_OSPF_ISIS is the best one.

The RIPv2_EIGRP_ISIS scenario is the slowest response time in terms of the HTTP page service as shown in Fig. 10. On the other hand, the three other scenarios have faster response time and their response time are the same; therefore, they are the best in this service.

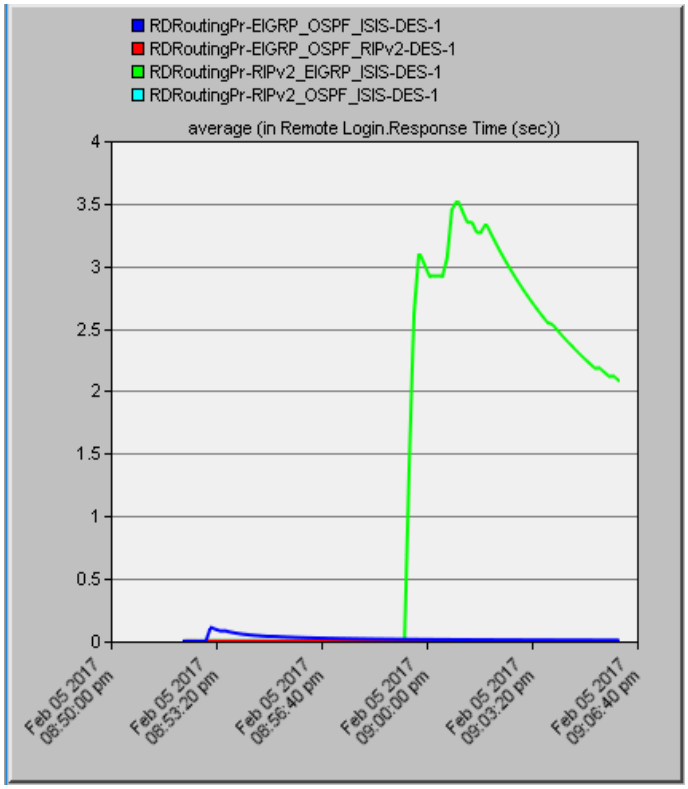


Fig. 7. The Remote Login Response Time in the four scenarios

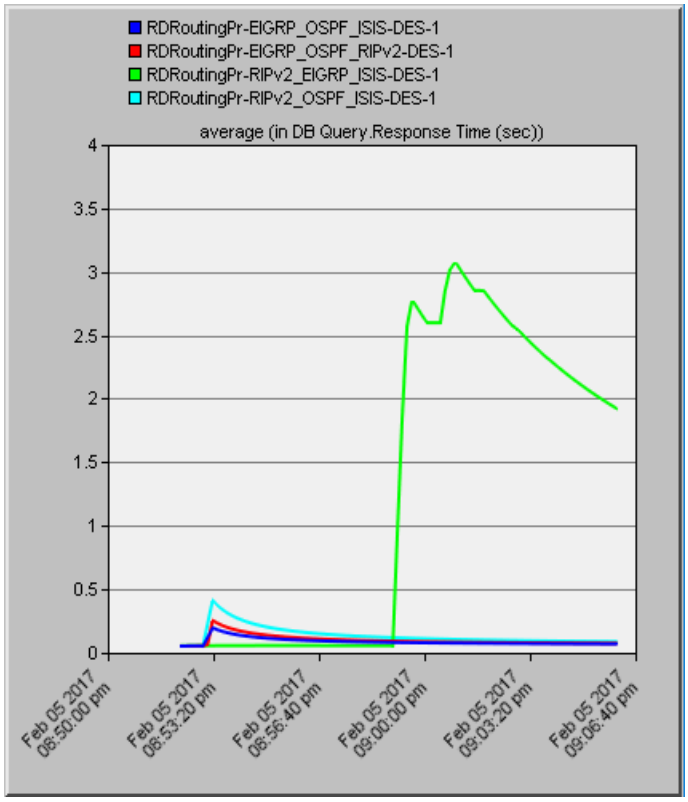


Fig. 8. The Data base Query Response Time in the four scenarios

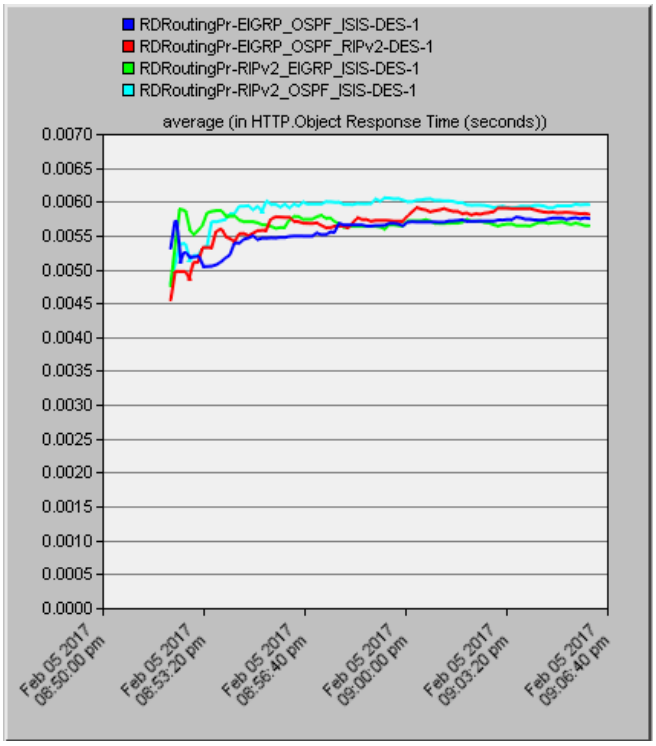


Fig. 9. The HTTP Object Response Time in the four scenarios

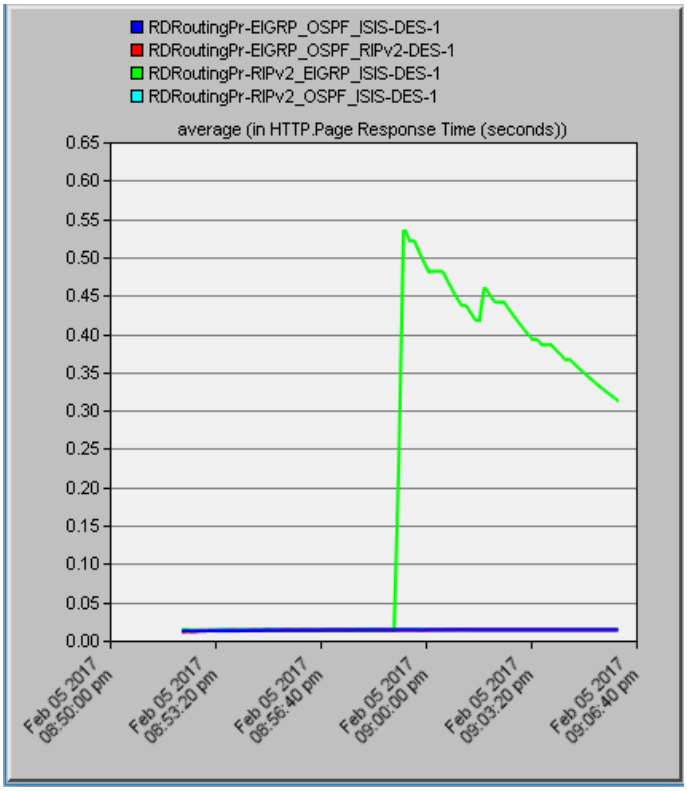


Fig. 10. The HTTP page response time in the four scenarios

Fig. 11 shows the RIPv2_EIGRP_ISIS scenario that has the worst response time compared with the three other scenarios. In contrast, the three other scenarios are faster than the RIPv2_EIGRP_ISIS scenario in terms of using the Email upload service that have the same response time thus they have the best response time.

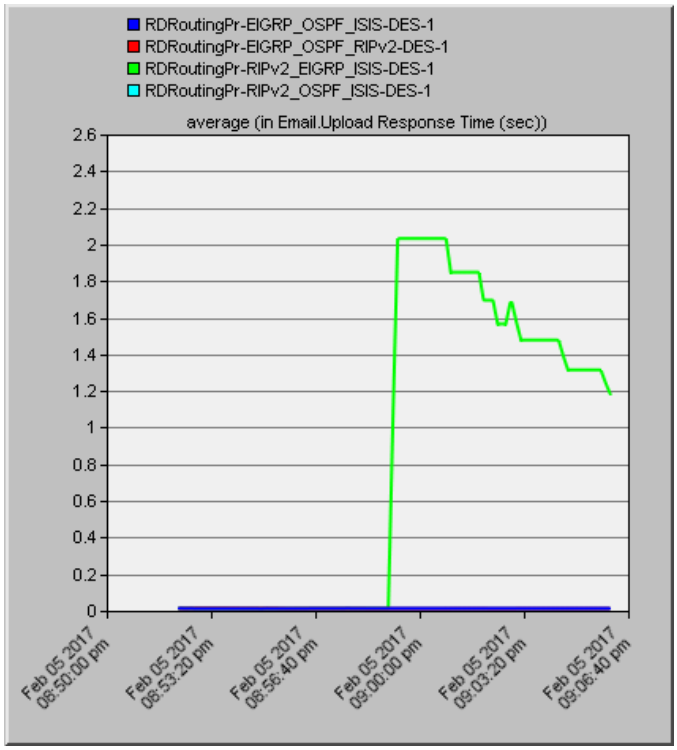


Fig. 11. The Email Upload Response Time in the four scenarios

The performance analysis in term of Email download response time is showed in Fig. 12, where the RIPv2_EIGRP_ISIS scenario has the worst result as compared with the three other scenarios in this simulation.

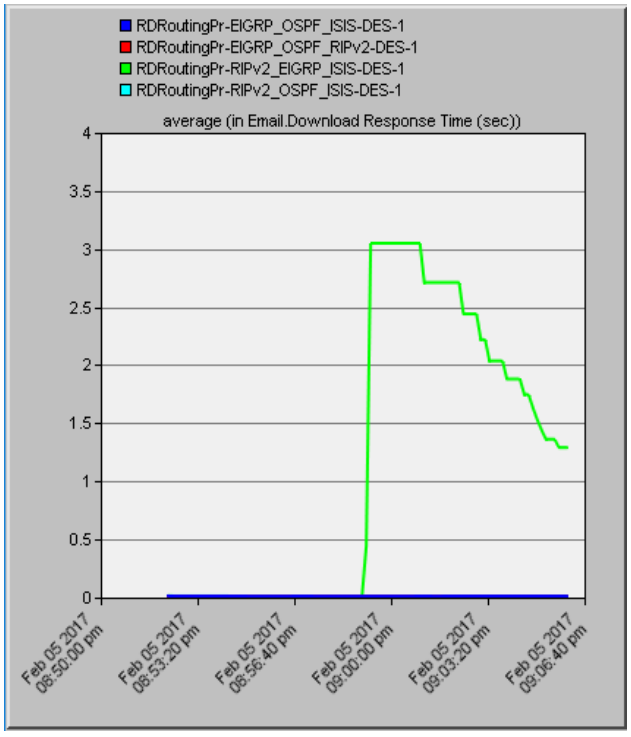


Fig. 12. The Email Download Response Time in the Four Scenarios

D. Voice Jitter

The performance analysis in terms of multimedia service, the voice and video conferencing services are used in order to demonstrate the results in this study. Fig. 13 shows that the RIPv2_EIGRP_ISIS scenario is the best voice jitter from among the three other scenarios, where a jitter is a variation in delay time of received packets. In contrast, the EIGRP_OSPF_RIPv2 scenario is the worst one.

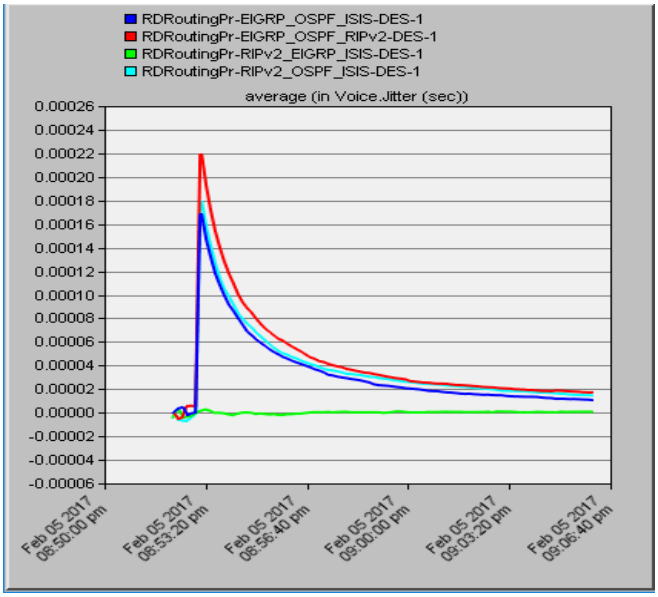


Fig. 13. The Voice Jitter in the Four Scenarios

E. Voice Packet Delay Variation

The RIPv2_EIGRP_ISIS scenario outperforms the three other scenarios in terms of the voice packet delay variation that has less delay variation as shown in Fig. 14, where a delay variation is a delay in receiving packets at the receiver. In the other hand, the EIGRP_OSPF_ISIS scenario has the worst result.

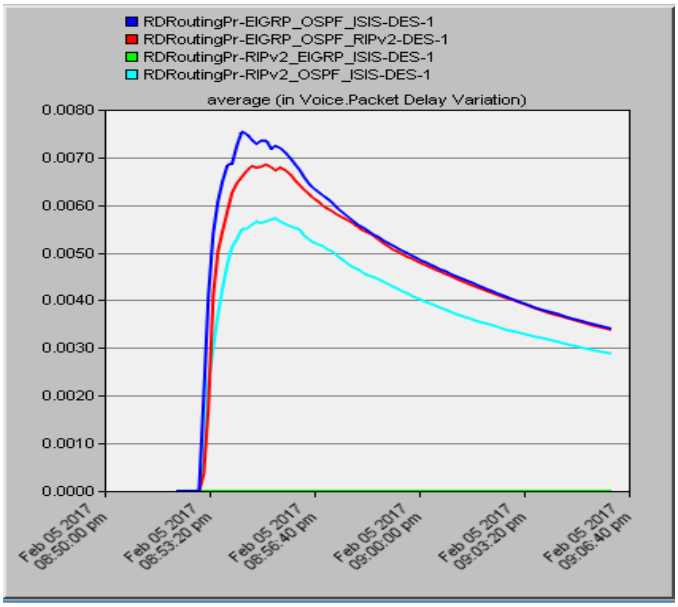


Fig. 14. The Voice Packet Delay Variation in the Four Scenarios

F. Voice Packet End to End Delay

The Fig. 15 shows the RIPv2_EIGRP_ISIS scenario that has less end to end delay time as compared with the three other scenarios, where end to end delay is defined as the time taken for a packet to be sent via a network from sender to receiver.

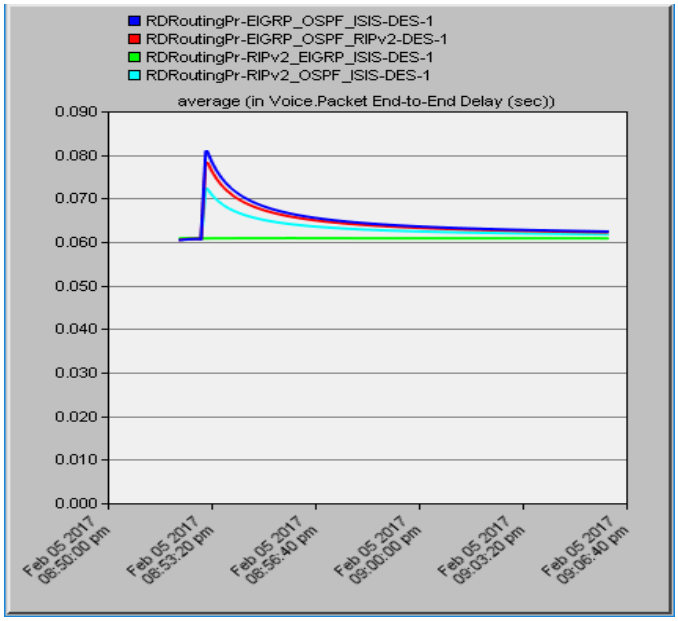


Fig. 15. The Voice Packet End to End Delay in the Four Scenarios

G. Video Conferencing Packet Delay Variation

The RIPv2_EIGRP_ISIS scenario has less delay variation as compared with the three other scenarios, therefore, it is the best one in terms of video conferencing service. On the other hand, the route redistribution among three protocols EIGRP, OSPF, and RIPv2 in the EIGRP_OSPF_RIPv2 scenario that has the worst result in the simulation as shown in Fig. 16.

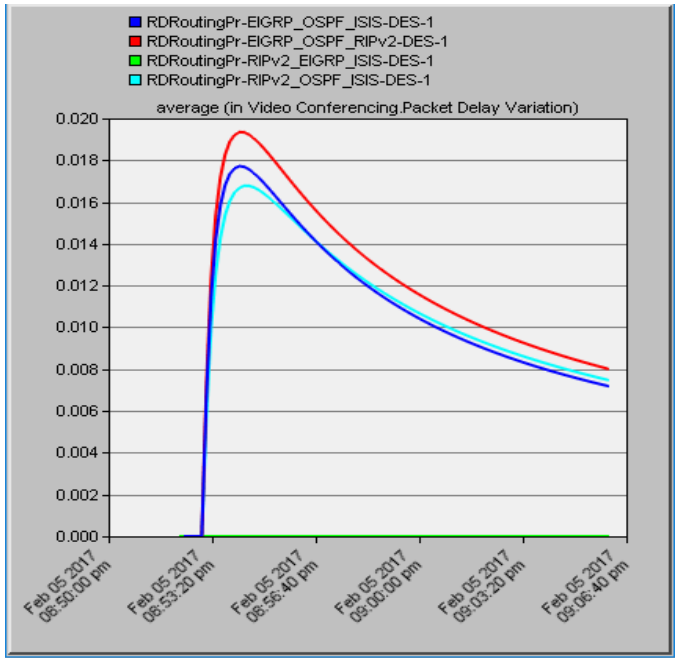


Fig. 16. The Video Conferencing Packet delay Variation in the Four Scenarios

H. Video Conferencing Packet End to End Delay

The video conferencing packet end to end delay is demonstrated in Fig. 17. According to this figure, the RIPv2_EIGRP_ISIS scenario, before of failure occurrence and after network recovery that it has the lowest value and stable compared with the other scenarios consequently it is the best in terms of real time application.

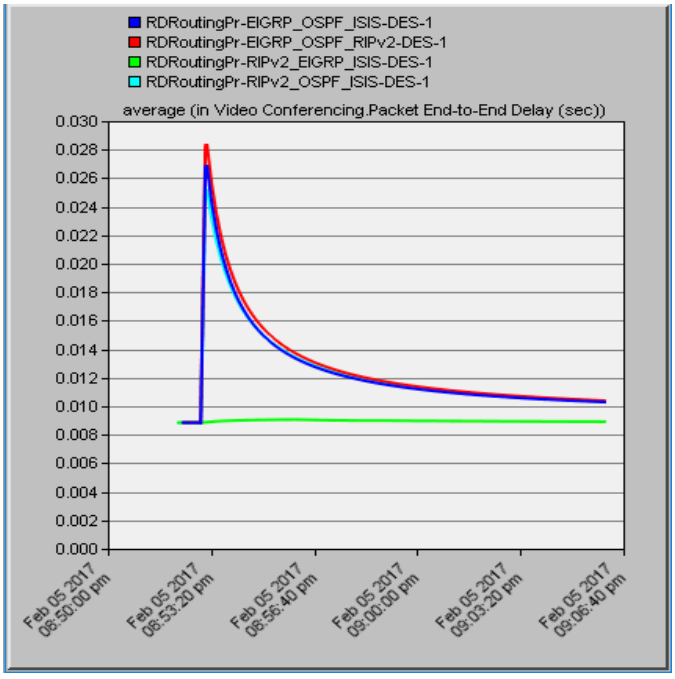


Fig. 17. The Video Conferencing Packet End to End Delay in the Four Scenarios

V. Conclusion and Future Work

The Large network infrastructure consists of multiple routing protocols in order to advertise different routing information, therefore, the network border routers should be configured in order to keep the network connected. This paper has analyzed the performance of the route redistribution among three different routing protocols. The four scenarios are created and configured on the same network from different dynamic routing protocols. The first scenario is configured from EIGRP, OSPF, and IS-IS, the route redistribution among EIGRP, OSPF, and RIPv2 that is configured in the second scenario, the third scenario is the combination of RIPv2, EIGRP, and IS-IS, and the last scenario is configured from RIPv2, OSPF, and IS-IS routing protocols. The simulation showed the third scenario RIPv2_EIGRP_ISIS that is the best in terms of real time application such as voice and video conferencing as compared with the other scenarios in this study. The RIPv2_EIGRP_ISIS scenario has the optimal path of the hops number and minimal value of convergence duration time. In contrast, the EIGRP_OSPF_ISIS scenario has fast response time in terms of using Email, database query, and web browsing services.

In the future work, the route redistribution among three different dynamic routing protocols in this paper which will be tested on the GNS3 emulator.

References

[1] F. Le, G. G. Xie and H. Zhang, "Understanding Route Redistribution," 2007 IEEE International Conference on Network Protocols, Beijing, 2007, pp. 81-92.

[2] Xin Sun, Sanjay G. Rao, and Geoffrey G. Xie. "Modeling complexity of enterprise routing design." In Proceedings of the 8th international conference on Emerging networking experiments and technologies, Nice, 2012, pp. 85-96.

[3] David A. Maltz, Geoffrey Xie, Jibin Zhan, Hui Zhang, Gísli Hjálmtýsson, and Albert Greenberg, "Routing design in operational networks: a look from the inside", In Proceedings of the 2004 conference on Applications, technologies, architectures, and protocols for computer communications, Portland, 2004, pp. 27-40.

[4] Cisco, Redistributing Routing Protocols, Mar, 2012.

[5] Cisco, OSPF Design Guide, Aug, 2005.

[6] Cisco, Intermediate System-to-Intermediate System Protocol.

[7] Mustafa Abdulkadhim, "Routing Protocols Convergence Activity and Protocols Related Traffic Simulation With It's Impact on the Network", IJCSET. International Journal of Computer Science Engineering and Technology, vol. 5, no. 3, pp. 40-43, March, 2015.

[8] Anibrika Bright Selorm Kodzo, Mustapha Adamu Mohammed, Ashigbi Franklin Degadzor and Dr. Michael Asante, "Routing Protocol (EIGRP) Over Open Shortest Path First (OSPF) Protocol with Opnet", IJACSA. International Journal of Advanced Computer Science and Applications, vol. 7, no. 5, pp.77-82 , 2016

[9] Lalu Zazuli Azhar Mardedi and Abidarin Rosidi, "Developing Computer Network Based on EIGRP Performance Comparison and OSPF", IJACSA. International Journal of Advanced Computer Science and Applications, vol. 6, no. 9, pp. 80-86, 2015.

[10] Richard John Whitfield and Shao Ying Zhu,"A Comparison of OSPFv3 and EIGRPv6 in a Small IPv6 Enterprise Network", IJACSA. International Journal of Advanced Computer Science and Applications, vol. 6, no. 1, pp.162-167 , 2015.

[11] G. K. Dey, M. M. Ahmed and K. T. Ahmmed, "Performance analysis and redistribution among RIPv2, EIGRP & OSPF Routing Protocol", International Conference on Computer and Information Engineering, Rajshahi, 2015, pp. 21-24.

[12] Haresh N. Patel and Prof.Rashmi Pandey, "Enhanced Analysis on Route Summarization and Route Redistribution with OSPF vs. EIGRP Protocols Using GNS-3 Simulation", IJCTA. Int.J.Computer Technology & Applications, vol. 5, no. 5, pp. 1682-1689, ,Sept-Oct, 2014.

[13] S. Farhangi, A. Rostami and S. Golmohammadi, "Performance Comparison of Mixed Protocols Based on EIGRP, IS-IS and OSPF for Real-time Applications", Middle-East Journal of Scientific Research, vol. 12 , no. 11, pp. 1502-1508, 2012.

[14] Syed Yasir Jalali, Sufyan Wani and Majid Derwesh, "Qualitative Analysis and Performance Evaluation of RIP, IGRP, OSPF and EGRP Using OPNET™", Advance in Electronic and Electric Engineering, vol. 4, no. 4, pp. 389-396, 2014.

[15] Asmaa Shaker Ashoor, "Performance Analysis Between Distance Vector Algorithm (DVA) & Link State Algorithm (LSA) For Routing Network", International Jiurnal of Scientific & Technology Research, vol. 4, no. 02, pp. 101-105, February, 2015.

[16] Martin Kuradusenge and Damien Hanyuwimfura, "Operation and Comparative Performance Analysis of Enhanced Interior Routing Protocol (EIGRP) over IPv4 and IPv6 Networks", International Journal of Advanced Research in Computer Science and Software Engineering, vol. 6, no. 7, pp.174-182 , July, 2016.

[17] Sukhkirandeep Kaur and Roohie Naaz Mir, "Performance Analysis of Interior Gateway Protocols", Advanced Research in Electrical and Electronic Engineering , vol. 1, no. 5, pp. 59-63, 2014.

[18] Kuwar Pratap Singh, P. K. Gupta and G. Singh, "Performance Evaluation of Enhanced Interior Gateway Routing Protocol in IPv6 Network", International Journal of Computer Applications ,vol. 70, No.5, pp. 42-47, May, 2013.

[19] M. Pavani, M. Sri Lakshmi and Dr. S. Prem Kumar, "A Review on the Dynamic Routing Protocols in TCP/IP", The International Journal Of Science & Technoledge, vol. 2, no. 5, pp.227-234, May, 2014

[20] P.Priyadhivya and S.Vanitha, "PERFORMANCE ANALYSIS OF INTERIOR GATEWAY PROTOCOLS", International Journal of Advanced Technology in Engineering and Science, vol. 3, No. 2, pp. 693-700 ,February, 2015.

[21] Shah.A and Waqas J.Rana, "Performance Analysis of RIP and OSPF in Network Using OPNET", IJCSI. International Journal of Computer Science Issues, vol. 10, Issue 6, No 2, pp. 256-265, November, 2013.

[22] S. Vissicchio, L. Vanbever, L. Cittadini, G. G. Xie and O. Bonaventure, "Safe routing reconfigurations with route redistribution", IEEE INFOCOM 2014 - IEEE Conference on Computer Communications, Toronto, 2014, pp. 199-207..

[23] IKram Ud Din, Saeed Mahfooz and Muhammad Adnan, "Analysis of the Routing Protocols in Real Time Transmission: A Comparative Study", Global Journal of Computer Science and Technology, vol. 10, no. 5, pp. 18-22, July, 2010.

[24] Jagmeet Kaur and Er. Prabhdeep Singh, "COMPARATIVE STUDY OF OSPFV3, IS-IS AND OSPFV3_IS-IS PROTOCOLS USING OPNET", IJARCET. International Journal of Advanced Research in Computer Engineering & Technology, vol. 3, no. 8, pp. 2656-2662, August, 2014.

[25] ShewayeSirika and SmitaMahajine, "Performance Evaluation of Dynamic Routing Protocols for Real time application", IJETT. International Journal of Engineering Trends and Technology,vol. 32 no. 7, pp. 328-337 ,February, 2016.

[26] Komal Gehlot and N.C. Barwar, "Performance Evaluation of EIGRP and OSPF Routing Protocols in Real Time Applications", IJETTCS. International Journal of Emerging Trends & Technology in Computer Science, vol. 3, no. 1, pp. 137-143, January – February, 2014.

[27] Y. N. Krishnan and G. Shobha, "Performance analysis of OSPF and EIGRP routing protocols for greener internetworking", ICGHPC. International Conference on Green High Performance Computing, Nagercoil, 2013, pp. 1-4..

[28] Moh'd Rasoul Ahmad Al-Hadidi, Mohammed Yousef Al-Gawagzeh, Nayel Al-Zubi,Bayan Al-Saaidah and Mohammed Alweshah, “Performance Analysis of EIGRP via OSPF Based on OPNET and GNS3”, Research Journal of Applied Sciences, Engineering and Technology, vol. 8, no. 8, pp. 989-994, 2014.

[29] Jaswinder Kumar, Samiksha, Susil Bhagat and Karanjit Kaur," Route Redistribution between EIGRP and OSPF Routing Protocol in Computer Network Using GNS3”, IJCNWMC. International Journal of Computer Networking, Wireless and Mobile Communications, vol. 5, no. 1, pp. 27-34, Feb, 2015.

Automatic Image Annotation based on Dense Weighted Regional Graph

Masoumeh Boorjandi
Department of Computer,
Aliabad Katoul Branch,
Islamic Azad University,
Aliabad Katoul, Iran

Zahra Rahmani Ghobadi
Faculty of Computer,
Ramsar Branch,
Islamic Azad University,
Ramsar, Iran

Hassan Rashidi
[3]Faculty of Mathematics and
Computer Science,
Allameh Tabataba'i University,
Tehran

***Abstract*—Automatic image annotation refers to create text labels in accordance with images' context automatically. Although, numerous studies have been conducted in this area for the past decade, existence of multiple labels and semantic gap between these labels and visual low-level features reduced its performance accuracy. In this paper, we suggested an annotation method, based on dense weighted regional graph. In this method, clustering areas was done by forming a dense regional graph of area classification based on strong fuzzy feature vector in images with great precision, as by weighting edges in the graph, less important areas are removed over time and thus semantic gap between low-level features of image and human interpretation of high-level concepts reduces much more. To evaluate the proposed method, COREL database, with 5,000 samples have been used. The results of the images in this database, show acceptable performance of the proposed method in comparison to other methods.**

Keywords*—automatic annotation; dense weighted regional graph; segmentation; feature vector

I. Introduction

Due to the growing use of digital technologies, image data generated and stored every day in large numbers, and using this data as text data has become commonplace. Hence the need to search for video data according to different demands increased. One of the traditional methods for image retrieval, is content based image retrieval [1,2,3]. But these systems are not able to understand the meaning. Also in this system, user must express your wishes with the visual properties of image expression, which in turn is difficult for users [1]. This is a formidable challenge in content-based image retrieval, called semantic gap. Semantic gap, the gap between low-level visual content of the image and human interpretation of it, is a high-level concept. [2]Methods including automatic image annotation that have been proposed in recent decades to reduce semantic gap.[4] computer in automatic image annotation is used to describe words that are suitable for producing images. In this case, to recover the image of a set of annotated images, using text request is also possible. Using text of the question is far easier than using a sample image or characteristics of the image. [1] Great deal of research has been done in the field of image annotation that can be grouped into three models: probabilistic models, model-based on categories and models based on the nearest neighborhood. [5] Most probabilistic models [6,7,8,9] joint probability estimate on the image content and keywords. Model-based categories [10,11], the image annotation to be an issue with the supervisor behave category. Models based on the nearest neighborhood, is one of the oldest, simplest and yet most efficient models in the category, the model is k-nearest neighbor. This model is growing, especially in the absence of training samples, efficiently. One of the methods in this area includes paper [12] cited. But in all of the presented methods in the field of automatic annotation there are two challenging problems: First, annotation techniques available generally are a feature of regional or global brand used to describe alone. But national and regional features focused on different aspects of an image complement each other, so combining them together to describe the images will be beneficial. Second, in all delivery methods based on characteristics of the area, only the direct distribution areas as areas used for image annotation via image segmentation based on region or the nature of objects are obtained, and the relationship between areas does not be paid attention. If given the link between areas, each of which represents a word or concept, we can help improve final margin words. in order to fix the problem on the basis of a weighted graph, in this article we offer an area dense. So that the proposed method uses the theory of Rough and Fuzzy feature vector for regions resulting from segmentation to effectively classified images and graphs do make up a dense area of both national and regional characteristics for use together to annotate and significantly enhanced accuracy. It also uses weighting the edges between vertices in the graph area proposed for communication between areas of images detected by the system.

This article is organized as in Section 2, the image automatic annotation method based on weighted graph describes dense in an area, in section 3, the proposed algorithm simulation results and comparison with other methods in this area are provided. In the fourth overall conclusion of the article offered.

II. Image Annotation based on Dense Weighted Regional Graph

In this section of the paper, the proposed method for annotating images is explained.

The construction of an area of dense graph is as follow:

1) First collect a free enlarge dataset of annotated images, and classify values into different categories according to their annotation keywords.

2) Pictures related to any particular class are divided with accurate and effective method to pieces Rough separate charges.

3) For Category parts resulting from segmentation of images related to any particular group, first we used the method of k-means, so that same parts are classified in the same group. But because of some shortcomings in this method and more accurate grouping, again we will classify low-density lightweight piece set of groups of pictures (which are included large amounts of each class of images) by considering fuzzy feature vector associated with them, and compare it with the original image feature vector corresponding to the category of other classes, categories reclassification of the charges.

4) At this point we create dense an area graph so that we put fences and high density near the center of the image in a category and get the fuzzy feature vector associated with the label of their respective classes, images are annotated.

For groups with low density and outlying parts by determining the fuzzy feature vector and determine its similarity to other video groups from other classes, in Group of parts that are most similar to them categorized and tagged with annotations are related to their respective floors.

5) Then each vertex of the graph, which represents dense cluster of image segments with high similarity which are connected to each other by weighted edge with other categories that include other pieces from the collection of images related to each specific circuit. We are considering the joint probability of more accurate values for weight gain so that by taking pictures of each class, the weaker groups object to be removed from any particular class.

After creating an area of dense graph of images related to the training data, in order to annotate new image, first we will segmentation and then due to dense parts closer to the center of the image and fuzzy characterization of their right to obtain the original image, and in the following, based on Weighted area of dense graph, Image annotation associated by taking in account similarity of other image parts which are connected with more weighted vertices of main group of images.

In the graph area of dense, lightweight piece groups are annotated with names of collected images. Thus, the number of categories in training data should be large enough to cover the meanings of these groups and pieces of the image.

A. Image Segmentation in every particular class

Maximum image components can be extracted through various ways such as image segmentation [13], dense samples [14], and recognize a specified area [15] and etc. But most of these methods are very expensive in terms of time and calculation. Among the proposed methods, Rough set theory has the ability to cluster profitably data that comes from image analysis [16] as efficiently identify the edge that is one of the effective methods of segmentation, convergence time So we proposed this theory we use for segmentation of images. Rough set more detailed description about the image segmentation in the paper [17] can see. After extracting the piece of the picture, we produce dense piece band, we take action the k-means clustering pieces by the charges. However, in clustering k-means, the number of clusters manually set-up, and to produce pieces that are freely distributed is such as parts noise or background unstable and cluttered, so the results so not ideal. So after the initial clustering for producing dense, we have tried to use the feature vector effective, low-density parts with high density split into smaller pieces and fit them with similar criteria in groups with high similarity of our Categories. In this way text communication between groups are clearly to be identified, so that the results show high accuracy of the proposed method in the categories of different parts of images.

B. Produce dense areas and annotate

Most of the annotations provided in part due to lack of identification with high density of low-resolution images. Since the identification of areas with low density, similar to other image areas are densely populated in the group, we have tried to use the feature vector effective, low-density parts with high density split into smaller pieces and fit them with similar standards in groups with high similarity of our markets. For this purpose, we used a mask with size 60 × 60 on low-density areas.

We use to identify groups of two characteristic color and edge video compression phase for more accurate the obtained fuzzy logic.

More details on how to determine vector features can be found in Articles [18] and [19] pictures. Since fuzzy feature vector includes three properties in each area of the image are color, location and edge color, therefore, the proposed fuzzy similarity measure using data from the three phase characteristics for more accurate similarity or dissimilarity to appoint different images.

The following formula shows the proposed fuzzy similarity measure:

$$sim(q,t) = (w_c \times (\frac{\sum_{j=1}^{10} \min(H_c^{q_i}(j), H_c^{t_i}(j))}{\min(\sum_{j=1}^{10} H_c^{q_i}(j), \sum_{j=1}^{10} H_c^{t_i}(j))} \times \sum_{j=1}^{10} (1 - (\sqrt{\frac{(\overline{x_{q_j}} - \overline{x_{t_j}})^2 + (\overline{y_{q_j}} - \overline{y_{t_j}})^2}{2}}))) + (\sum_{i=1}^{4} w_{ei} \times \frac{\sum_{j=1}^{4} \min(H_e^{q_i}(j), H_e^{t_i}(j))}{\min(\sum_{j=1}^{4} H_e^{q_i}(j), \sum_{j=1}^{4} H_e^{t_i}(j))}) \quad (1)$$

$H_c^{q_i}(j)$: J-th column of fuzzy color histogram, i-th class query image

$\overline{x_{q_j}}, \overline{y_{q_j}}$: Average Location j-th column of the query image color histogram

W_c: the importance of class background

$H_e^{q_i}(j)$: J-th bar of the histogram of the i-th edge to edge located

W_{ei}: the importance of the i-th to edge

As the numerical value for each component fuzzy weight in the literature [18] and [19] were determined as the following:

Very large = 1, large = 0.8, medium = 0.55, small = 0.3, very small = 0.1

This method to a large extent is resistant to the problems that the majority of edge detection methods are faced, such as sensitivity to noise and with thick lines. So low-density areas are divided into several regions with higher density and tagged with annotations are the most similar. In this way, for different images in the database area of dense graph is created.

C. Annotation

After you remove the piece of background object slice group dense pieces, we attempted to annotate them. To do this, we first collect band name as a label. Since a large number of large-scale collections are on the floor, we assume that most of groups extracted can be annotated with the name of the group. We do annotation groups based on two criteria:

1) Visually similar groups must be annotated with similar tags.

2) Groups that are distinct and belong to a particular class, are likely annotate with this class. This idea is illustrated in Figure1.

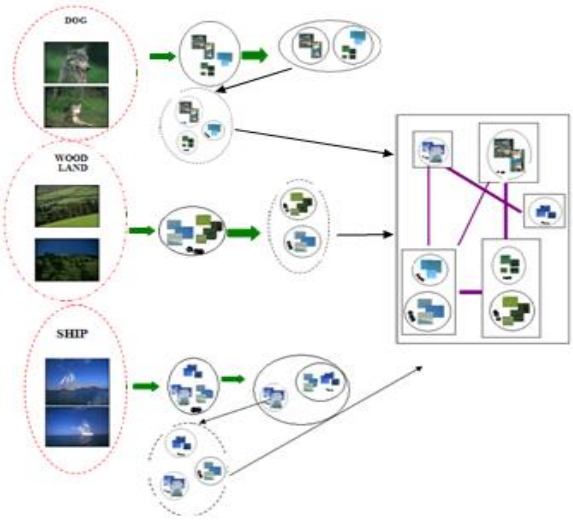


Fig. 1. Image of the proposed method to annotate Group

For example, in Figure 1 piece for the dog, after compression zones were discovered containing "dog", "woods", "sky" and so on, that piece band dogs due to high density and proximity to the center of the image the dog is the label class that is annotated. The following groups are fences "sky" and "woodlands" by comparing their characteristics with the vector of the main themes in classes other picture, as a new group piece which most closely resembles groups may be categorized with them And then the edge between the two vertices in the graph occurred that indicate the presence of similar images is common in these group.

For example in Figure 1 edge between vertices of dogs and woodlands parts thicker than the edge between the vertices of the parts is dog heaven and this occurred due to high joint between the two parts of the head. To determine the edge weight between vertices, following formula is used:

$$W=k/n \qquad (2)$$

W is Weight between two vertices in the graph, an area corresponding to each of the respective condensing And K, is the number of subscribers took place labels on two related helm annotation of images that class and n, is the number of images of respective class.

So common they both took the helm that is more likely that more weight be given indicator is thicker than the edges.

D. Experimental results and evaluation

The system uses a data set of 5,000 images from Corel database for evaluating the proposed approach. This collection of images includes 50 main groups that each group consisted of 100 images. Each image is accompanied by a set text labels include 1-5 word. Overall, there are 360 the meaning of the set of images. To perform the test, the complex image is divided into two parts: an image training set of 4500 images and a test set of 500 images. Figure 2 shows a few examples of annotated images from the database of our proposed method for Corel database. Tags are attributed to each image so that false labels marked with red color.

Based on images, it is clear that some of the labels we determined by the proposed approach them, are correct, but Due to poor images in the database annotation tags have been considered as inappropriate.

For example in the third image (eagle), the sea is also part of the picture shown. But the label is not determined by man. That's why this tag as a tag identified by the proposed method is considered to be false

image						
Correct lable	Cat,tiger,grass	Flower,leaf,grass,clouds ky	Train,bus,grass,tree,sky	Eagle,bird,sky,cloud,mountain	Horse,tree,grass,sky,fance	Man,sea,beach,sky
Proposed model	Tiger,grass,tree	Flower,grass,leaf,tree,sea	Train,sky,grass	Eagle,sea,sky,tree	Horse,bush,fance	Man,sea,cloud,beach,sky

Fig. 2. The words predicted by the proposed method for multi-image annotation database of Corel

Then for more precise evaluation of effectiveness of the proposed model, the average precision, recall and F1 parameters have been calculated per word.

$$\text{Precision } (v_i) = N_C / N_S \quad (3)$$

$$\text{Recall } (v_i) = N_C / N_R \quad (4)$$

$$F_1 = 2 \times \text{Precision} \times \text{Recall} / \text{precision} + \text{Recall} \quad (5)$$

Accuracy, is the ratio of N_C, the ratio of the number of images in the test phase to the N_S, the number of images than in the test phase. Calling the ratio of N_C, the ratio of the number of images in the test phase N_R, number of images in the database for each v i is the word.

Annotations to assess the quality, accuracy and calling for each and every word database (v i) are calculated.

Table 1, shows mean precision, recall and F_1 for the proposed methodology and the appropriate annotation recently presented three methods (IAGA -2014 [20], Feature fusion and semantic similarity-2014 [5], MLRank -2013 [21]). More areas have higher density than other areas in each image by weighting the edges of the graph. Less important areas are removed so that it causes a system closer to annotate people, in our proposed method is compared to other methods.

TABLE. I. COMPARISON OF PRECISION, RECALL AND F 1OF THE PROPOSED METHOD AND OTHER METHODS

F1	recalling	accuracy	Model name
0.34	0.31	0.34	IAGA-2014
0.31	0.32	0.3	Feature fusion and semantic similarity-2014
0.29	0.28	0.3	MLRank-2013
0.36	0.33	0.39	Proposed method

The results of this graph shows that, using accurate segmentation according to the theory of Rough and strong feature vectors that lead to the formation of dense graph and highlight an area densely populated areas in each image, is the more precise identification of concepts such as stairs, flowers and lawns aircraft per cent lower compared to IAGA that have been identified.

In fact, one of the main problems in annotation methods including IAGA non-designated areas in each image is important, we provided an area graph density to accurately classify areas according to each class video By weight of feature vectors appropriate and accurate to solve them.

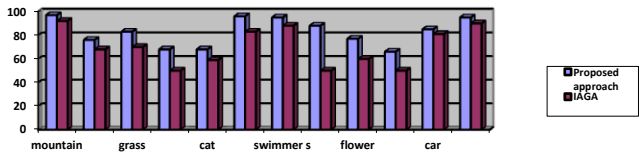


Fig. 3. Percentage identify 12 words sample to the suggested model and model IAGA

Figure 4, the results retrieved for a query on the database shows Corel. For any query shows five images with the highest similarity. According to the results in the figure below, we conclude that the graph based image retrieval dense area for suggestions, will guide us to recover the database with accuracy and efficiency.

Therefore, in the proposed method, we were able to accurately and more efficient than the methods suggested in the annotation to achieve by Categorizing more accurate images with different areas of dense graph and also remove less important areas in each class by weighting the edges.

Fig. 4. Results semantic image retrieval database corel. Each row of five top result query semantic meaning in accordance with the left-most column shows

III. CONCLUSION

In this paper, a graph-based method for automatic image annotation densely an area is provided.

In most of the methods presented in the context annotations there are two basic challenging problems including Lack of integration of national and regional characteristics for each of the images and lack of attention to the relationship between different areas in Pictures.

In this article we formed an area graph, the relationship between different areas in the image are considered and weighted the edges in the graph, and done compression of areas so that Prominent areas on each floor image in low-density areas considered less important and have to be removed. Also by Using strong fuzzy feature vectors based on color and edge features for the considered areas we have done Aggregation of national and regional action features lightweight and have improved Annotations practice considerably.

At the end we have implemented the proposed approach on Corel database. The results provided on the database show acceptable performance of the proposed method compared with other methods in this field.

ACKNOWLEDGEMENT

This work has been supported by research contract of the Islamic Azad University of Aliabad Katoul branch.

REFERENCES

[1] T. Pavlidis , Limitations of Content-based Image Retrieval, invited plenary talk at the 19 th International Conference on Pattern Recognition, Tampa, Florida, Dec. 8-11, 2008.

[2] R. Datta, D. Joshi, J. Li, J.Z. Wang, Image Retrieval: Ideas, Influences, and Trends of The New Age, ACM Transactions on Computing Surveys, Vol. 40, No. 2, April 2008.

[3] A.W.M. Smeulders, M. Worring, S. Santini, A. Gupta, R. Jain, Content-based image retrieval at the end of the early years, ieee transactions on pattern analysis and machine intelligence, vol. 22, no. 12, december 2000,p. 1349-1380.

[4] Jiayu Tang ,.Automatic Image Annotation and Object Detection, A thesis for the degree of Doctor of philosophy , University Of Southampton, 2008.

[5] X. Zhang and C. Liu, "Image annotation based on feature fusion and semantic similarity," in Neurocomputing - Volume 149, Part C, Elsevier – ScienceDirect, 3 February 2015, Pages 1658–1671.

[6] P. Duygulu, K. Barnard, J. F. de Freitas, and D. A. Forsyth, Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary, in Computer Vision—ECCV 2002, Springer, pp. 97–112, 2002.

[7] J. Jeon, V. Lavrenko, and R. Manmatha, Automatic image annotation and retrieval using cross-media relevance models, in Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, pp. 119–126, 2003.

[8] V. Lavrenko, R. Manmatha, and J. Jeon, A model for learning the semantics of pictures ,in Advances in neural information processing systems, 2003.

[9] S. L. Feng, R. Manmatha, and V. Lavrenko, Multiple bernoulli relevance models for image and video annotation, in Computer Vision and Pattern Recognition, Proceedings of the 2004 IEEE Computer Society Conference on, vol. 2, pp. II–1002, 2004.

[10] J. Li, J. Wang, Automatic linguistic indexing of pictures by a statistical modeling approach, IEEE Transactions on Pattern analysis and Machine Intelligence, Volume: 25 Issue: 9,2003,p. 1075-1088.

[11] E. Chang, K. Goh, G. Sychay, G. Wu, CBSA: CBSA: content-based soft annotation for multimodal image retrieval using Bayes point machines, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2003,pp. 26--38.

[12] J. Verbeek, M. Guillaumin, T. Mensink, and C. Schmid, Image annotation with tagprop on the MIRFLICKR set, in Proceedings of the international conference on Multimedia information retrieval, pp. 537–546, 2010.

[13] P. Felzenszwalb, D. Huttenlocher, Efficient graph-based image segmentation, International Journal of Computer Vision, Volume 59, Issue 2,2004, pp 167–181 .

[14] S. Lazebnik, C. Schmid, J. Ponce, Beyond bags of features: spatial pyramidmatching for recognizing natural scene categories, IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2006.

[15] J. Matas, O. Chum, M. Urba, T. Pajdla, Robust wide baseline stereo frommaximally stable extremal regions, In British Machine Vision Conference ,2002.

[16] B. Frey, D. Dueck, Clustering by passing messages between data points, journal of Science, Vol. 315, Issue 5814, pp. 972-976,2007.

[17] T. Chua, J. Tang, R. Hong, H. Li, Z. Luo, Y. Zheng, NUS-WIDE: a real-world web image database from National University of Singapore, Conference: Proceedings of the 8th ACM International Conference on Image and Video Retrieval,2009.

[18] M. Bourjandi, Image Retrieval Based on Eten Fuzzy Color Histogram, International Journal of Mathematics and Computer Sciences (IJMCS), Vol.23 November 2013.

[19] M. Bourjandi , A.M.Eftekhari, Image Retrieval Based on Fuzzy Weight Vector in Competative Fuzzy Edge, 18th Iranian International Power System conference ,2010.

[20] S. Bahrami, M. Saniee Abadeh, "Automatic image annotation using an evolutionary algorithm (IAGA), In 7th International Symposium on Telecommunications (IST), pp. 320-325. IEEE, 2014.

[21] Z. Li, J. Liu, C. Xu, and H. Lu, MLRank: Multi-correlation Learning to Rank for image annotation, journal of Pattern Recognition., vol. 46, no. 10, pp. 2700–2710, 2013.

A New Comment on Reinforcement of Testing Criteria

Monika Singh
College of Engineering and Technology,
Mody University of Science and Technology,
Lakshmangarh, Rajasthan, India

Vinod Kumar Jain
College of Engineering and Technology,
Mody University of Science and Technology,
Lakshmangarh, Rajasthan, India

***Abstract*—This paper presents the formal aspects of testing criteria for Safety Critical Systems. A brief review of testing strategies i.e. white box and black box is given along with their various criteria's. Z Notation; a formal specification language is used to sever the purpose of formalization. Initially, the schemas are formed for Statement Coverage (SC), Decision coverage (DC), Path Coverage (PC), Equivalence Partition Class (EPC), Boundary Value Analysis (BV) and Cause & Effect (C&F). The completeness and correctness of test schema are enriched by verifying these with Z/EVES; a Theorem Prover tool for Z specification.**

Keywords—Formal Methods; Safety Critical System; Z Notation; Schema

I. Introduction

Testing [1] plays an important role for checking the correctness of system implementations. To test system, test cases are formed and system behavior has been observed during execution. Based on test execution, the decision is made for the correctly functioning of the system. However, the criterion for the correctness of test cases has been specified in the system specification. A specification prescribes "What" part of the system i.e. the function that a system supposes to do and accordingly forms the foundation for testing criteria. As system specifications are documented in natural language (informal), which is generally incomplete and ambiguous in nature, due to this many problems may occur in testing processes such as incompleteness, ambiguous and inconsistency in test specifications. With an unclear specification, it is next to impossible to predict how the implemented system will behave; consequently testing will be difficult as it is not clear what to test. This become more severs specifically in case of Safety Critical System [2]. An ambiguous system specification which further forms the root for test specification may raise many problems such as misinterpretation and therefore needs explanations of specification's purpose. This requires rework of the system specification during the testing phase of software development. The rework process takes too much time, money and efforts which ultimately delay the process of deployment of system. Therefore, there is an utter need of usage of the formal model [3] for testing criteria of Safety Critical Systems [4] for test case's completeness and correctness. Formal methods are equipped with rich mathematical axioms and tool support. This rich tool support will help further verification of test specification in automated environment. In this paper, the purpose of formalization has been accomplished by Z Notation [5] and simulation has been done with Z/EVES [6]: an automated Theorem Prover.

Formal methods: Formal methods [3] are the methods which use mathematical techniques as their foundation pillars and are used to develop the software systems. They can be applied at any phase of software development process, but highly recommended to apply in early phases. By using formal methods, one can reduce the chances of ambiguities and incompleteness in requirements documents, design specification and the test case specification. There is a range of formal specification languages available to design the software system such as Z notation [5], B-methods [7], VDM [8] etc which are further verified by Theorem Prover [9] and Model Checker [9]. Broadly, formal methods are categorized into two groups:

a) Model based Formal methods: In this group, the formal specifications are consisting of mathematical structures such as relations, functions, sets and sequences to design software system model. The members of this group are: Z Notation [5], VDM [8], B-Methods [7], Petri net [10], Communicating Sequential processes (CSP) [11].

b) Property oriented formal methods: Property oriented formal methods, on the other hand, the specifications of system are defined in terms of its properties, generally in form of axioms which satisfied by the system. For example, OBJ, LOTOS [12], Larch lies in this group.

In this paper, we use Z notation to write done the test documents which is further analyzed by Z/EVES Theorem Prover tool. Rest of the paper is organized as follow: Section 2 represents the methodology and research components. Section 3 presents the Formal aspect of testing strategies for Safety Critical Systems. Section 4 advocates the simulation results and discussions. At last, the conclusion is given base on section 4 analysis in Section 5.

II. Methodology and Research Components

Initially the schemas of testing criteria i.e. SC, DC, PC, EPC, BV and C&F are formed by using Z Notation. Once the schemas are formed, they are checked for their completeness and correctness using Z/EVES; automated Theorem Prover tool for Z specification. If errors occur, corrections are made in respective schema and again execute on Z/EVES. This process is repeated until error free schemas are come as an output. Figure 1 presents the formal model of testing strategies which composed of following research components:

A. White Box testing

White box testing [13] is more concerned about the implementation details such as: programming style, control methods: statements coverage, decision coverage, condition coverage etc. It is also known as structural testing. It emphasizes on internal structure of software artifact. The internal structure mainly tested by using the following scenarios:

- Statement coverage: Test cases are executed in such a way that all statements have been covered once.
- Branch/decision coverage: Test cases are executed in such a way that both if-branch and else –branch covered.
- Path coverage: Test cases are executed in such a way that each possible path has been executed once.

B. Black Box Testing

Black box testing [14] focuses on functional/ behavioral testing of system without peeking into internal structure of system. It is also known as functional testing. It can be done by following ways:

- Equivalence partition classes: The input set is partitioned into equivalence classes and a single test case is executed for each class. The single test case is valid for all the elements of a given class. However, the classes chosen should be disjoint to avoid redundancy.
- Boundary value Analysis: In boundary value analysis rather than taking input from the partition classes, test cases are executed for boundary value points or near the boundary of partition classes.
- Cause & Effect Graph: In cause and effect (CF) graph, the combinations of inputs are analyzed. The cause is a representation of inputs and effect is a symbol of resultant output. Boolean graphs are used to link various causes and their respective effects.

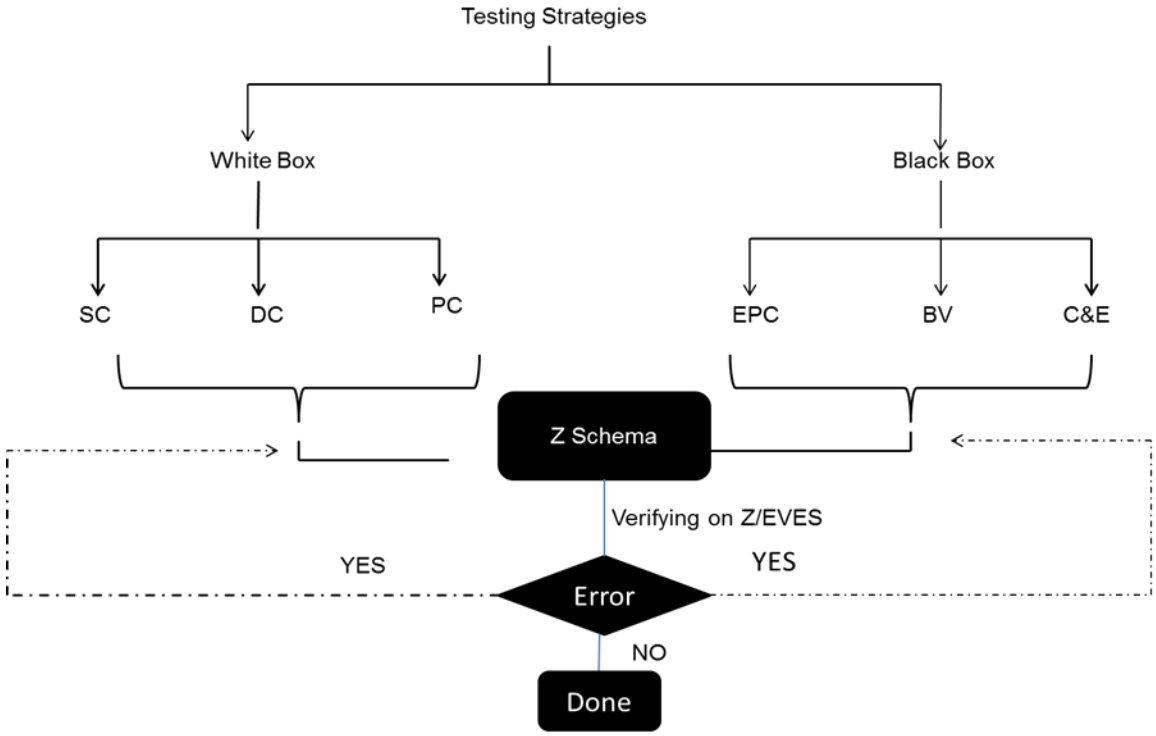


Fig. 1. Formal model of Testing Strategies

C. Z Schema

Schema is the notion used to structure the specification written in Z notation. It's composed of three parts: schema name, variable and constraints.

The generic structure of schema which showed in figure 2 consists of three parts as:

- Schema Name
- Variables declaration
- Constraints

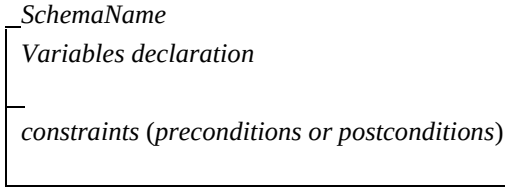


Fig. 2. Basic Schema structure

D. Z/EVES

Z/EVES toolset is an interactive tool for composing, checking, and analyzing Z specifications. It is based on the

EVES system, and uses its proof checker to carry out its proof steps. The language accepted by Z/EVES is a LATEX markup form [19]. This toolset helps in the analysis of Z specifications in several ways: (1) syntax and type checking, (2) schema expansion, (3) precondition calculation, (4) domain checking, (5) and general theorem proving [7]. The model checker of the Z/EVES is considered as user friendly and simple, especially when compared with other related tools such as Isabelle-HOL or Proof Power-Z. It could also prove its merit and popularity; due to its power in proving the specifications of critical systems written using the Z notation.

III. Formal Aspect of Testing Strategies

This section composed of two parts: Formal transformation of White box testing and Formal transformation of Black box testing.

A. Formal transformation of White box testing

White box testing focuses on internal structure of software artifact. One of the ways to test internal structure is to use either of following scenarios: Statement coverage, decision coverage or path coverage (Figure 1). However, various definitions of these scenarios may raise ambiguity. One possible solution is the elaboration of formalized definition of testing criteria by using rigorous mathematics such as set theory graph theory, predicates logics etc. In this paper, Z notation (Formal Specification Language) has been used to serve the purpose.

To check the completeness and correctness of above mention scenarios, mathematical structure i.e. Z –schema has been used. For any testing criteria, the two basic sets are required i.e.

[INPUT, STATEMENT]

Where INPUT is the set of all possible values of input variable and STATEMENT is the set of all program statements. Since in Statement coverage, every statement in the program has been executed at least once, therefore we define a function *path* from INPUT to STATEMENT as

Path i: INPUT →STATEMENT

Along with path function, we need to define two other set i.e. BOOL, INPUT-PART and COND as follow:

BOOL= {0, 1}

Which are respective values of executed conditions (as 1) and non-executed conditions (as 0)

INPUT-PART= P INPUT \ {INPUT}

i.e. non-empty set of input variables which yet not executed. Moreover, INPUT-PART is a subset of INPUT.

COND is a non-empty set which contain values by mapping an input i ∈ INPUT to true condition (as 1) and false condition (as 0).

COND== INPUT → BOOL

Now the formal definition of Statement coverage is given by using Z schema as:

SC

$$decinput \,!: \mathbb{P}_1 INPUT$$
$$decst \,?: STATEMENT$$
$$decinput\ 0, decinput\ 1: INPUT\text{-}PART$$

$$decinput = \{i: INPUT \mid decst \in path\ i\}$$
$$\langle decinput\ 0, decinput\ 1\rangle\ partitions\ decinput$$
$$Dom\ value = decinput$$

The constraints are: (i) the domain all values should be the set of input; (ii) The input values partition the set of input and (iii) For each i, the function path maps the input to respective statement. Therefore based on this, test_ data has been built which satisfy or not satisfy testing criteria.

Now the Decision Coverage (DC) is formally defined as:

DC

$$\Delta SC$$

$$\forall\ d{:}dec \bullet (test\text{-}data \cap d \bullet\ decinput\ 0 \neq \emptyset) \wedge$$
$$(test\text{-}data \cap d \bullet\ decinput\ 1) \neq \emptyset$$

The constraints of DC schema are defined as: if there is if-else condition, both of the decision will execute which consequently satisfy the definition of decision coverage. However, there would be change in statement coverage schema which has been shown by ΔSC.

The next schema is Path Coverage (PC)

PC

$$\Delta SC$$
$$\Delta DC$$

$$\forall\ i, j \in \mathbb{N}, (path\ i \cap path\ j = \emptyset) \wedge$$
$$(path\ i \cup path\ j = test\text{-}data)$$

The constraints of DC schema are: (i) all the possible paths are covered at least once and if and two paths are identical.

B. Formal Transformation of Black box testing

The three black box testing criteria which are considered here are:

- Equivalence partitions class (EPC)
- Boundary Value Analysis (BV)
- Cause & Effect (C&E)

As mentioned in section 2 (b), the test data is partitioned into equal classes and for each class only one value from test data is tested. Therefore for schema, two basic sets are:

[TEST_DATA, CLASS]

Now the schema of EPC is as follow:

EPC

tstdat: *TEST_DATA*
cls: *CLASS*

$\forall\ i, j \in \mathbb{N}, tst1, tst\ 2 \in TEST\_DATA \land$
$tst\ 1 \cap tst\ 2 = \emptyset \land$
$\forall i \in \mathbb{N}, cls1, cls2 \in CLASS \land$
$cls1 \cap cls2 = \emptyset \land$
$\cup cls\ i = CLASS$

The constraints are: All the partitions are disjoint and one test value should be chosen from one class. For Boundary value analysis, the boundary values of each class are tested. In other words, we need to test the five values for each class i.e. (i) Minimum (ii) Just above the minimum (iii) A nominal value (iv)Just below the maximum and (v) Maximum. Therefore the schema for boundary value is:

BV

Δ *EPC*

min, max, nominal $\in$ *TEST-DATA* $\land$
jst-abv-min, jst-blw-max $\in$ *TEST-DATA*

IV. Simulation and Discussion

Although the formal specification languages uses mathematics notation (in this paper Z notation has been used), yet chances of ambiguities are still there. Automated or semi-automated tool are used to check the Z specification. Z/EVES; a Theorem Prover tool for syntax, type checking and domain checking is used for checking the Z specification. The graphical interface of Z/EVES tool consists of two columns: Syntax and Proof. The columns with 'Y' value show that there is no error. Once the specification written, the file has been stored with extension ".zev". Figure 3 depicts the execution of Statement Coverage (SC) specification for syntax and type checking. It is cleared from figure 3 that both the columns have value 'Y' indicating that SC schema is free from syntax and domain errors. Similarly, Fig. 4, 5 represents the formal part of Path coverage (PC) and Equivalence Partition Class (EPC) respectively.

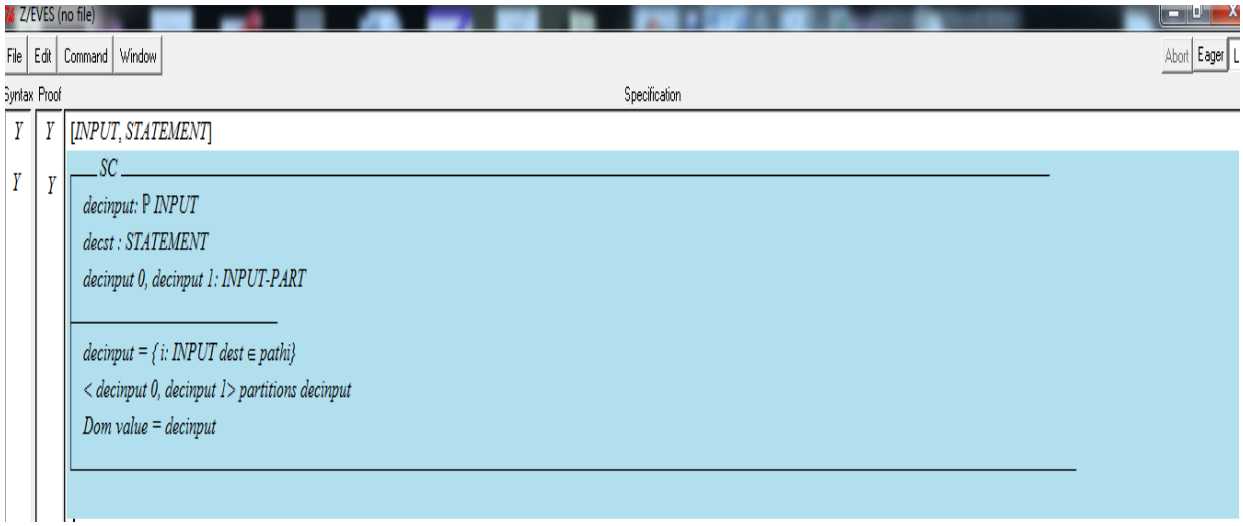


Fig. 3. Formalization of SC schema

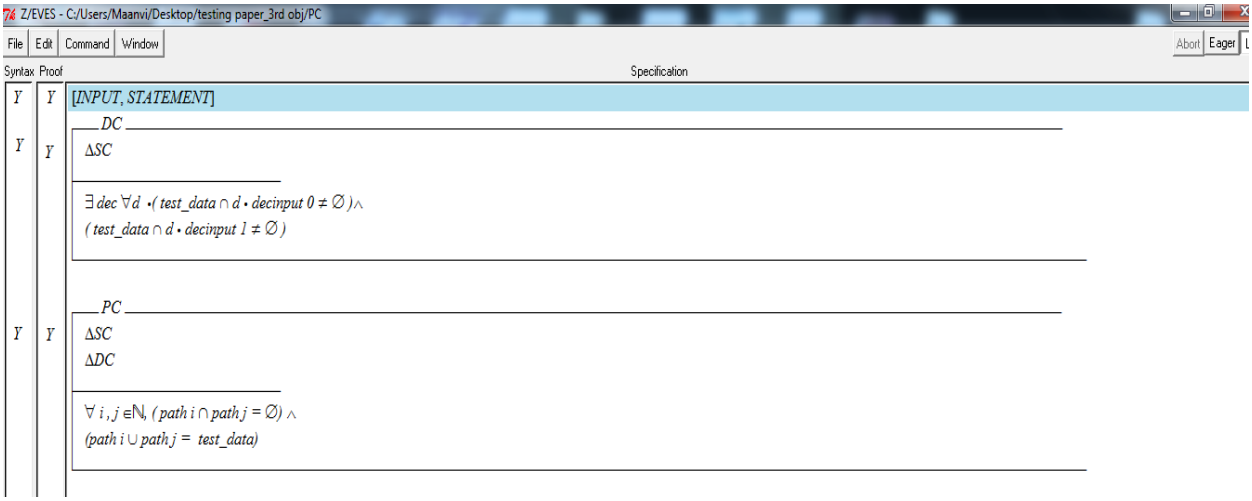


Fig. 4. Formalization of Path Coverage specification

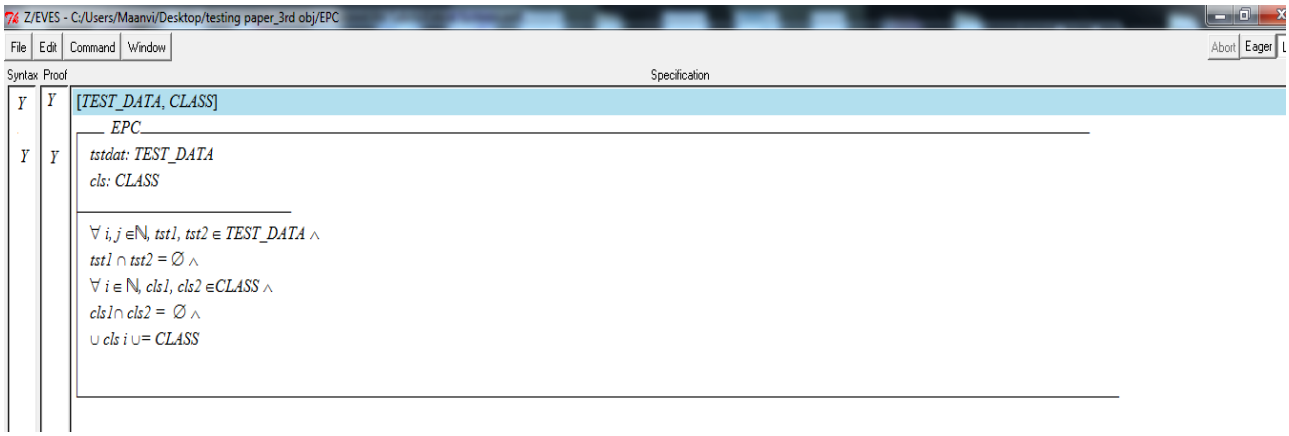


Fig. 5. Execution of EPC for syntax and Domain checking

V. CONCLUSION

The main idea of this article is formalization of testing criteria for safety critical systems. Software Testing Techniques are broadly partitioned into two groups i.e. white box testing and black box testing. For white box testing, three criteria's are used i.e. Statement Coverage (SC), condition Coverage (CC), Path Coverage(PC) and for Black Box testing, the criteria's which has been used are Boundary Value Analysis (BV), Equivalence Partition (EP)class, Cause & Effect (C&E). All these criteria used to figuring out the branch and loop structure using logical expressions in program. For fulfilling the definition of formalization, Z notation is used. Initially Z schemas are formed for each criterion's .i.e. SC, PC, BC, EP, BV and C&F. To check the correctness and completeness of schemas, Z/EVES tool is used further. The findings of Z/EVES are syntax checking, domain checking and type checking.

REFERENCES

[1] Glenford J. Myers. The Art of Software Testing, John Wiley & Sons, 2nd Edition, 2004.

[2] IPL. An Introduction to Safety Critical System, executive summary, 1997.

[3] Monin, Jean-Francois. Understanding Formal Methods, 2003, Springer.

[4] Jonathan Bowen. Safety Critical Systems, Formal Methods and Standards, 1992, Software Engineering Journal.

[5] J. Michael Spivey. The Z Notation: A Reference Manual, 2001, 2nd eds. Prentice Hall.

[6] Saaltink, M. The Z/EVES 2.0 User's Guide, Technical Report TR-99-5493-06a, ORA Canada, One Nicholas Street, Suite 1208 - Ottawa, Ontario K1N 7B7 - CANADA, 1999.

[7] S. Schneider. B Method- an Introduction Palgrave, Cornerstones of Computing series, 2001.

[8] C. B. Jones. Systematic Software Development using VDM, 1990, In Prentice Hall.

[9] Hasan Amjad. Combining model checking and theorem proving, 2004, technical report, University of Cambridge, computer Laboratory. UCAM-CL-TR-601, ISSN 1476-2986.

[10] Reisig, Wolfgang. Understanding Petri Nets: Modeling Techniques, Analysis Methods, Case Studies, 2013, 1st Eds. Springer-Verlag Berlin Heidelberg, ISBN 978-642-33277-7.

[11] C. A. R. Hoare. Communicating Sequential Processes, 1985, In Prentice Hall.

[12] Howard Bowman. A LOTOS based tutorial on formal methods for object-oriented distributed systems, New Generation Computing, Vol. 16, Issue 4, pp 343-372.

[13] Glenford J. Myers, Corey Sandler, Tom Badgett. The Art of Software Testing, 2011 3rd Eds. Wiley.

[14] Paul C. Jorgensen. Software Testing: A Craftsman's Approach, 2013, 4th Eds. Auerbach Publications.

Appraising Research Direction & Effectiveness of Existing Clustering Algorithm for Medical Data

Sudha V
Asst. Prof: Dept of Information Science & Engg.
RNS Institute of Technology
Bangalore, India

Girijamma H A
Prof.: Dept. of Computer Science & Engg.
RNS Institute of Technology
Bangalore, India

***Abstract*—The applicability and effectiveness of clustering algorithms had unquestioningly benefitted solving various sectors of real-time problems. However, with the changing time, there is a significant change in forms of the data. This paper briefs about the different taxonomies of the clustering algorithm and highlights the frequently used techniques to understand the research popularity. We also discuss the existing direction of the research work and find that still there is a significant amount of open issues when it comes to clustering medical data. We find that existing techniques are quite symptomatic in nature on local problems in clustering while problems associated with complex medical data are yet to be explored by the researchers. We believe that this manuscript will give a good summary of the effectiveness of existing clustering techniques towards medical data as a contribution.**

Keywords*—Medical Data; Clustering Algorithm; k-Means Clustering; Fuzzy; Classification

I. INTRODUCTION

In the list of challenges about unsupervised learning techniques, clustering is one of the biggest challenges till date [1] [2]. Clustering deals with exploring an elite structure from a given set of the raw database. Theoretically, the technique of organizing the objects into the group where the member of the group's bears certain similarity score with each other is known's as clustering. A good clustering technique always identifies the internal grouping from a given set of raw data. The user frames the effectiveness of the clustering performance. The user provides such forms of converging criterion. The applications of the clustering algorithm observed in many places e.g. biology, city planning, libraries, marketing, studying natural calamities, etc. [3]. For a clustering algorithm to be robust, needed that it should explore random-shaped clusters, should possess scalability, and should have high dimensionality. It should have better usability and interoperability characteristics along with insensitive features towards inputs. Most important, it should also have the potential to counter-measure the adverse effect of noise as well as outliers. A robust clustering algorithm can also state if it possesses the capability to manage higher and diversified number of attributes. Finally, it should have lower demands for domain knowledge in order to evaluate input attributes [4] [5]. However, there are also certain pitfalls associated with conventional clustering techniques, for example:

1) Higher dependencies of the spatial feature is the prime criteria of effectiveness (usually, such forms observed over distance-based clustering.

2) All the clustering and classification demands cannot be fulfilled using existing clustering techniques.

3) Defining a specific measure of the distance in case of multi-dimensional spaces is quite a challenging task,

4) Existing techniques suffers from problems with the larger dimension of the data owing to the greater extent of time complexity.

Although the outcomes of any clustering algorithm can have multiple inferences, it is hardly possible to even identify the correct number of outcomes for higher dimensional data. The clustering algorithms used over the various field but the applicability of the clustering in medical science is highly challenging. The input for clustering techniques could be any form of medical data, where the purpose could be anything right from segmentation to the classification of a specific disease condition. The original of medical data could have diversified forms (signal, image, dataset, wavelet, etc.). The medical images as quite different from the natural images as they captured from a specific data capturing device. Hence, their formats are very different that causes to implement specific forms of medical image processing. There is also a possibility of inclusion of the higher amount of noises and distortion that potentially affect the data quality. Hence, performing clustering of the medical data is one of the challenging problems in medical image processing.

In most recent times, there has been a significant amount of research work being carried out in introducing clustering techniques using various forms of data. However, with the evolution of complex medical data capturing devices and analysis, the inputs of medical data are no simpler than ten years ago. They will be required to analyze in the perfect manner to assist in effect clustering algorithm. The prime aim of this is to present a discussion about the effectiveness of the existing clustering techniques towards medical data. The discussion has been carried out using standard research papers and its contribution towards solving clustering problems.

Section II discusses the fundamental briefing of the clustering techniques followed by existing research trends in Section III. Section IV discusses the recent techniques about

clustering techniques and studying its effectiveness. The open research problems have been discussed in Section V while the summary of the work and future direction of the work is briefed in Section VI.

II. About Clustering Techniques

Clustering is a mechanism that allows the grouping of the data in the form of logical groups of certain significance [6]. One of the prime beneficial characteristics of clustering is its adaptability feature. The prime goal of the clustering algorithm is to carry out a transformation of the group of data into further meaningful data in order to ensure that the data residing in the similar group or cluster offers certain logic [7] [8]. The majority of the clustering algorithms aims to reduce the distance between two similar clusters (intra-cluster distance) and increase the distance between two different clusters (inter-cluster distance) [9]. The mechanism of clustering also termed as data segmentation owing to its characteristics of differentiating objects that also results in the identification of outliers. The usage of clustering observed in various fields e.g. machine learning, pattern recognition, false detection, analysis of the business market, etc. In the majority of the analysis, clustering tree is represented using dendrogram. The clustering technique is also frequently called as data mining technique using unsupervised approach applied for clustering (or grouping) data. For a given set of unlabeled data, mainly clustering technique explores the internal grouping.

As per theory, there are five types of clustering techniques i.e. i) Hierarchical methods, ii) Partitioning Methods, iii) Grid-based methods, iv) Machine Learning methods, v) algorithms for high dimensional data [10]. Hierarchical clustering methods are again divided into two types i.e. Agglomerative Algorithms and Divisive Algorithm. The partitioning methods of clustering are mainly of 5 type's viz. relocation algorithms, probabilistic clustering, k-medoids methods, k-means methods, and density-based algorithms. The density-based algorithms are further divided into density based connectivity clustering and density functions clustering. The clustering techniques using machine learning are again classified into type's viz. gradient descent and Artificial Neural Network and evolutionary methods. Finally, the clustering algorithms for high dimensional data is divided into three types i.e. subspace clustering, projection techniques, and co-clustering techniques. Although there are five types of clustering techniques, Fig.1 shows the frequently used clustering techniques practiced in existing research work.

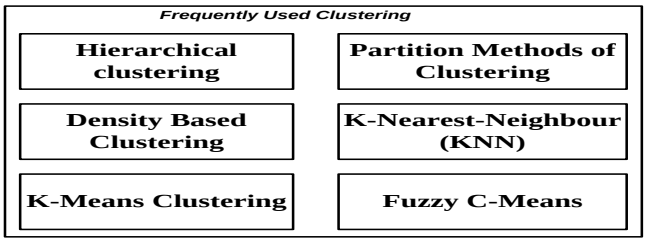


Fig. 1. Frequently used Clustering Techniques

The frequently exercised clustering mechanisms shown in Fig.1 are briefed as follows:

- ***Overlapping Clustering:*** - Normally, such technique uses fuzzy logic over grouped data in order to incorporate fuzzy membership function for the clustered data.
- ***Exclusive Clustering:*** - This technique uses a particular technique to perform grouping of data. It ensures that a data should occupy only one cluster during the grouping operation.
- ***Probabilistic Clustering:*** - Normally, such techniques are applied for optimization techniques in order to ensure the best fit between the experimental value and framework. It uses probabilistic theory. A parametric distribution is used to represent each cluster.
- ***Hierarchical Clustering:*** - This forms of clustering normally constructs or agglomerates or performs breaking up or performs the divisive operation to form a cluster hierarchy.

The brief discussions of the different forms of the clustering techniques are as follows:

A. Hierarchical clustering

A pre-determined order of cluster is formulated either from top to bottom (divisive) or vice-versa (Agglomerative) in hierarchical clustering. It is normally represented by the dendrogram. Fig.2 shows the two forms of hierarchical clustering technique.

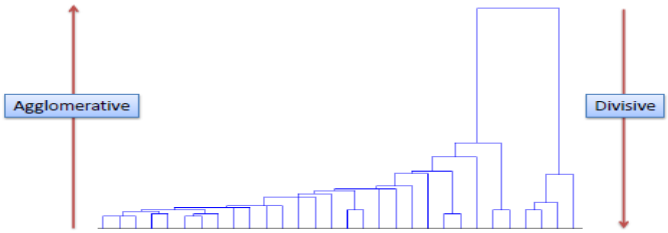


Fig. 2. Hierarchical Clustering

An agglomerative clustering initiate with the one-point group and then iteratively combines two or more precisely determined clusters. It performs the computation of all pair wise patterns for evaluating similarity coefficient. After analyzing it's each pattern in one class, it than combines the clusters to form new clusters and compute the respective distances of similarity score. This step is repeated until it ends up in k-cluster that can be one also. Similarly, divisive clustering initiates with one cluster and then iteratively divide the precise cluster. It starts its divisive operation from the top of the cluster that is distributed with the aid of flat clustering algorithm [11]. This mechanism is repeated till it reaches the singleton pattern of a cluster. Research papers use the terms called as Agglomerative Hierarchical clustering algorithm (AGNES) and Divisive Hierarchical clustering algorithm (DIANA) for agglomerative nesting and divisive analysis) respectively [12]. Both AGNES and DIANA are opposite of each other. The existing research studies that have discussed AGNES and DIANA [12]. The beneficial attributes of such

algorithms are - i) simplified implementation to offer better outcomes and ii) Lesser pre-defined information about demanded number of clusters. The limitation of such techniques would be -

1) The algorithms cannot be effectively controlled to return to its prior state if required.

2) Increase of computational resources with an increase of data points.

3) Due to the form of the spatial factor selected for combining, this algorithm is witnessed with troubleshooting while splitting larger size of clusters, higher sensitivity to outliers, challenging to manage heterogeneity in clusters. In many cases, determining the precise number of clusters is highly difficult one.

B. Partition Methods of Clustering

This technique of clustering is used for partitioning database consisting of a specific number of clusters and objects. An optimization of iterative nature is used in partitioning technique between k-number of clusters. Such technique is further classified in the form of k-means as well as k-medoids approaches. Usage of *k*-means is seen in maximum research work as it is quite simple to be incorporated in a majority of research problems. It is also one of the simple algorithms for extracting the demanded cluster number using centroid. Fig. 3 shows the conventional representation of partitioning process. The technique doesn't have any pitfalls on the types of parameters that are governed by the location of the predetermined fraction of the coordinates within the cluster location. Therefore, the grouping of the nationalities by food habit as shown in Fig.3 can be easily done using k-means clustering.

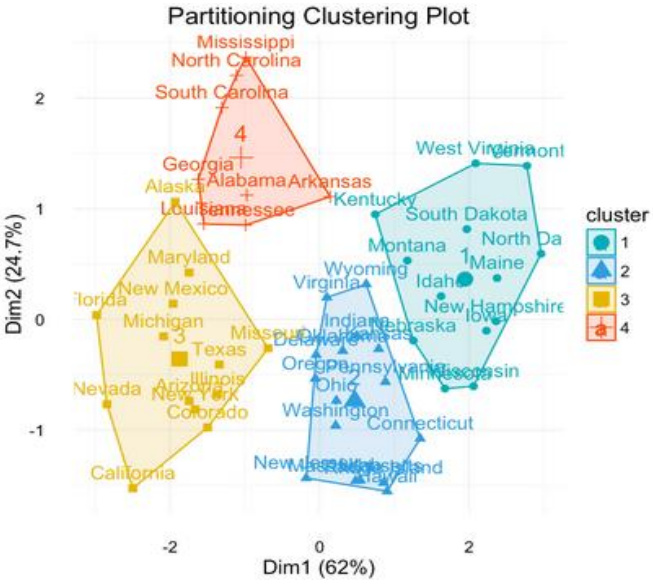


Fig. 3. Partitioning Clustering

C. Density Based Clustering

This is another frequently used clustering technique of existing system that is more inclined towards densities of the data point. This technique is more interested in exploring the random shapes cluster along with noise over a distance-based dataset. It always ensures that neighbor quantity is more than minimum data points in case the cluster is constructed. Fig.4 highlights a typical case of density-based clustering. It uses iterative processes for forming a cluster. One of the prominent pitfalls of this technique is that it cannot perform grouping of the data over the dataset of the larger dimension of differences in the cluster densities. The technique uses three different classified forms of objects e.g. classified, non-classified, and noise. A respective id of a cluster is always used for every classified object as well as noise object. However, this technique doesn't use any form of cluster id for non-classified objects. The example cited in Fig.4 shows implication of density-based clustering technique to categorize unhealthy tissue or a lesion from health tissue. It could further explore the sub-regions of different colors within the unhealthy tissue that could be again benefitted for association or classification operation. The advantage of using density-based clustering is to identify the cluster number as apriori in order comfortably manage the clusters with random dimension. However, it also suffers from the pitfalls as its inapplicability in heterogeneous densities. Moreover, its outcome highly depends on spatial measures.

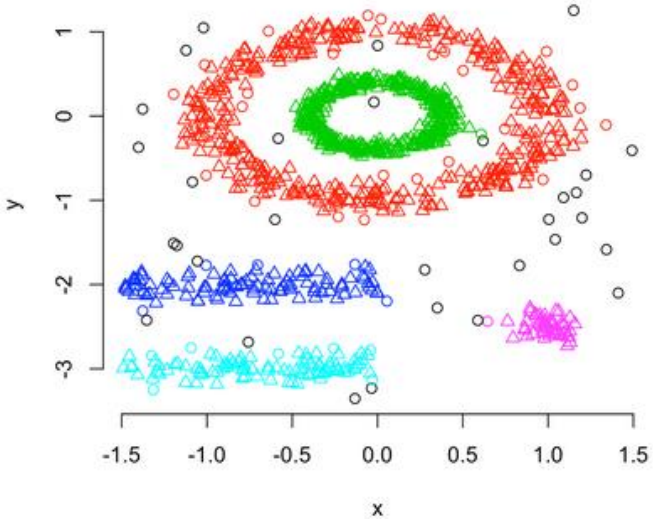


Fig. 4. Density-based Clustering

D. K-Nearest-Neighbour (KNN)

KNN algorithm is also known as memory-based clustering technique as it needs prior feeding of the samples required for training while performing processing at run time. The algorithm used in the mining operation. Different forms of the continuous parameters can be managed by the KNN algorithm although it can also work with similar capability over discrete-based properties during clustering. All the parameter in this algorithm associates distance and considers the maximum of them as far as possible. However, relationships of the parameters are not considered in this technique for computing similarity metric. This is the prime cause of errors in distance measures that significantly affects the classification accuracy. The beneficial factor of KNN algorithm is its simple implementation procedure accompanied by faster training steps. The issues in this algorithm are its dependencies of the

larger database, slower validation process, and have higher noise sensitivity.

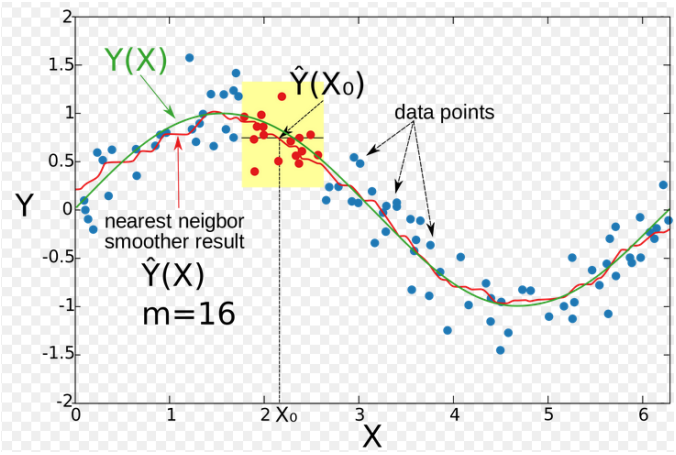


Fig. 5. KNN Clustering

E. K-Means Clustering

Usage of k-means clustering is seen in the majority of the clustering techniques. This technique is quite iterative in nature that classifies the given data in order to form k-disjoint clusters. Fig.6 shows the technique of KNN clustering for a given set of original data. The effectiveness of k-means clustering is normally assessed using squared error factor within a cluster.

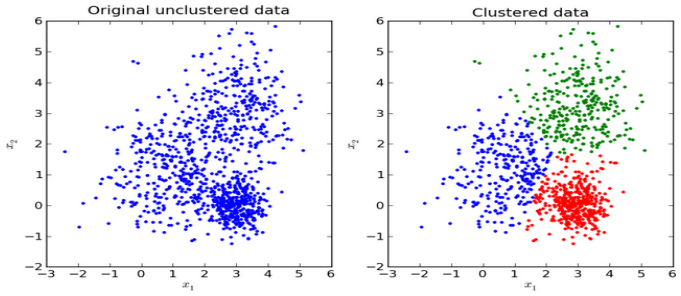


Fig. 6. K-Means Clustering

It was noted that adoption of k-means clustering forms cluster of compact form but it choose not to consider the distance between two clusters. From theoretical viewpoint, adoption of squared l_2-normalization leads to higher sensitivity in the case of maximized errors. This will eventually mean that such formulation is quite less robust from the statistical viewpoint. Only because of its simple implementation and efficiency towards computational performance, k-means algorithm is frequently used clustering technique. It also has very low memory utilization and relatively easier to understand compared to other existing clustering techniques. For distinct dataset, it offers higher precision result and offers better compactness in the cluster as compared to hierarchical clustering technique. However, it also suffers from limitations e.g. it doesn't resolve any overlapping clusters, higher dependencies of pre-determined information, random selection of clusters, the applicability only in case of presence of the mean value, and it cannot be used for outliers as well as noisy data.

F. Fuzzy C-Means

Usage of fuzzy logic over clustering has been started witnessing since last decade. Such form of the algorithm uses spatial attribute for assigning membership function mapping with the data points which is considered equivalent to the center of each group. If the nearness of the data is more towards the center of the cluster than the ability of the membership function is also more towards the cluster center. Using probabilistic approach, the sum of all the involved membership function is equivalent to 1. Fig.7 shows the mechanism of clustering in this case. The beneficial factors of using fuzzy c-means clustering are that its applicability of assigning membership functions at the center of the cluster. Moreover, fuzzy c-means algorithm is highly applicable for the dataset that is in overlapping form, and it works better as compared to a conventional k-means algorithm. However, the limitation of this technique does also exist e.g. usage of Euclidean's distance is not proportionate with the unequal weight and it involves more iterative steps. Predetermined information dependencies are another pitfall of this algorithm.

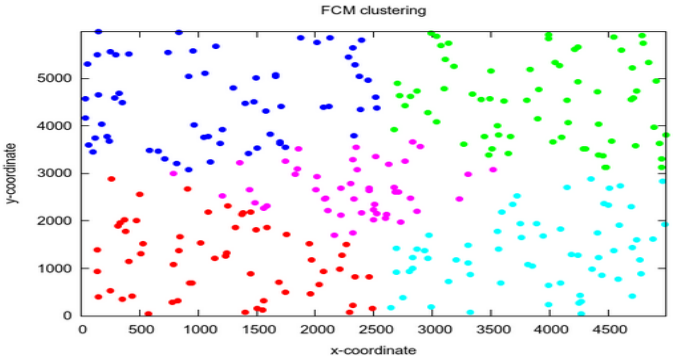


Fig. 7. Fuzzy c-Means Clustering

III. Existing Research Trend

This section discusses the existing research trends towards clustering techniques. For this purpose, we prune the research papers published between 2010 to till date from IEEE Xplore. We find that there are 14,730 conference paper and 2090 Journals associated with the problems and enhancement techniques of clustering. For an elaborated understanding, we use Fig.8 that basically furnishes two types of information i.e. i) complete classification of clustering algorithm and ii) a total number of research papers specific to each type of the clustering algorithm. It is widely known that clustering is specifically useful for performing pattern recognition, Spatial Data Analysis, Image Processing, Economic Science, document classification, data mining, etc. [13]. The survey outcome basically shows that k-means (No. of published Journal: 399, No. of Published Conference: 4226) algorithm is the highly adopted technique in classification followed by probabilistic technique (No. of published Journal: 216, No. of Published Conference: 804). Although there are other significant types of a clustering algorithm, they are less explored by the research community since 2010. Another significant trend is that all the investigation was carried out by diversified forms of the data, where maximum data is in the form of an image. There is also less specialization work of

clustering towards detection and diagnosis of the complex medical condition. Most recently, there are certain standard review papers e.g. [14] [15] [16] that has reviewed over different research work being carried out over clustering techniques. But nowhere it is found how strongly clustering technique is used over medical data or any other form of complex data. With the increasing usage of the dynamic user, the formation, processing, and distribution process of such data would be quite complex to solved. Even the frequent usage of the k-means algorithm was not much seen to address the complicated problems associated with medical images. On the other hand, there has been considerable amount of work being carried out using Artificial Neural Network (No. of published Journal: 174, No. of Published Conference: 1237), Evolutionary technique (No. of published Journal: 174, No. of Published Conference: 1229) on clustering problems, probabilistic techniques (No. of published Journal: 216, No. of Published Conference: 804), and Co-clustering technique (No. of published Journal: 216, No. of Published Conference: 1723). Hence, it can be easily said that maximum research work till date from 2010 has used k-means clustering algorithm followed by the probabilistic approach, co-clustering approach, neural network, and evolutionary techniques. Apart from this, other techniques have received less attention till date. Therefore, it can be said that usage of machine learning and portioned-based clustering techniques are predominantly used in the existing system and can also be represented as existing research trends. However, the existing survey papers don't speak about predominant clustering techniques of recent time, and hence it is quite challenge to understand the effectiveness of existing clustering techniques.

The next section discusses the existing research techniques accompanied by brief highlights of existing problems, the technique adopted to solve them with associated advantages and limitation of existing techniques.

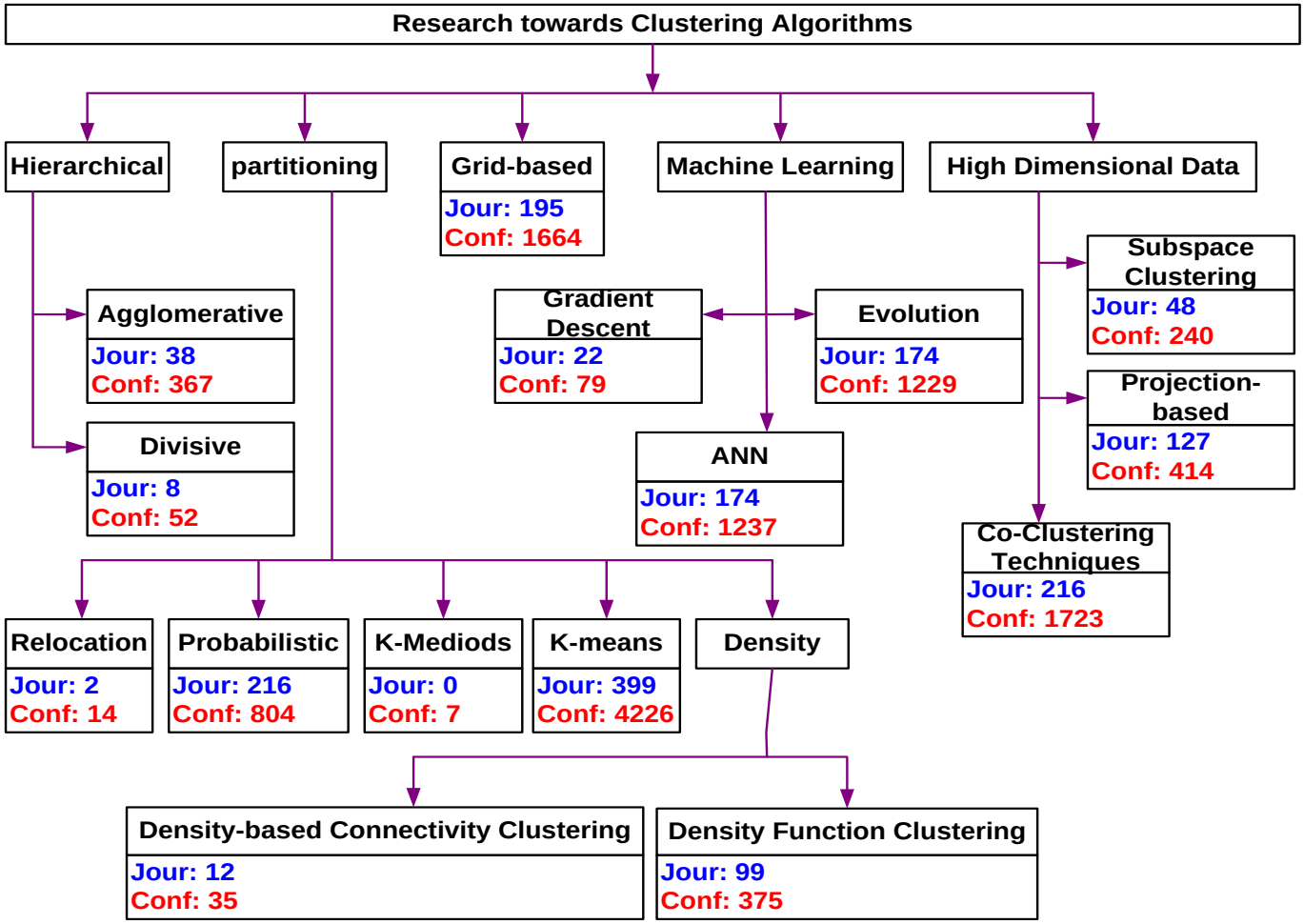


Fig. 8. Research Trend towards Clustering Algorithm

IV. Existing Research Work

This section discusses the existing research technique that has been used for enhancing clustering performance. Usage of clustering technique towards medical data is mainly associated with improving the image processing operations e.g. segmentation. The work carried out by Al-Dmour and Al-Ani et al. [17] have combinedly used unsupervised and semi-supervised classification approach in order to perform involuntary segmentation. The authors have also used the median filter as well as Fuzzy-c-means attributes for performing clustering. A technique called as subtractive clustering is used for minimizing computational complexity. Adoption of fuzzy clustering technique was also seen in the

work of Proietti et al. [18] that applies membership function of kernel-based. The study claims to extract unconstrained structure. Clustering also plays a significant role in maintaining the resolution of an image. Al-Qizwini et al. [19] have used similarity of the subspace as well as manifold clustering. Applying subspace clustering assists in extracting low ranks clusters along with usage of Principal Component Analysis (PCA). Finally, training and testing are carried out on natural images where the outcomes were testified using Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM). Although, clustering techniques is beneficial with the abundance of data, but could encounter a significant problem if data is incomplete. One of such investigation towards implementing clustering operation for a given set of impartial data was carried out using Li et al. [20]. The authors have used K-means clustering algorithm as well as a k-median method in order to perform clustering. The technique also performs minimax optimization technique for reduced complexities. The study outcome was assessed using numbers of wrongly estimated values for different clustering mechanism. Ahmad [21] have applied fuzzy clustering algorithm for breast cancer detection. The technique has applied fuzzy c-means clustering and applied an existing technique for computing the distance between two values of features. Clustering approaches were also studied with respect to the transfer function. Such direction of research work was carried out by Zhang et al. [22] where affinity-based propagation is studied over histograms of intensity gradient magnitude in order to generate transfer function. The study proved that such clustering technique assists in accomplishing better accuracy in clustering outcomes as well as it also achieves convergence point faster over medical images.

El-Khamy et al. [23] have presented a study that performs clustering of brain images in order to identify the suspected mass. The technique uses the fuzzy c-means algorithm as well as conformed threshold in order to enhance the clustering performance. The study outcome shows higher accuracy and lower processing time. Kitrungrotsakul et al. [24] have used clustering approach in order to perform segmentation that significantly minimizes the graph scale for increasing the optimization speed. Shabanzadeh et al. [25] have used biogeography-based optimization in order to perform data clustering over the real-life dataset. The study outcome was found to have better performance compared to existing clustering and optimization techniques. Haraty et al. [26] have enhanced k-means clustering for extracting diversified patterns from the medical data. The algorithm also uses greedy approach, where the outcomes of the study have been evaluated with respect to a number of items in dataset and f-measure, a coefficient of variance, etc. Hou and Lin [27] have used subspace clustering in order to carry out image retrieval. The technique uses low-rank representation and a matrix completion algorithm for performing involuntary tag completion. Usage of sparse subspace clustering was also seen in the work carried out by Wen et al. [28]. The technique also utilizes total variation method and forms a non-convex optimization model. The technique is mainly used for recovering image as well as performing clustering over the images that has incomplete information. Zhan et al. [29] have presented a clustering technique for medical images using graph-based theory. The technique uses the weighted representation of the medical image to give a shape of a completed graph that is further subjected to pruning. The study outcome was assessed using f-score. Usage of subspace clustering was also seen in the work carried out by Ziko et al. [30] where a visual descriptor was created. A supervised data is added during the clustering process that further minimizes the errors in the results. Aghabozorgi et al. [31] have used time-series data to formulate a unique hybrid clustering algorithm along with k-medoids. The study outcome of the presented work is assessed using accuracy over the cardinality of the datasets. Harchaoul et al. [32] have used the fuzzy c-means algorithm for overcoming the problems of overlapping clustering. Schultz and Kindlmann [33] have presented a technique for three-dimensional image analyses using spectral clustering. Using medical images, the technique was implemented. Boulemnadjel and Hachouf [34] have presented a technique of subspace clustering considering medical images. Paul et al. [35] have presented a simplified clustering technique that assists in the detection of specific diseases. The authors have used constraint k-means and k-mode clustering technique to achieve this. Sulaiman and Isa [36] have presented a technique of image segmentation using fuzzy k-means clustering. The interesting point is its applicability on different forms of images.

TABLE I. SUMMARY OF EXISTING CLUSTERING TECHNIQUE

Authors	Problem	Technique	Advantages	Limitation
Al-Dmour and Al-Ani et al. [17]	Segmentation	Median filter, fuzzy c-means, subtractive clustering	Low complexity	Not applicable to complex medical data.
Proietti et al. [18]	Minimizing error rate	Kernel-based construction of fuzzy members	Minimal Error rate	-No benchmarking -cannot perform classification of a complex disease condition.
Al-Qizwini et al. [19]	To attain super-resolution	Subspace clustering, PCA	Good Image quality	-Not testified over medical data -Not compared with existing techniques.
Li et al. [20]	Impartial data availability	Minimax Optimization, k-means, k-median	Minimizes complexity	Less likely to work of complex medical data.
Ahmad [21]	Enhancing clustering performance	Fuzzy c-means clustering	Only5% error in clustering	-Less likely applicable on complex medical data -computational performance not stated.
Zhang et al. [22]	Enhancing clustering performance	Affinity propagation clustering	Faster convergence	-Discussion restricted to volume visualized data.
El-Khamy et al. [23]	Adaptive clustering	Fuzzy c-means conformed thresholding	Higher accuracy, and faster processing time	Less effective benchmarking, complexity performance not discussed.
Kitrungrotsakul et al. [24]	Segmentation	Linear iterative clustering	Lower computational time	-Less likely applicable on complex medical data -Higher complexities for multi-modal image
Shabanzadeh et al. [25]	Clustering optimization	Biogeography-based optimization	Lower error rates	-uses of the recursive function to increase the complexity -less scalable approach
Haraty et al. [26]	Enhancing clustering performance	k-means, Greedy approach	Higher and stable F1-score	No-benchmarking
Hou and Lin [27]	Image retrieval	Subspace clustering	Benchmarked outcome	No tested over a medical image.
Wen et al. [28].	Clustering performance	Sparse Sub-space clustering	Efficient image recovery	No tested over a medical image.
Zhan et al. [29]	Medical image clustering	Undirected graph, sparsification	Minimal run time involved	Less Effective benchmarked outcomes.
Ziko et al. [30]	Constructing visual dictionary	Subspace clustering	Minimized Error	Complexity performance not stated.
Aghabozorgi et al. [31]	Cluster enhancing	Hybrid clustering	Good accuracy in classification	No tested over a medical image.
Harchaoul et al. [32]	Data analysis	Fuzzy C-means, probabilistic	Achieved good clustering accuracy, applicable to brain MRI	Not tested over a complex form of data.
Schultz and Kindlmann [33]	3D Medical Image Analysis	Spectral Clustering	Simplified usage, extensive operation	-N/A-
Boulemnadjel and Hachouf [34]	Clustering enhancement	Subspace Clustering	Applicable on original data	-Not benchmarked -More iteration leads to complexity.
Paul et al. [35]	Disease detection	k-means, k-mode	Involves for discrete and continuous data	-Less effective benchmarking -Less likely to work on complex medical data.
Sulaiman and Isa [36]	Image segmentation	Fuzzy k-means clustering	Applicable for post image processing stage.	-Doesn't address complexity problems -Applicability on complex medical image is not discussed

V. Open Research Issues

This section discusses the open research issues after reviewing the standard clustering techniques as well as some of the significant research being carried out by recent times.

- *Less emphasis on classifying medical data:* Without precise classification, complex medical data cannot be subjected for diagnosis. Such complex medical data are often High-dimension and difficult to perform clustering. Owing to data complexity, existing classification techniques cannot be applied. The biggest challenge is to select one smaller set of highly precise data (suitable for diagnosis) from the massive volume of complex medical data.
- *Few works towards Multi-tier Clustering:* The majority of the Advanced Radiological images (e.g. MRI etc.) are gray scale and not a true color which poses challenges towards the investigation. The majority of the existing diagnosis from medical data is based on a region of interest. Accurate labeling of a region of interest is not feasible in real-time, and hence it demands multi-tier clustering. Existing clustering techniques uses single-tier approach (i.e. using the single template). Some of the challenges in the medical data that are not addressed are i) automatic detection of metastatic stages in medical images, ii) large-scale evaluation of disease detection followed by classification, and iii) automated segmentation
- *Less work towards Clustering Complex Disease Pattern:* Frequently used medical data doesn't exhibit heterogeneous symptoms associated with the particular disease. Existing cluster analysis is not effective towards identifying disease heterogeneity.
- *More inclination towards recursive-based approach:* It has also been seen that maximum studies in the existing system have been used the recursive function which calls for more number of iterative steps to achieve the stage of convergence or meet the objective function. Existing studied has been only testified with respect to time complexity and very few studies to be testified for space complexities. There is less availability of studies that considers using the non-recursive approach in the process of optimization.

Although there is the certain level of work being carried out towards enhancing clustering techniques, it can be easily seen that majority of them are associated with limitations (Table 1 of Section IV). Classification of the disease condition with faster response time and lower computational complexity is the critical demands of clustering techniques over medical images. There is a less number of analytical modeling designed using any of the existing clustering techniques for enhancing the classification performance. Moreover, usage of multi-dimensional technique can further leverage the disease classification while formulating novel clustering technique. Such technique can be used for performing clustering of the medical data with the complex disease condition. However, in an existing system, the term medical data is found maximum corresponding to image only. A closer look at the existing system also shows that there are various clustering techniques that offer lower time complexities. However, there is no such evidence if such claims will be applicable while changing the environments. It will also mean lower applicability in a physical world and more on research work. A closer look at the existing system shows that adoption of the complex medical dataset is few to find. Even with the general medical data, the multiple modalities among the images are quite less to find.

It was also observed that there had been various clustering techniques presented in the past with MRI image that are normally bigger in size using k-means clustering algorithm. In fact, the majority of such scheme is similar to this. Maximum of such techniques are found to provide non-intuitive outcomes of classification. Such outcomes are never considered to be understood completely by the radiologist or attending physician in real-time practices. For better outcomes, it is necessary to perform inference of the clinical outcomes using simple rules. Unfortunately, the complex medical data e.g. that of gene expression data are normally collected in the form of high dimensional format. Such data not only have the higher value of veracity but it also has a greater extent of outliers and noise. Therefore, it is quite a challenging task to design and develop a technique that can deal with such complicated issues of clustering.

VI. Conclusion & Future Work

Clustering is the better way to deal with the classification of the higher number of data by performing logical groups. This paper discusses the theoretical aspects of clustering and its applications and taxonomies. By reviewing the existing clustering schemes, we find that it uses the common database with no clustering algorithm to represent disease heterogeneity. Moreover, existing algorithms are quite specific to the medical database. However, to cope up the rising demands of clustering, it is required that it should start analyzing the database of complex disease condition as well as it should also address disease heterogeneity. It is also required that the algorithm should be working on multiple forms of a complex dataset with nearly similar outcomes. Finally, the paper highlights the open research problem associated with clustering of medical data. Our future work will be in the direction to find the certain robust solution for open resource issues. Our first approach will be to develop a novel prioritizing scheme to select the best sub-cluster from complex medical data followed by application of an enhanced fuzzy logic on the informative sub-cluster extracted from complex medical data. A novel labeling technique would be formulated to assists in extraction of normal as well as the abnormal region. Its consecutive approach will be to formulate a framework for disease classification to address problems associated with multi-tier clustering. A novel multi-modal scheme will be developed for extracting significant features from complex medical data. A study-specific optimal pattern selection strategy will be designed to obtain multiple patterns from data. This step could be further enhanced by performing extraction of multi-modal regional feature representation for each subject from multiple pattern spaces. We will also develop a new technique of Sub-class Clustering-Based Feature Selection by applying supervised learning to perform classification. Our final phase of the study will be to formulate clustering framework for

complex disease pattern to address the problem of clustering complex disease pattern. A generative scheme with probability theory is the best way to start this for designing an iterative method for finding maximum likelihood or maximum a posteriori (MAP) estimates of complex disease patterns and parameters. Finally a novel clustering technique can be designed for extracting the patterns of complex disease.

REFERENCES

[1] M. E. Celebi, K. Aydin, Unsupervised Learning Algorithms, Springer-Technology & Engineering, 2016

[2] Y. Yang, Temporal Data Mining via Unsupervised Ensemble Learning, Elsevier, 2016

[3] C. C. Aggarwal, C. K. Reddy, Data Clustering: Algorithms and Applications, CRC Press, 2016

[4] C. Karlsson, Handbook of Research on Innovation and Clusters: Cases and Policies, Edward Elgar Publishing, 2008

[5] J. N. Sheth, Lawrence Sherman, Cluster Analysis and its Applications in Marketing Research, Marketing Classics Press, 2011

[6] S. Dua, P. Chowriappa, Data Mining for Bioinformatics, CRC Press, 2012

[7] R. Lee, Applied Computing & Information Technology, Springer, 2015

[8] S. N. Bhaduri, D. Fogarty, Advanced Business Analytics: Essentials for Developing a Competitive Advantage, Springer, 2016

[9] O. Maimon, L. Rokach, Data Mining and Knowledge Discovery Handbook, Springer Science & Business Media, 2010

[10] K. R. Venugopal, K.G. Srinivasa, L. M. Patnaik, Soft Computing for Data Mining Applications, Springer, 2009

[11] http://nlp.stanford.edu/IR-book/html/htmledition/flat-clustering-1.html

[12] H. J. Miller, J. Han, Geographic Data Mining, and Knowledge Discovery, Second Edition, CRC Press, 2009

[13] J. Han, J. Pei, M. Kamber, Data Mining: Concepts and Techniques, Elsevier, 2011

[14] N. Y. Saiyad, H. B. Prajapati and V. K. Dabhi, "A survey of document clustering using semantic approach," *2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*, Chennai, 2016, pp. 2555-2562

[15] A. Fahad, N. Alshatri, Z. Tari," A Survey of Clustering Algorithms for Big Data: Taxonomy & Empirical Analysis", IEEE Transactions on Emerging Topics in Computing, 2014

[16] J. Li; H. W. Lewis, "Fuzzy Clustering Algorithms – Review of theApplications", IEEE International Conference on Smart Cloud, 2016

[17] H. Al-Dmour and A. Al-Ani, "MR Brain Image Segmentation Based on Unsupervised and Semi-Supervised Fuzzy Clustering Methods," *2016 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, Gold Coast, QLD, 2016, pp. 1-7.

[18] A. Proietti, L. Liparulo and M. Panella, "2D hierarchical fuzzy clustering using kernel-based membership functions," in *Electronics Letters*, vol. 52, no. 3, pp. 193-195, 2 4 2016.

[19] M. Al-Qizwini, C. Dang, M. Aghagolzadeh and H. Radha, "Image super-resolution via Dual-Manifold Clustering and Subspace Similarity," *2016 IEEE International Conference on Image Processing (ICIP)*, Phoenix, AZ, 2016, pp. 1429-1433.

[20] J. Li, S. Song, Y. Zhang, and Z. Zhou, "Robust K-Median and K-Means Clustering Algorithms for Incomplete Data, Hindawi Publishing Corporation Mathematical Problems in Engineering, 2016

[21] A. Ahmad, "Evaluation of Modified Categorical Data Fuzzy Clustering Algorithm on the Wisconsin Breast Cancer Dataset", Hindawi Publishing Corporation Scientifica, 2016

[22] T. Zhang, Z. Yi, J. Zheng, D. C. Liu, W-M Pang, "A Clustering-Based Automatic Transfer Function Design for Volume Visualization", Hindawi Publishing Corporation, Mathematical Problems in Engineering, 2016

[23] S. E. El-Khamy, R. A. Sadek and M. A. El-Khoreby, "An efficient brain mass detection with adaptive clustered based fuzzy C-mean and thresholding," *2015 IEEE International Conference on Signal and Image Processing Applications (ICSIPA)*, Kuala Lumpur, 2015, pp. 429-433.

[24] T. Kitrungrotsakul, X. H. Han and Y. W. Chen, "Liver segmentation using superpixel-based graph cuts and restricted regions of shape constrains," *2015 IEEE International Conference on Image Processing (ICIP)*, Quebec City, QC, 2015, pp. 3368-3371.

[25] P. Shabanzadeh and R. Yusof1, "An Efficient Optimization Method for Solving Unsupervised Data Classification Problems", Hindawi Publishing Corporation, Computational and Mathematical Methods in Medicine, 2015

[26] R. A. Haraty, M. Dimishkieh, and M. Masud, "An Enhanced k-Means Clustering Algorithm for Pattern Discovery in Healthcare Data", Hindawi Publishing Corporation, International Journal of Distributed Sensor Networks, 2015

[27] T. Kitrungrotsakul, X. H. Han and Y. W. Chen, "Liver segmentation using superpixel-based graph cuts and restricted regions of shape constrains," *2015 IEEE International Conference on Image Processing (ICIP)*, Quebec City, QC, 2015, pp. 3368-3371

[28] X. Wen, L. Qiao, S. Ma, W. Liu and H. Cheng, "Sparse Subspace Clustering for Incomplete Images," *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, Santiago, 2015, pp. 859-867.

[29] Y. Zhan, H. Pan, Q. Han, Xiaoqin Xie, Zhiqiang Zhang and Xiao Zhai, "Medical image clustering algorithm based on graph entropy," *2015 12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, Zhangjiajie, 2015, pp. 1151-1157.

[30] I. M. Ziko, E. Fromont, D. Muselet and M. Sebban, "Supervised spectral subspace clustering for visual dictionary creation in the context of image classification," *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, Kuala Lumpur, 2015, pp. 356-360.

[31] S. Aghabozorgi, T. Y. Wah, T. Herawan, H. A. Jalab, "A Hybrid Algorithm for Clustering of Time Series Data Based on Affinity Search Technique", Hindawi Publishing Corporation The Scientific World Journal, 2014

[32] N-E El Harchaoui, M. A. Kerroum, A. Hammouch, "Unsupervised Approach Data Analysis Based on Fuzzy Possibilistic Clustering: Application to Medical Image MRI", Hindawi Publishing Corporation, Computational Intelligence, and Neuroscience, 2013

[33] T. Schultz and G. L. Kindlmann, "Open-Box Spectral Clustering: Applications to Medical Image Analysis," in *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 12, pp. 2100-2108, Dec. 2013.

[34] A. Boulemnadjel and F. Hachouf, "A new method for finding clusters embedded in subspaces applied to medical tomography scan image," *2012 3rd International Conference on Image Processing Theory, Tools and Applications (IPTA)*, Istanbul, 2012, pp. 383-390.

[35] R. Paul and A. S. M. L. Hoque, "Clustering medical data to predict the likelihood of diseases," *2010 Fifth International Conference on Digital Information Management (ICDIM)*, Thunder Bay, ON, 2010, pp. 44-49.

[36] S. N. Sulaiman and N. A. Mat Isa, "Adaptive fuzzy-K-means clustering algorithm for image segmentation," in *IEEE Transactions on Consumer Electronics*, vol. 56, no. 4, pp. 2661-2668, November 2010.

Improvement of Data Transmission Speed and Fault Tolerance over Software Defined Networking

SM Shamim
Dept. of Information and Communication and Technology
Mawlana Bhashani Science and Technology University
Dhaka, Bangladesh

Mohammad Badrul Alam Miah
Dept. of Information and Communication and Technology
Mawlana Bhashani Science and Technology University
Dhaka, Bangladesh

Nazrul Islam
Dept. of Information and Communication and Technology
Mawlana Bhashani Science and Technology University
Dhaka, Bangladesh

***Abstract*—Software Defined Networking (SDN) is a new networking paradigm where control plane is decoupled from the forwarding plane. Nowadays, for the development of information technology large number of data traffic has been added in the global network each day. Due to proliferation of the Internet, e-commerce, video content and personalized cloud-based services higher channel bandwidth required to deliver larger data from one center to others. Lower data communication speed and fault tolerance are major factors for SDN which degrades network performance. This paper presents enhancement of data communication speed and fault tolerance over SDN using Link Aggregation Control Protocol (LACP. The result of this paper shows network performance has been improved by increasing approximately 31% data transmission speed over SDN using LACP. Moreover, this paper shows fault tolerance have been improved by LACP which prevents failure of any single component link from leading to breakdown of the entire communications.**

Keywords—Fault Tolerance; Link Aggregation Control Protocol (LACP); OpenFlow; Mininet Emulator; Software Defined Networking (SDN)

I. INTRODUCTION

Software Defined Networking (SDN) [1, 2, 3] is a new approach for managing, building and designing computer networks which decouple the network's control plane from the forwarding planes. It has emerged as new paradigm in networking which has the possibility to enable ongoing network innovation and enable the network as a programmable, pluggable component of the larger cloud architecture [4]. SDN is being strongly considered as the next promising networking platform. In recent years, SDN has been developing tremendously in different organizations [5]. In order to reduce operational costs and strengthen network architecture different companies are planning or deploying SDN in their network [6]. In the next five years, SDN will be considered one of the most advanced information technologies over the world [7, 8]. About US $2 billion has been estimated to invest in SDN for knowledge discovery [9].

In order to handle the larger data high configuration router and switch are needed. Server and storage resources are interconnected via switches and routers [10]. Adding more switch and router will increase operational cost which reduces the network performance. In addition, network path failure one of the major problem which reduce network efficiency. An efficient routing, sever load balancing, access control and traffic monitoring system has to be designed to overcome these limitation. One of the possible solutions of these problems is Link Aggregation control protocol where two or more ports in an Ethernet switch are combined together to operate as a single virtual port. It increases available bandwidth by aggregating two or more links between network devices.

Due to programmability of Software Defined Networking, standard mechanisms needed for achieving higher data transmission speed and fault tolerance. Different researchers propose few techniques [11, 12, 13, 14 and 15] to improve data transmission speed and fault tolerance over Software Defined Networking. In this consequence, the paper [11] analyzed Bandwidth and latency aware routing using OpenFlow over SDN which improve network performance. Paper [12] proposed a novel architecture BRAS (Broadband Remote Access Server) which could enhance data transmission speed according to users' preference in specific applications.

In [13] authors presents Software Defined Networking based on OpenFlow can be used to build efficient solutions in order to handle fault-tolerant multicast in substation environments. Their implementation handles single link failure and also indicates how their approach can be expanded to handle multiple link or node faults. Fault tolerance issue with systematically review the existing methods has been proposed in [14] which are useful in failure recovery. In [15] proposes new architecture to strengthen the reliability and fault-tolerance over SDN in terms of network operations and management. After studying related works realized that several researchers improve data transmission speed and fault tolerance over SDN separately with different technique.

However, there needed further studies to improve data transmission speed and fault tolerance over SDN in order to increase network performance. Yet, there has been lack of studies, which can enhance both of these two major facts at same time. Though Link Aggregation control protocol is known for traditional network architecture, no implementation has been done over Software Defined Networking. Link Aggregation [16] is a technology defined in IEEE802.1AX-2008, which is a method of combining multiple physical lines to be used as a logical link. It increases capacity and

availability of the network between specific devices (both switches and end stations) with the help existing Fast Ethernet and Gigabit Ethernet technology. It has higher potential transmission speed and higher accessibility in contrast to conventional connections using an individual cable. The purpose of this paper is to enhance data communication speed and fault tolerance over SDN at same time using Link Aggregation control protocol (LACP).

The rest of the paper is organized as follows. Section II detail describes Link Aggregation with the types in background. Section III describes the research methodology. Section IV presents the experiment setup in details. Experimental results and discussion are evaluated in Section V. Section VI concludes the paper and deliberates the future perspectives.

II. Background Study

In Link Aggregation, multiple parallel physical links are combined together between two devices in order to form single logical link. This function also provides load balancing where the processing and communications activity distributed over multiple links to avoid single link overwhelmed. The architecture diagram of Link Aggregation function is shown on Fig-1.

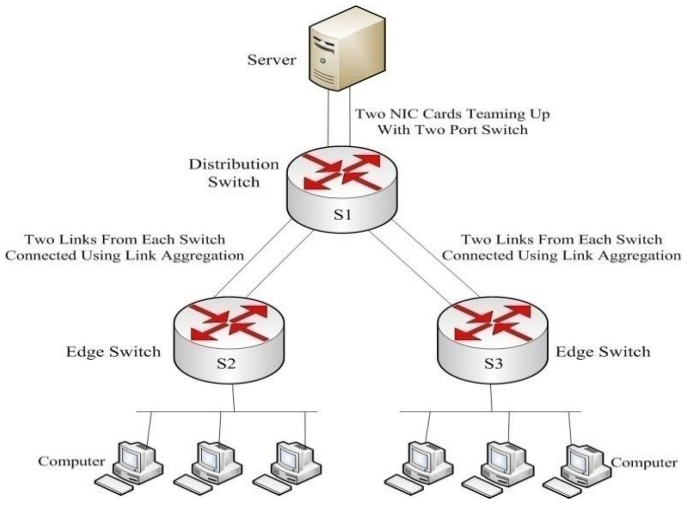


Fig. 1. Link Aggregation Architecture Diagram [17]

Two Edges switch are connected to distribution switch and distribution switch connected to server using link aggregation. In order to form a higher data transmission capacity several link from server may connect with different switch ports. Link Aggregation provides higher link availability, increase link capacity and arrogates replace upgrading over conventional network [17].

III. Research Methodology

A literature review has been performed in Software Defined Networking research scope and challenges. After the review, fault tolerance and data communication speed option arrived over SDN. A handy simulation tool was needed to analyze the SDN. Different experimental studies have been performed among OMNET++ [18], EstiNet [19], OFNet [20], Maxinet [21], NS-3 [22] and Mininet [23, 24]. The studies appear that open-source network simulator Mininet has good potential for simulating Software Defined Networking. In order to evaluate designed network topology with OpenFlow virtual switch Mininet has been installed over Ubuntu 14.04. Its installation and configuration is easy and straightforward than other simulators. Virtual Software Defined Networking can be designed using Mininet which consists of OpenFlow [25] controller, OpenFlow-enabled Ethernet switches and multiple hosts connected to those switches. OpenFlow is one of the most widely deployed SDN communications standards protocols. This protocol used in order to communicate between controller and other networking devices i.e. switch, router etc. A component based Software Defined Networking framework Ryu has been used as OpenFlow controller. Ryu Controller managed and maintained by open Ryu which is written in Python [26]. After careful study of the experiment different network analysis graph has been plotted which shows the expected results.

IV. Experiment Setup

Custom network topology has been designed using Mininet API which is shown on Fig-2. Designed network consist of one OpenFlow switches, an OpenFlow Ryu controller and three hosts. All the host h1, h2 and h3 are connected with the OpenFlow switch s1. Link Aggregation function has been implemented between OpenFlow switches s1 and host h1.

All of the hosts have assigned unique IP address and MAC address. The IP address and MAC address for host h1 are ’10.0.0.1/24’ and ’00:00:00:00:00:01’. For all the other host corresponding IP and MAC address is also assigned i.e. host h2 (’IP=10.0.0.2/24’ and MAC=’00:00:00:00:00:02’) , host h3 (IP=’10.0.0.3/24’ and MAC=’00:00:00:00:00:03’) , and host h4 (IP=’10.0.0.4/24’ and MAC=’00:00:00:00:00:04’).

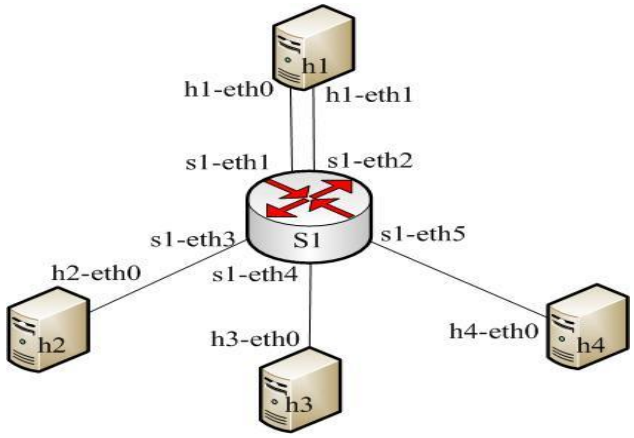


Fig. 2. Designed Network Topology

The Linux bonding driver [27] provides a method for combining more than one network interface controllers (NICs) into single logical bonded interface. Initially, bonding driver module has been loaded in host h1 to perform link aggregation. There are two interfaces in host h1 which are h1-eth0 and h1-eth1. These two interfaces are bond together in order to form one logical interface, i.e. bond0One of the commonly used network analysis tool iperf has been used to measure the performance. Network analysis tool Wireshark formerly known

as Ethereal has been used which captures packets in real time and display in human-readable format.

Two scenarios have been executed to evaluate the performance where deigned network topology has been executed without LACP implementation and secondly topology executed with LACP implementation. The corresponding result of each execution has been captured by Wireshark. For each corresponding result three performance analysis graphs throughput graph, time sequence graph and round trip time graph has been drawn. Details comparison among these graphs has been shown in result section.

V. Experiment Result

A. Throughput Graph

In data communication network throughput refers to average rate of successful message delivery over a transmission channel which measured in bits per second or in data packets per second or data packets per time slot. Data may be delivered over a physical or logical link, or pass through a certain network node. Figure-3 shows throughput graph with implementation of LACP over SDN and Figure-4 shows another throughput graph for without implementation of LACP over SDN. Throughput graph is valuable in understanding end-to-end performance.

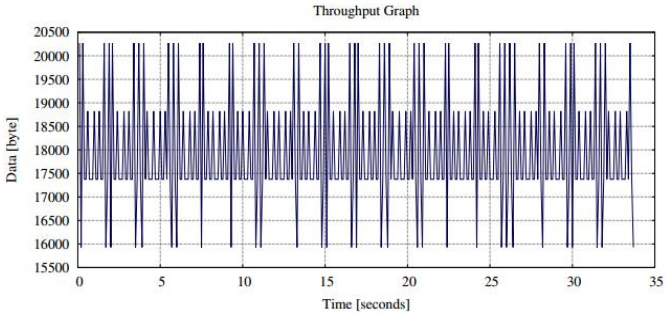


Fig. 3. With LACP SDN Throughput Graph

From the With LACP SDN throughput graph Fig-3, highest throughput in bytes approximately 20250 bytes and lowest is 16000 bytes. There are about 4202 packets has been transmitted between sender and receiver. The size of average packet and average Bytes per second are 108.765 and 125004.869.

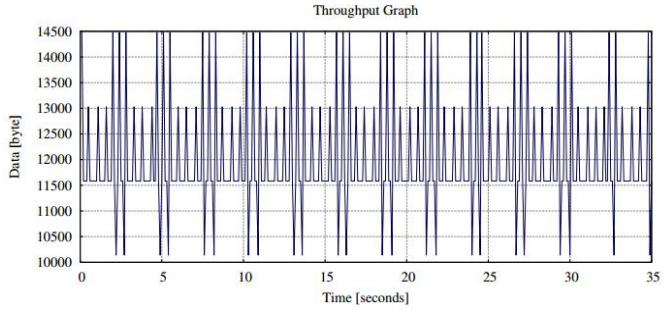


Fig. 4. Without LACP SDN Throughput Graph

From the Without LACP SDN throughput graph Fig-4, there are about 2896 packets has been exchanged between sender and receiver. The size of average packet and average Bytes in per second are approximately 82.699 and 82701.

Without LACP SDN throughput begins from the lower value approximately 10,100 after that it keeps on increasing and decreasing within some specific range. At the same time, the LACP SDN throughput graph begins with large value approximately 16,000. After some time it decreases sharply and again increase sharply within some specific range. A details comparison between two throughput graphs has been shown on table-1.

TABLE I. Comparison of Throughput Result between LACP SDN and without LACP SDN

Features	LACP SDN	Without LACP SDN
Total Packets	4202	2896
Time Duration Between First And Last (sec)	54.864	35.019
Avg. Packets/Sec (bytes)	108.765	82.699
Avg Packet Size	1506.591	1511.572
Bytes	6330694	4377512
Avg Bytes/Sec	163864.50	125004.87
Avg Mbit/Sec	1.311	1.00

From the Table-1, average megabyte per second for the without LACP SDN is 1.00MB and for the LACP SDN is 1.311 MB which is approximately 31% higher. The simulation graph and data table-1 are shown that LACP SDN has higher rates of throughputs than without LACP SDN. Data communication speed has been improved approximately 31% for LACP SDN compare to without LACP SDN.

B. TCP Time Sequence Graph

Time-Sequence graphs visualize TCP-based traffic. In an ideal situation, the graph plots from the lower left corner to the upper right corner in a smooth diagonal line.

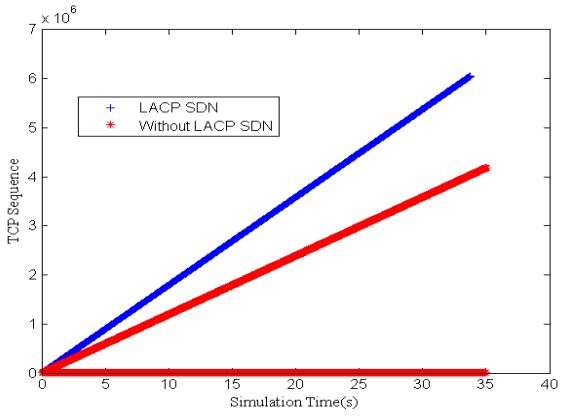


Fig. 5. TCP Time Sequence Graph

Fig-5 shows comparison of two time sequence graph where LACP SDN time sequence graph identified by blue lines and without LACP SDN time sequence graph indicated by red lines. The Y axis defines the TCP sequence numbers and X axis defines the simulation time. The slope of the line would be the theoretical bandwidth of the pipe. The steeper the line, the

higher the throughput. From the Fig-5 more packets has been transmitted for LACP SDN compare to without LACP SDN.

C. Round Trip Time Graph

Round-trip time (RTT) is the length of time takes a data packet to be sent plus length of time takes for acknowledgment to be received of that packet to be received. RTT Graph depicts round trip time from a data packet to corresponding ACK packet. The Y axis is created based on the highest round trip latency time. Latency times are calculated as the time between a TCP data packet and the related acknowledgment.

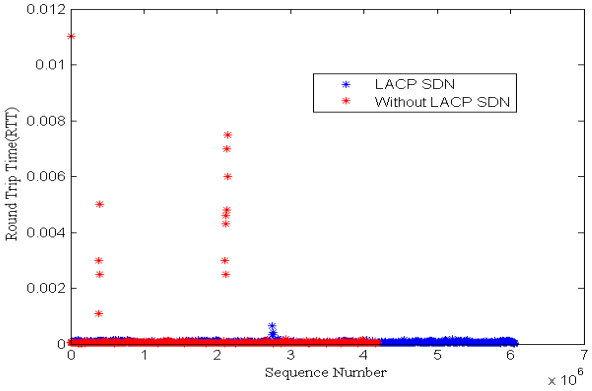


Fig. 6. TCP Round Trip Time Graph

The Y axis defines the Round Trip Time (RTT) in seconds and X axis defines TCP Sequence Numbers. Fig-6 shows latency times are very high at few points in the trace file and there are specific moments when the traffic is bursty in nature. For without LACP SDN RTT graph, some packet has higher latency time compare to average latency time. There are also some packets for LACP SDN RTT graph has higher latency time. Lower round trip time for the corresponding sequence number always expected. After comparing two graphs, LACP SDN RTT graph shows better output which has less round trip time than without LACP SDN RTT graph.

D. Improvement of Fault Tolerance

Fault tolerance is the ability of the system to perform its function even in the presence of one of multiple link failures. Fault tolerance setup or configuration prevent computer network device from failing in the event of an unexpected problem or error among connected devices. It is one of the major problems for networking application which can be improved by using Link Aggregation function. Designed network topology in Fig-2, two aggregated link are available between host h1 and switch s1 where each side link has separated port numbers (h1-eth0, h1-eth1, s1-eth0, and s1-eth1). All of the host (h2, h3 ad h4) can communicate with the host h1 using either port s1-eth1 or port h1-eth0. Each of the host can also communicate by using port s1-eth2 in switch and port h1-eth1 in host h1. Now disable one communication channel from aggregated group where disabled channel port number h1-eth0 which is opposite interface of s1-eth1.

After separating one channel, test the network connectivity by using ping command which sends ICMP echo request message and wait for corresponding reply between defined nodes. Now check connectivity between host h2 and host h1 and captured the corresponding result using Wireshark. The result shows in Fig-7, host h2 still communicate with host h1 using port s1-eth2 and port h1-eth1.

No.	Time	Source	Destination	Protocol	Length	Info
19	3.632590	10.0.0.2	10.0.0.1	ICMP	98	Echo (ping) request
20	3.632635	10.0.0.1	10.0.0.2	ICMP	98	Echo (ping) reply
26	4.631424	10.0.0.2	10.0.0.1	ICMP	98	Echo (ping) request
27	4.631459	10.0.0.1	10.0.0.2	ICMP	98	Echo (ping) reply
30	5.631364	10.0.0.2	10.0.0.1	ICMP	98	Echo (ping) request
31	5.631395	10.0.0.1	10.0.0.2	ICMP	98	Echo (ping) reply
32	6.631437	10.0.0.2	10.0.0.1	ICMP	98	Echo (ping) request
33	6.631469	10.0.0.1	10.0.0.2	ICMP	98	Echo (ping) reply

Fig. 7. Ping Result from host h2 to host h1

If we remove port h1-eth1 from the aggregation, h2 will still communicate to the host h1 using port s1-eth1 in the switch and port h1-eth0 in the host h1. Instead of failure occurs one links, Link Aggregation function able to check and automatically recover the communication using other links.

VI. Conclusion

This paper presents improvement of data transmission speed and fault tolerance over Software Defined Networking. The result obtained after extensive simulation study which has evaluated by TCP time sequence graph, throughput graph and round trip time graph. Throughput graph shows data communication speed has improved approximately 31% over SDN by using Link Aggregation control protocol. Simulation result in TCP time sequence graph and RTT graph shows network performance also improved for LACP SDN. LACP ensure failure safety systems which are crucial for every network administrator. The automatic configuration protocol LACP provides redundancy with dynamic switching to the standby link in case the active link fails. Moreover, it can be implemented in SDN using existing hardware which decreases the operational cost for upgrading the performance and resiliency of a system. Our future works involves improvement of fault tolerance and data transmission speed over Software Defined Wireless Networking (SDWN).

References

[1] B. Nunes, M. Mendonca, X.-N. Nguyen, K. Obraczka, T. Turletti, et al., "A Survey of Software-Defined Networking: Past, Present, and Future of Programmable Networks," *Communications Surveys & Tutorials, IEEE*, vol. 16, no. 3, pp. 1617–1634, 2014.

[2] Y. Jarraya, T. Madi, and M. Debbabi, "A Survey and A Layered Taxonomy of Software Defined Networking," *Communications Surveys & Tutorials, IEEE*, vol. 16, no. 4, pp. 1955–1980, 2014.

[3] D. Kreutz, F. M. Ramos, P. Esteves Verissimo, C. Esteve Rothenberg, S. Azodolmolky, and S. Uhlig, "Software-Defined Networking: A Comprehensive Survey," *Proceedings of the IEEE*, vol. 103, no. 1, pp. 14–76, 2015.

[4] M. Kobayashi, S. Seetharaman, G. Parulkar, G. Appenzeller, J. Little, J. Van Reijendam, P. Weissmann, and N. McKeown, "Maturing Of OpenFlow and Software-Defined Networking through Deployments," *Computer Networks*, vol. 61, pp. 151–175, 2014.

[5] B. Leng, L. Huang, X. Wang, H. Xu, and Y. Zhang, "A Mechanism For Reducing Flow Tables in Software Defined Network," *IEEE International Conference on Communications (ICC)*, pp. 5302–5307, IEEE, 2015.

[6] C. Yoon, T. Park, S. Lee, H. Kang, S. Shin, and Z. Zhang, "Enabling Security Functions With SDN: A Feasibility Study*," Computer Networks*, 2015.

[7] "Network world, gartner: 10 critical it trends for the next five years." Online Available: http://www.networkworld-.com/news/2012/102212gartner-trends-263594.html. Accessed: 2015-11-20.

[8] M.t. review, 10 emerging technologies: Tr10: software-defined networking." Online Available: http://www2.technologyreview.com/article/412194/tr10software-defined-networking. Accessed: 2015-11-20.

[9] "Enterprise networking, IDC: SDN a 2 billion market by 2016," Online Available: http://www.enterprisen-etworkingplanet.com/datacenter/-idcsdn-a-2-billion-market-by-2016.html. Accessed: 2016-10-25.

[10] TBitar, N., Gringeri, S., & Xia, T. J. "Technologies and protocols for data center and cloud networking". *IEEE Communications Magazine*, 51(9), 24-31. (2013).

[11] Pongsakorn, U., Kohei, I., Putchong, U., Susumu, D., & Hirotake, A. (2014). "Designing of SDN-Assisted Bandwidth and Latency Aware Route Allocation". *Journal of Information Processing Society of Japan*, 2014(2), 1-7.

[12] Chicago Li, K., Guo, W., Zhang, W., Wen, Y., Li, C., & Hu, W. (2014, May). "QoE-based bandwidth allocation with SDN in FTTH networks". *In 2014 IEEE Network Operations and Management Symposium (NOMS)* (pp. 1-8). IEEE.

[13] Pfeiffenberger, T., Du, J. L., Arruda, P. B., & Anzaloni, A. (2015, July). "Reliable and flexible communications for power systems: fault-tolerant multicast with SDN/OpenFlow". *7th International Conference on New Technologies, Mobility and Security (NTMS) in 2015* (pp. 1-6). IEEE.

[14] Chen, J., Chen, J., Xu, F., Yin, M., & Zhang, W. (2015, November). "When Software Defined Networks Meet Fault Tolerance: A Survey". *In International Conference on Algorithms and Architectures for Parallel Processing (pp. 351-368). Springer International Publishing*.

[15] Kim, D., & Gil, J. M. (2015). "Reliable and Fault-Tolerant Software-Defined Network Operations Scheme for Remote 3D Printing". *Journal of Electronic Materials*, 44(3), 804-814.

[16] "Link Aggregation Control Protocol (LACP) (802.3ad) for Gigabit Interfaces." Online Available: http://www.cisco.com/c/en/us/td/docs/ios/12_2sb/feature/guide/gigeth.html. Accessed: 2016 − 08 − 27.

[17] "link aggregation according to IEEE standard 802.3ad online." Online Available:https://www.google.com.-bd/url?sa=trct=jq=esrc=ssource=webcd=3cad=rjauact=8ved=0ahUKEwjEk5GQ7aJ AhV XCI4KHR7pAjUQF gg3MAIurl = http, Accessed: 2015 − 11 − 20.

[18] "OMNeT++ Discrete Event Simulator" Online Available: https://omnetpp.org/, Accessed: 2016-09-25

[19] "EstiNet 9.0 | Simulator", Online Available: http://www.estinet.com/ns/?page_id=21140, Accessed: 2016-09-30

[20] "OFNet- Quick User Guide "Online Available: http://sdninsights.org/, Accessed: 2016-10-19

[21] "MaxiNet: Distributed Network Emulation", Online Available: https://maxinet.github.io/, Accessed: 2016-11-08

[22] " NS-3", Online Available: https://www.nsnam.org/, Accessed: 2016-11-08

[23] "Mininet: An Instant Virtual Network on your Laptop (or other PC) - Mininet" Online Available: http://mininet.org/. Accessed: 2016-06-15.

[24] F. Keti and S. Askar, "Emulation of Software Defined Networks Using Mininet in Different Simulation Environments," *6th International Conference on in Intelligent Systems, Modeling and Simulation (ISMS)*, 2015, pp. 205–210, IEEE, 2015.

[25] N. McKeown, T. Anderson, H. Balakrishnan, G. Parulkar, L. Peterson, J. Rexford, S. Shenker, and J. Turner, "OpenFlow: Enabling Innovation in Campus Networks," *ACM SIGCOMM Computer Communication Review*, vol. 38, no. 2, pp. 69–74, 2008.

[26] "osrg/ryu . github." Online Available: https://github-.com/osrg/ryu. Accessed: 2015-11-01

[27] Bonding, Online Available: http://www.linuxfoundation-.org/collaborate/workgroups/-networking/bonding, Accessed: 2016-06-15.

Water Quality Monitoring based on Small Satellite Technology

N. Gallah[1], O. b. Bahri[1], N. Lazreg[1], A. Chaouch[1]
[1]Microelectronic and Instrumentation Lab
[1]Faculty of sciences, University of Monastir
Monastir, Tunisia

Kamel Besbes[1,2]
[2]Center for Research on Microelectronics and Nanotechnology, CRMN
Sousse, Tunisia

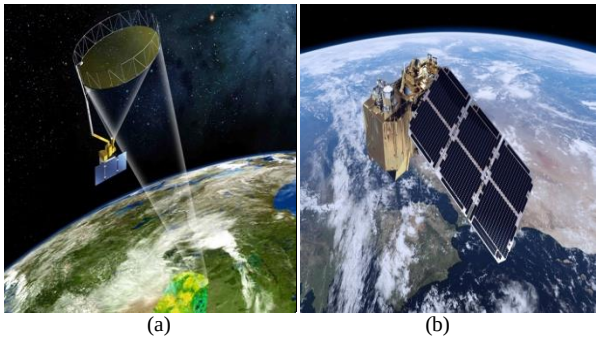
***Abstract*—In order to improve the routine of water quality monitoring and reduce the risk of accidental or deliberate contaminations, this paper presents the development of in-situ water quality monitoring and analysis system based on small satellite technology. A space mission design and analysis was performed in this work. The system consists of three segments; space segment including a constellation of nano-satellites, ground segment entailing the ground station and user-segment containing the in-situ water quality sensors. The authors studied the orbit characteristics and the number of nano-satellites required to cover the Middle East and North Africa (MENA) which presents the most water scarce region of the world. Thus, 9 nano-satellites distributed in 5 low earth orbits (600 km) are needed to cover MENA region continuously for about all the day. Data collected by various sensors such as pH and temperature are sent through Software Defined Radio (SDR) module which is responsible for the satellite communication.**

Keywords—space mission; nanosatellite; SDR; autonomous; on-line water monitoring

I. INTRODUCTION

According to the Intergovernmental Panel on Climate Change (IPCC, 2016), the Middle East and North Africa (MENA) region is projected to experience an increase of 3°C to 5°C in mean temperature and 20% decline in precipitation by the end of this century. The consequence of this climate change on water run-off is projected to drop by 20% to 30% in this region by 2050. In MENA region, further 83 million persons (about 27% of population) need to be supplied with safe water. Clearly, the implementation of carefully real-time and remotely water-quality monitoring systems is becoming necessary for large water distribution networks [1], to prevent accidental or deliberate contamination [2]. It is undeniable that water is an important element of economy and a vital resource for the world.

As an example of interest, the World Bank, USAID and NASA are collaborated for establishing a scientific program based on satellite earth observations (images, maps…) for water management in this region (fig. 1a). This initiative was launched in October 2011 and it is scheduled by the end of 2015 [3]. In addition, European Space Agency (ESA) launched satellites aiming for earth observation with the capability of inland water quality monitoring (fig. 1b)



(a) (b)

Fig. 1. Satellite earth observation. (a) NASA project (Courtesy of NASA[1]). (b) ESA project (image by ESA/ATG medialab[2])

However, the satellite earth observation systems present always gaps. The earth observation methods used for water quality monitoring are based on sensors resolution. Indeed, almost of this projects focus on ocean monitoring. The spatial resolution not suitable for lakes and rivers monitoring (water quality parameters measurement) that are need more accuracy. It is based on spectral response, thermal and scattering reflexion from water or by combination of spectral and microwave techniques. These methods need accuracy improvement in order to effectively fill observational gaps such as the determination of water quality parameters based on texture analysis method [4]. In addition, the launch of a satellite for earth observation purpose requires big budgets when good resolution and accuracy needed. The best way to improve the water quality monitoring is by combining data between the satellite earth observations and the in-situ measurement system.

The traditional method of in-situ measurement is based on water sampling which could become very expensive. It is in addition slow and unreliable due to the analysis time consuming in laboratory. Indeed, water-quality monitoring must be continuous and alerts should be sent whenever it's necessary. It becomes possible to properly register the annual evolution but also to foresee risks and to plan new development models. The number of scientific research and development systems is extremely large in this area. Thus, several researches focused on the development of in-situ water-quality measurement systems avoiding the traditional method.

[1] https://www.nasa.gov/smap
[2] http://www.odermatt-brockmann.ch/sponge/

The development of environmental sensing technology offers a promising application for future evolution [5], with different data sharing methods such as, Wireless Local Area Network (WLAN) [6], Wireless Sensor Network (WSN) [7-12], including different communication technologies such as Zigbee transmission method [13-16], or GPRS network [17-19].

These communication methods present a limitation to monitor our water resources on a large scale and in area without infrastructure, especially in developing countries. Nevertheless, Transboundary water issues will continue to plague the MENA region and others in terms of sharing data for better management of water resources. Indeed, the fast developments of the small satellite technology open a new era in this field. The nano-satellite technology can offer to the region and others an important contribution both to the autonomy of the management and sharing data related to water quality [20]. Furthermore, it can provide the international cooperation in the field of environment monitoring of our planet.

This work is one part of the small satellite mission for water telemetry and control. The entire mission consists of the cubesat constellation, the ground station and the in-situ water-quality sensor system. The main objective of this paper is the development of the in-situ water-quality measurement system reducing the mission cost. In this regard, a microcontroller board is used at node, which is a low cost device intentionally made as easy to program. Two important water quality parameters were measured: pH and temperature. The system developed in this work is reliable for adding other water quality sensors. Indeed, a wireless sensor network is proposed made up of set of sensor nodes. The topology and the techniques used in this network provide easy deployment of the nodes anywhere in the distribution water system without any particular limitation. The microcontroller unit (MCU) collects the information and sends them to the Software Defined Radio (SDR) module which is responsible for the communication with the nano-satellites such as HumSat constellation [21]. The SDR module adapts the signal received from the MCU to the appropriate protocol of communication and transfers it to the nano-satellite constellation. This network architecture is proposed in order to reduce the conflict and improve the communication between the sensor network and the nano-satellites. A prototype system was developed and tested at real water resources by measuring the pH and the temperature parameters continuously.

II. Entire Mission Description

Figure 2 illustrates the entire mission architecture. It consists of three main segments; the space segment, the ground and user segments. Indeed, water-quality monitoring has a promising application for the nano-satellite technology. Thus, the objective of the proposed system is to monitor water quality remotely and in a real-time. Therefore, the nano-satellite technology was considered as it is a better solution for wireless water-quality monitoring in zone without infrastructure. In addition, it facilitates sharing data regarding the transboundary water issues.

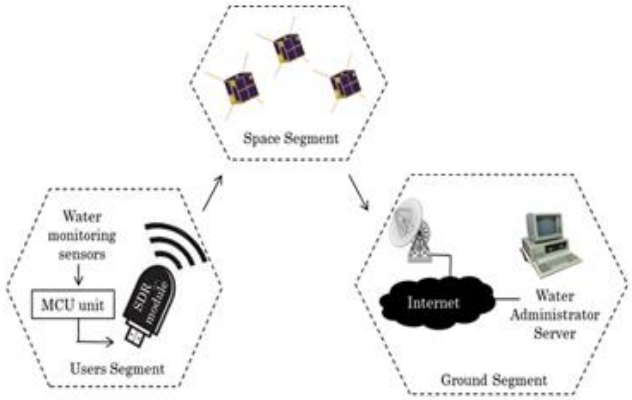


Fig. 2. Entire mission architecture

Accordingly, the integration of the small satellite technology in the routine of water-quality monitoring and management is very interesting. This work is one part of the general project which consists of three segments: space segment including a constellation of nano-satellites offered by the International University of space projects, ground segment entailing the ground stations included in radio amateur University Network such as GENSO [22] and user-segment which contains the in-situ multi-parameter water-quality sensors. The main contribution of this work is the development of a low-cost water-quality system. Thus, an MCU board with an Atmel microcontroller was used at node for transmitting data collected by sensors. The choice of the SDR technology is based on the entire project and the integration of the small satellite technology for water quality monitoring. A wireless sensor network which is based on USB-UHF Bridge was proposed based on SDR module capabilities. This bridge is a wireless communication device where the transmitter and receiver operations are changed or modified by software alone without making any (physical) changes to the hardware, which can reduce costs, improves reliability and allows to process real-time high bandwidth and simultaneously performs the decoding multiple emissions being in this band [23].

The communication scenario is listed below:

- In-situ water quality sensors collect the water parameters,
- The microcontroller board acquires these parameters and sends them through the SDR module,
- The USB-UHF Bridge receives data, adapts to the appropriate communication protocol and sends them to the nanosatellite constellation.

Finally, the data received from the nanosatellite through the ground station will be distributed through internet.

III. Nano-Satellite Constellation

The Middle East and North Africa is the region with the most water stress in the world. Therefore, the coverage of MENA region along about all the day was studied. In fact, one nano-satellite distributed in a low orbit, has coverage of

discontinuity for a day. Furthermore, it provides a short access unlike a satellite at high altitude. These consequences require many satellites in Low Earth Orbit (LEO) to provide continuous coverage of a specific area.

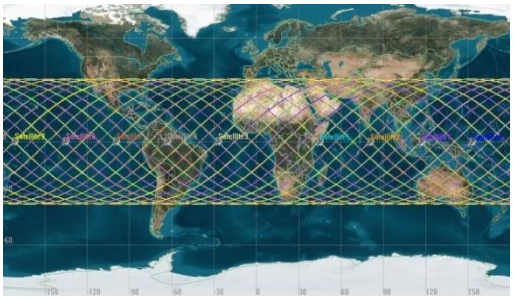


Fig. 3. Coverage of nano satellites constellation

TABLE I. ORBIT PARAMETERS

	Sat-1	Sat-2	Sat-3	Sat-4	Sat-5	Sat-6	Sat-7	Sat-8	Sat-9
a (Km)	600	600	600	600	600	600	600	600	600
e	0	0	0	0	0	0	0	0	0
I (°)	36	36	36	36	36	36	36	36	36
Ω (°)	0	0	72	72	144	144	216	216	288
TA (°)	0	180	0	180	0	180	0	180	0

In this part, the number of nano-satellites necessary to cover MENA region in a low orbit was investigated. Following Walker constellation architecture, 9 nano-satellites were chosen to accomplish this mission, which are distributed in 5 orbits as shown in table 1. With this approach, the MENA region can be continuously covered for about all the day as indicated in Fig.3 [24].

IV. GROUND SEGMENT

The ground segment consists of a ground station connected with the water administrator server which encompasses a network of water monitoring service.

The space-ground interface based on UHF band. The nano satellites constellation uses the UHF band for downlink with 9600 bps at frequency of 437 MHz. After being received across the ground station, the data are distributed through Internet (fig. 2). The VHF band was proposed for the uplink commands with 9600 bps at frequency of 145 MHz.

V. USER SEGMENT

Actually, there are several commercial multi-parameter sondes for water quality parameters measurement. These industrial instruments with high accuracy are very expensive, however, for the detection of water quality changes in water distribution system a less accurate instrument is enough. In this work, a simple cheap sensors system was chosen. It is a block made up two sensors. The measured variables are: temperature (k type thermocouple) and pH (combined electrode, KCL 3M electrolyte).

A. Hardware system architecture

The designed system is used to develop an application for real-time and remote water-quality monitoring in large volume water samples. In this paper, we focused on the user-segment. Figure 4 shows the bloc diagram of the segment. A microcontroller board was used at node to process sensor data, while the SDR module was used to send sensors information to the PC. The data was transmitted and received through the UHF band using two SDR modules. The first one was integrated with the user-segment as a transmitter, while the second one was integrated with a PC as a receiver, in order to test the proposed communication architecture based on a bridge type SDR module. Two commercial water-quality sensors were used to measure the pH and temperature parameters. The microcontroller board used is a tool for making small computers that can sense and control more things of materials as your desktop. It is a programmed electronic open-source platform that is based on Atmega328 microcontroller (AVR family). It is used at nodes which has 14 digital input/output pins, 6 analog inputs, 16 MHz crystal oscillator, a power jack, an ICSP header and a reset button. The features of this controller are 5 V operating voltage, flash memory of 32 KB, SRAM of 2 KB, EEPROM of 1KB. In addition, this board support UART, USB, SPI, and I2C communications. In this project, an SDR module is used to create a connection between the microcontroller board and a PC. The MCU receives sensors data collected by the water-quality sensors then sends them to the computer via UHF through the SDR modules. Accordingly, a software program was developed in order to acquire sensor data, convert to pH and temperature values and send them continuously.

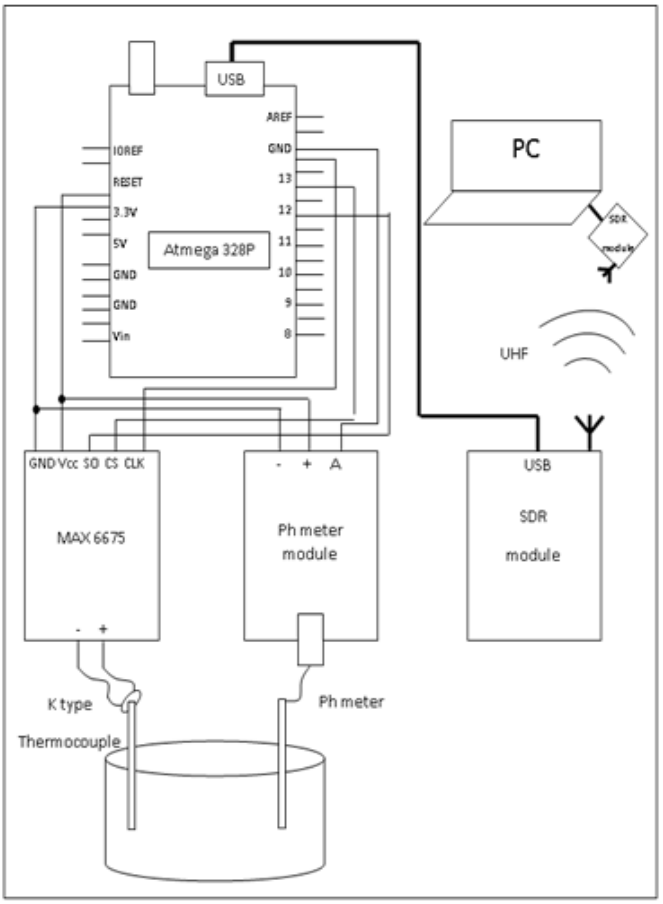


Fig. 4. Hardware system architecture

B. Temperature sensor

A K-type thermocouple is used to measure temperature. It's inexpensive, accurate, reliable and has a wide temperature range (-270 – 1260°C). The K-type thermocouple has an almost linear part between 0 and 1000°C with a see beck coefficient fluctuating around 40ΔV/°C. In our application, we focused only on the range of 0 – 100°C. The temperature interfacing circuit is shown in Fig. 5. The output voltage (mV) measured by the thermocouple in 0-100°C range is converted into 0-5 V range. The cold junction compensation is affected by the MAX6675. Figure 6 presents the characteristic curve of temperature.

C. pH sensor

High accuracy pH electrode is used as a pH sensor. It includes a measuring electrode which measures the potential related directly to the hydrogen ion concentration and a reference electrode which provides a stable potential to be compared to the measured electrode. The transfer function of the pH sensor is presented in Equation 1.

$$pH(x) = pH(s) + \frac{(Es - Ex)F}{RT\ln(10)} \quad (1)$$

where pH(x) is water solution's pH, pH(s) = 7, is the pH of standard solution, ES is the electric potential at reference, EX is the electric potential at pH measuring, F = 9.6485309*104 $Cmol^{-1}$, is the Faraday constant, R = 8.3145 J $K^{-1}mol^{-1}$, is the Universal gas constant and T is the temperature in Kelvin. The theoretical output is approximately 59.16 mV/pH at 25°C. Temperature affects the output voltage, thus it is required to compensate. The pH sensor used is the pH probe E201. This electrode operates in the full pH range from 0 to 14 and operates in the temperature range of 0° to 60°C. The sensor is terminated with BNC connector. Thus an adaptor was used to convert BNC to analog voltage to be acquired by the MCU pin.

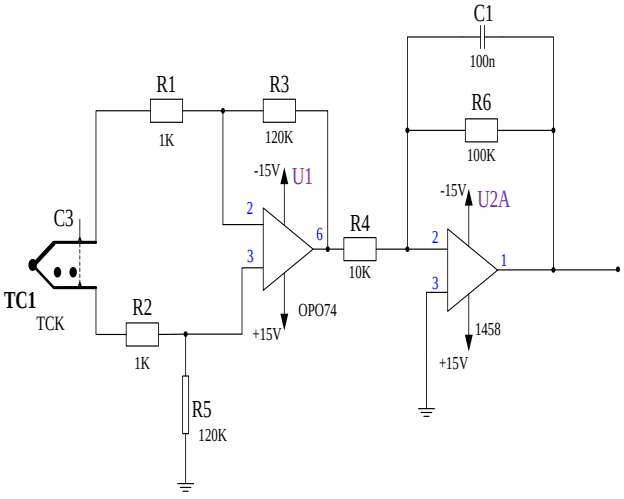


Fig. 5. Temperature interfacing circuit

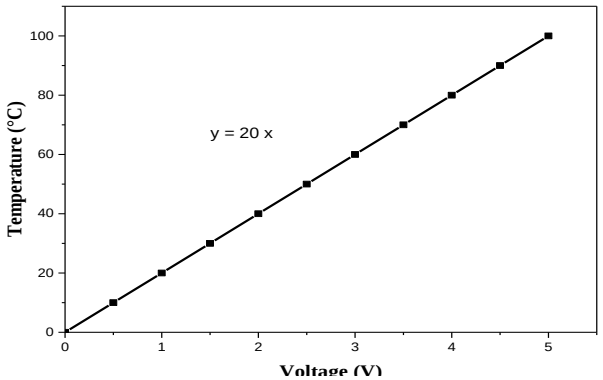


Fig. 6. Temperature characteristic curve

D. Software system architecture

The main purpose of this board is to acquire the analog value from the water quality sensors (pH and temperature) and transmit them through the SDR module. First, the MCU serial port connected with the MAX6675 and the pH adaptor circuits are initialized for acquiring the water quality parameters. Then, the microcontroller board converts the signals received from the sensors to pH and temperature values. If data is not completed, the MCU loop to acquire the sensor parameters. Upon converting data are completed, the MCU sends the information to computer via SDR module which is connected to this controller board. It sends data with a continuous manner. If the transmission data is not completed, it loop to acquire sensors data.

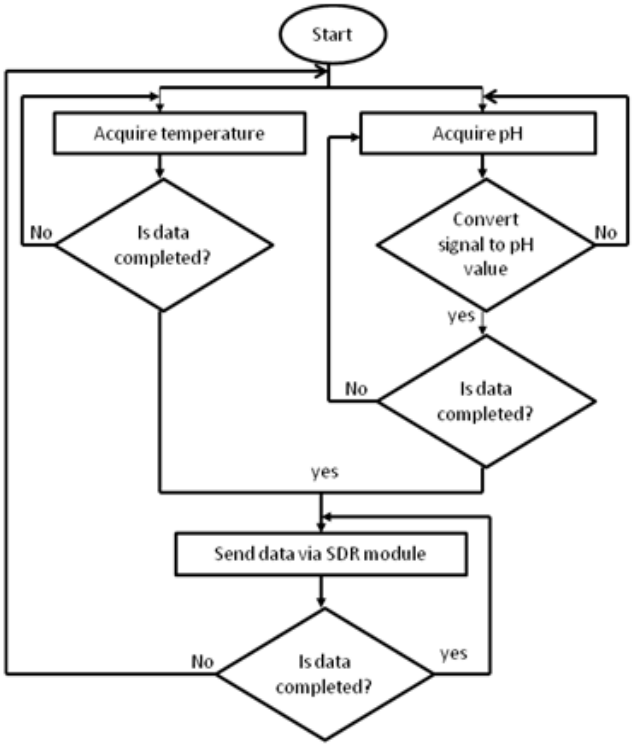


Fig. 7. Software system architecture

VI. Result and Discussion

The prototype version developed with the water-quality sensors is shown in Fig. 8. In order to test the prototype system, the pH and temperature sensors were emerged with another portable pH and temperature sensors on the same water samples with an accuracy of 0.05 and 0.5°C respectively. Thus, it allows the comparison between the data received through SDR module in computer and the direct data measured through the portable sensors. The comparison between the data received and the portable sensors of pH and temperature are shown respectively in Fig. 9 and 10. After acquiring data from the pH and temperature sensors, the MCU sends them with a continuous manner to computer. Figure 9 shows the pH values taking by the sensors system and the portable sensors during 10000 seconds. Except the first values which appear in large difference, the sensors system and portable sensors values difference are negligible as we progress in time axis. The response time of pH module cause the negligible accuracy difference between the pH sensors values. Figure 10 shows the temperature values from both the K-type thermocouple sensors which almost coincide with each other except that the sensors system data are deviated in time axis relative to portable sensor data.

The system cost including sensor devices, adaptor modules, MCU board and the SDR module stands at $270. In real world application, we need a high number of sensor nodes to monitor the water quality parameters on-line in large volume water (rivers, lakes, water distribution system). It is very important to reduce the cost of the sensor nodes which allows reducing the cost of the total system including the bridge USB-UHF, the nanosatellite constellation and the ground station. The system developed in this paper allows defining the delay time taken to transmit sensor data via the microcontroller board.

The best way for increasing the powerful of decision maker related to water management is by combining data between the satellite observations (images, maps…) and the in-situ water quality monitoring system (pH, temperature, conductivity, dissolved oxygen, turbidity…).

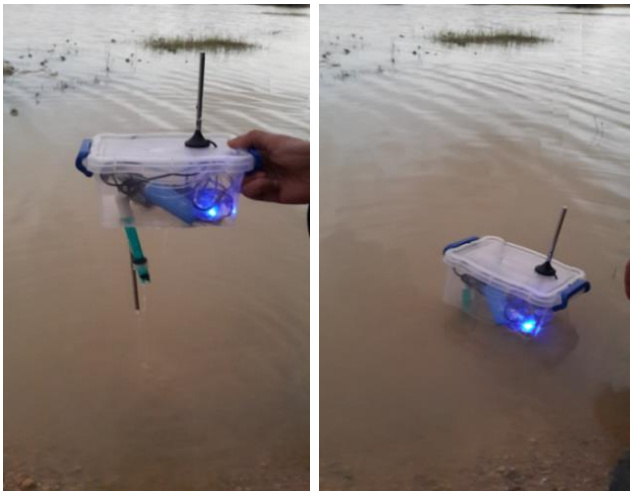


Fig. 8. Prototype of the system with microcontroller board, sensors and SDR module

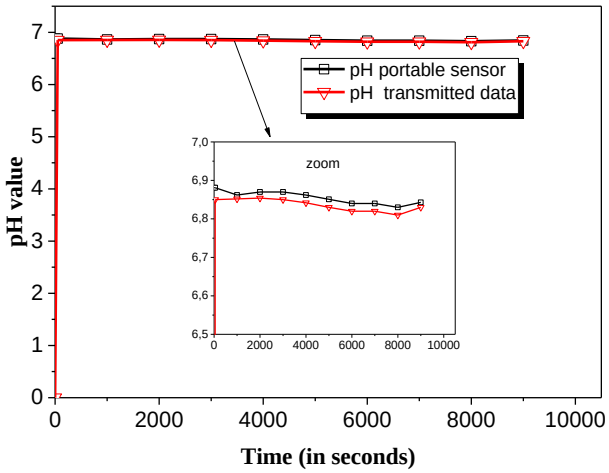


Fig. 9. pH parameters testing

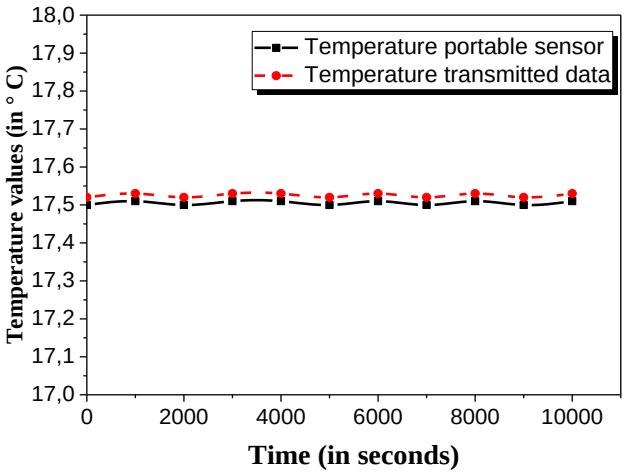


Fig. 10. Temperature parameters testing

Furthermore, an interface of this water quality system based on virtual instrument software was developed (Fig. 11). This interface is simple and clear for all users. It contains a start and exit module used to control the start and stop of monitoring system and graphs for measurement presentation of pH and temperature. An alarm will be trigger under the conditions discussed below:

- When the pH value is less than 6.5 or greater than 7.5, an alarm in the form of a red LED triggered to attract the attention of the users,
- When the temperature value exceeds 40°, an alarm in the form of a red LED triggered to attract the attention of the users.

The measured values are stored on specific measurement files, for future processing of data with other analysis software and to be published through browser to provide access to different users.

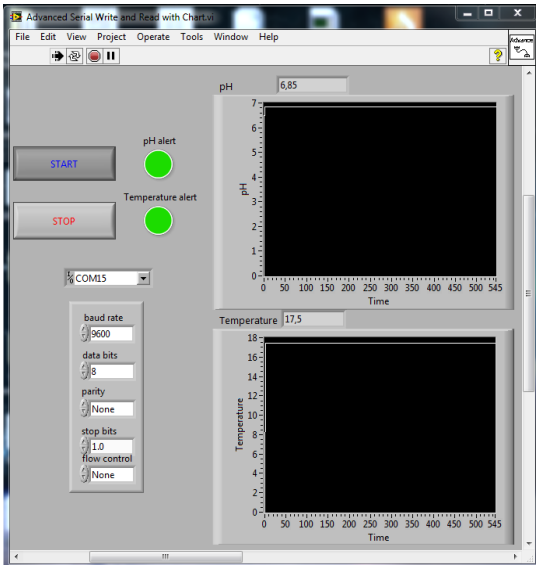


Fig. 11. Sensor measurement interface

VI. CONCLUSIONS

This paper presented a design of autonomous and in-situ water quality monitoring system based on SDR and small satellite technology. The main objective of this paper is to monitor the surface water quality in real time and remotely. Thus, a low cost system was developed which includes the multi-parameter sensors, acquisition card and the wireless sensor network based on SDR technology. A comparison was set between the results received from the developed system *and* the direct data measured through portable sensors which are found in a good agreement. In addition, we introduced the small satellite technology for water quality monitoring with new approach for a permanent coverage based on satellite constellation method. Future work will be related to the field of miniaturized water quality sensors.

REFERENCES

[1] Craun, G. F., Brunkard, J. M., Yoder, J. S., Roberts, V. A., Carpenter, J., Wade, T., …& Roy, S. L, "Causes of outbreaks associated with drinking water in the United States from 1971 to 2006", Clinical Microbiology Reviews, Vol. 23, no. 3, pp. 507-528, 2010.

[2] Clark, R. M., Geldreich, E. E., Fox, K. R., Rice, E. W., Johnson, C. H., Goodrich, J. A., …& Angulo, F. J, "A waterborne Salmonella typhimurium outbreak in Gideon, Missouri: Results from a field investigation", International Journal of Environmental Health Research, Vol. 6, no. 3, pp.187-193, 1996.

[3] Habib, S., Kfouri, C., & Peters, M, "Water information system platforms addressing critical societal needs in the MENA region", Geoscience and Remote Sensing Symposium (IGARSS), IEEE International conference, , pp. 2767-2770. July, 2012

[4] Muntadher, A. S., Abdelmalek, T. and Ali, K., "Estimation of Water Quality Parameters Using the Regression Model with Fuzzy K-Means Clustering" International Journal of Advanced Computer Science and Applications(IJACSA), Vol. 5, no. 6, 2014.

[5] Hart, J. K., & Martinez, K, "Environmental Sensor Networks: A revolution in the earth system science?", Earth-Science Reviews, Vol. 78 no. 3, pp. 177-191, 2006.

[6] Postolache, O. A., Girao, P. M. B. S., Pereira, J. D., & Ramos, H. M. G, 'Self-organizing maps application in a remote water quality monitoring system", IEEE transactions on instrumentation and measurement, Vol. 54, no. 1, pp. 322-329, 2005.

[7] Toran, F., Ramırez, D., Navarro, A. E., Casans, S., Pelegrı, J., & Espı, J. M, "Design of a virtual instrument for water quality monitoring across the Internet", Sensors and Actuators B: Chemical, Vol. 76, no. 1, pp. 281-285, 2005.

[8] Peixeiro, R., Postolache, O., & Dias Pereira, J. M, "Virtual instrument for water quality parameters measurement", Electrical and Power Engineering (EPE), International Conference and Exposition, , pp. 840-844. October 2012.

[9] Stamatescu, G., & Sgârciu, V, "Integration of wireless sensor networks with virtual instrumentation in a residential environment", arXiv preprint arXiv: 1305.6229, 2013.

[10] Simbeye, D. S., Zhao, J., & Yang, S, "Design and deployment of wireless sensor networks for aquaculture monitoring and control based on virtual instruments", Computers and Electronics in Agriculture, Vol. 102, pp. 31-42, 2014.

[11] Gómez, C., Aspas, J. P., & Herrero, J. E. C, "Sensors everywhere: wireless network technologies and solutions", Fundación Vodafone España, 2010.

[12] Ritter, C., Cottingham, M., Leventhal, J., & Mickelson, A, "Remote delay tolerant water quality montoring", Global Humanitarian Technology Conference (GHTC), IEEE, pp. 462-468. October, 2014.

[13] Du, Z. G., XIAO, D. Q., ZHOU, Y. H., & OU YANG, G. Z, "Design of water quality monitoring wireless sensor network system based on wireless sensor", Computer Engineering and Design, Vol. 17, 2008.

[14] Zhou, Y., Wen, D., Yuan, F., Li, J., & Li, M, "Research of Online Water Quality Monitoring System Based on Zibee Network", Advances in Information Sciences & Service Sciences, Vol. 4, no. 5, pp. 255-261, 2012

[15] Faustine, A., Mvuma, A. N., Mongi, H. J., Gabriel, M. C., Tenge, A. J., & Kucel, S. B, "Wireless Sensor Networks for Water Quality Monitoring and Control within Lake Victoria Basin: Prototype Development", Wireless Sensor Network, Vol. 6, no. 12,, p. 281, 2014.

[16] Prasanna, G. L., Prasad, S. R., Naidu, C. D., & Reddy, D. R, "Water Quality Monitoring And Controlling In Irrigation Using Zigbee Technology", International Journal of Science, Engineering and Technology Research, pp. 210-214. January, 2015.

[17] Lonel, R., Vasiu, G., &Mischie, S, "GPRS based data acquisition and analysis system with mobile phone control. Measurement", Vol. 45, no. 6, pp. 1462-1470. July, 2012.

[18] Xu, D., Li, D., Fei, B., Wang, Y., & Peng, F, "A GPRS-Based Low Energy Consumption Remote Terminal Unit for Aquaculture Water Quality Monitoring", Computer and Computing Technologies in Agriculture VII, pp. 492-503, 2014.

[19] Li, W., Pan, P., Tan, L. S., &Luo, X. K, "Remote On-Line Automatic Monitoring System of Reservoir's Water Regimen Based on WSN and GPRS Network", Applied Mechanics and Materials, Vol. 536, pp. 1223-1230. April, 2014.

[20] Gallah, N., & Besbes, K, "Small satellite and multi-sensor network for real time control and analysis of lakes surface waters", Recent Advances in Space Technologies (RAST), 6thIEEE International Conference, , pp. 155-158. June, 2013.

[21] Balogh, W, "Capacity building in space technology development: A new initiative within the United Nations programme on space applications", Space Policy, Vol. 27, no. 3, 2011, pp. 180-183. August, 2011.

[22] Leveque, K., Puig-Suari, J., & Turner, C, "Global Educational Network for Satellite Operations (GENSO)", 2007.

[23] T. Ulversøy, "Software defined radio: challenges and opportunities", IEEE Communications Surveys & Tutorials, Vol. 12, no. 4, pp. 531-550. 2010.

[24] Lazreg, N., & Besbes, K, "Design and architecture of Pico-satellites network for earth coverage", Advanced Technologies for Signal and Image Processing (ATSIP), 2nd International Conference on IEEE, , pp. 601-605. March, 2016.

Crowdsensing: Socio-Technical Challenges and Opportunities

Javeria Noureen
COMSATS Institute of Information Technology,
Sahiwal, Pakistan

Muhammad Asif
University of Lahore,
Pakpattan Campus, Pakistan

***Abstract*—With the advancement in mobile technology, the sensing and computational capability of mobile devices is increasing. The sensors in mobile devices are being used in a variety of ways to sense and actuate. Mobile crowdsensing is a paradigm that involves ordinary people to participate in a sensing task. This sensing model has the capability to provide a new vision of people centric sensing as a service. This research work reviewed different domains utilizing mobile crowdsensing for solving different domain specific problems. Mobile crowd sensing model is also posing different socio-technical challenges which needs to be addressed. The research work reviewed and analyzed a variety of socio-technical challenges of mobile crowdsensing and possible solutions presented by different studies. There are different socio-technical challenges but the challenge of privacy in crowdsensing requires extra measures to realize the vision of mobile crowdsensing.**

Keywords—Crowdsensing; sensing devices; privacy; smart phones

I. Introduction

The sensing capability of mobile devices is increasing day by day. The use of sensor enabled mobile devices is becoming ubiquitous. Researchers and engineers are seeking a variety of ways where sensing capabilities of mobile devices can be utilized. Mobile Crowdsensing (MCS) is an emerging sensing model which primarily depends on the strength of the people's sensor enable mobile devices to sense the data for a particular sensing task. Crowdsensing permits a huge number of sensing devices that share the collected data by the purpose to enumerate a phenomena of mutual interest [1]. Mobile devices are equipped with different sensors such as camera, GPS, digital compass, microphone, light sensor, accelerometer, and bluetooth as proximity sensor [2]. Crowdsensing empowers a large amount of mobile phones to be utilized for trading data among their clients, as well as for activities which might have an enormous societal impact. Mobile crowdsensing permits a large amount of mobile phone clients to share native knowledge (e.g., local information, ambient context, noise level, and traffic conditions) collected by their sensor-enhanced devices[3]. Mobile Crowdsensing has two distinct feature are as: 1) Implicit and explicit participation; 2) user-participant data sources [4] .

Mobile crowdsensing has various perspectives and defined in a variety of way as defined by Guo et al, *"a new sensing paradigm that empowers ordinary citizens to contribute data sensed or generated from their mobile devices, aggregates and fuses the data in the cloud for crowd intelligence extraction and people-centric service delivery" [5, 6]. The intrinsic nature of mobility in MCS allows a new and fast developing sensing model. It has the ability to acquire local knowledge through sensor-enhanced mobile devices – e.g., location, personal and surrounding context, noise level, traffic conditions, and in the future more specialized information such as pollution – and the possibility to share this knowledge within the social sphere, healthcare providers, and utility providers [5]*.

Mobile Crowdsensing (MCS) permits the large amount of cell phone clients share native knowledge such as (local information, ambient context, noise level, and traffic conditions) collected by their sensor-enhanced devices, and more information can be collects in the cloud for large scale sensing and community intelligence mining [7]. This model generally focus on the crowd powered data collection and processing [8][9].

Section II describes different application domains of crowdsensing and different socio-technical challenges are described in Section III. Privacy challenges and possible solutions are elaborated in Section IV and Section V concludes the paper.

II. Crowdsensing Applications Domains

Crowdsensing have different applications which is divided into three categories like (a) Infrastructure monitoring, (b) Social networking monitoring and, (c) Environmental monitoring [1]. In the infrastructure monitoring (Road monitoring, Traffic control/congestion, Road condition, and Individual travel planning and public transport) are further discussed. In Social networking monitoring (cinemas and historical places) and Environmental monitoring (natural environment, air pollution, walking, driving, level of water, wildfire habitats, noise pollution) are discussed [6].

A. Envirnmantal Monitoring

The crowdsensing paradigm is being utilized for environment monitoring, nature preservation, air pollution and many others. The Personal Environmental Impact Report (PEIR) project [10] utilize sensors in mobile phones to construct a framework which allows customized environmental effect reports, which follow how the activities of people's affect both their experience and their impact to troubles. The objective of the project was to evaluate the effect of individual user/public participation to observe the environment like contamination, climate and noise tracking and so on. Noise pollution creates problems in fitness and in quality of life, quoting high blood pressure, hearing damage,

frustration, and others [11]. The European Commission mandates the generation of noise to collect data and create noise maps. Yet, the government efforts are limited because the deployed sensing nodes cannot protect all regions of the city. A noise map is a graphic demonstration of the sound level distribution. To create a noise map, shared measurements were used. In their daily lives, NoiseTube [12] could measure personal exposure to environmental noise. EarPhone [13] was also a participatory noise mapping system. The END (European Noise Directive) [14] states environmental noise such as *"unwanted or harmful outdoor sound created by human activities, including noise emitted by means of transport, road traffic, rail traffic, air traffic, and from sites of industrial activity"*.

Mobile phones were also used to collect the information of on the road diesel trace to study community exposure to urban air pollution [15]. ExposureSense [16] explored the integration of Wireless Sensor Network and participatory sensing paradigms for personal air quality exposure measurement. The BikeNet application [17] could measure CO_2 level and also report the path of a cyclist activity.

B. Transportation and traffic planning

The traffic congestion remains a serious global problem; for example, congestion alone could impact both the environment and human productivity (e.g., wasted hours due to congestion) [3]. As GPS based vehicles which is equipped with computers travels, it periodically records the present time and location and use wireless network to send information to a server. GPS receiver on mobile phone can provide the location information. Wi-Fi can also be used to send data to a nearest wireless get to point. Traffic deferrals and congestion are a prime cause of disorganization, squandered fuel, and commuter frustration [18].

To report the road and traffic condition, mobile phones can be utilized. In Nericell [19], different embedded devices such as accelerometer, microphone, and positioning system being utilized to identify as well as focus on transportation and road situations, for example quality of road (potholes, bumps), and driving behavior (braking and honking or beeping) [20]. A potholes [21] application can find fleabags in streets using the crowdsourced shaking and position information collected from smart phones. VTrack [18] was a system that used mobile phones to correctly estimate the traffic time between different locations. WreckWatch [22]removing the interruption among accident occurrence and primary responder dispatcher and automatically detect the accidents and send the notifications to a server. T-Share [23] was a taxi ridesharing service that can produce optimized ridesharing schedules based on crowd-powered data.

C. Social Networking Monitoring

Social networks are popular way of communications with other who are members of the same social networking application and share information between the social groups [24]. Social media (i.e. Twitter, Facebook, MySpace, and LinkedIn) are used for communication. Millions of people take part frequently within online social networks and share their views, their ideas about any subject. Social sensing system used to get and share social information among friends, social clusters and communities [25]. There are two kinds of social sensing like implicit sensing and explicit sensing. In implicit social sensing always concerns on e-business sites line Amazon [26] which evaluate the purchasing behavior of their customers. While explicit social sensing concerns the existing study concentrates on the very famous tools for example, Flicker, Twitter and Facebook [27]. The Dartmouth CenceMe [28] development is *examining the utilization of* sensors in the mobile phone to mechanically categorize actions in *individuals'* existence, this known as *sensing existence.*

D. Health Care and Public Safety

In health care, public health and personal health is monitored. Mobile crowdsensing can be used to monitor the different diseases. Personal sensing systems collects people's data to monitor their health, routine life and physical activities such as heart rate, blood pressure, sugar level etc. [25]. Sensor-enabled cell phones utilized for the observation of physiological condition and well-being of patients utilizing inserted or exterior sensors such as wearable accelerometers, or air contamination sensors [29]. The DietSense [30] support the people who wants to lose their weight. This system allows the people to report or share their food choices via pictures and sound sample to get suggestions from online experts for weight loss. HealthAware [31] is also a similar kind of system that persuade people to participate in improving health through people centric feedback.

Public safety is about detecting or protecting the citizens from the events (e.g., crimes, disasters) that could be danger for the safety of the citizens. By evaluating the large number of geo tagged Twitter messages posted from mobile devices, Lee [32] proposed a method to detect the curious crowded spaces such as (a terrorist activity). SAIS (Smartphone Augmented Infrastructure Sensing) [33], also confirm public security in smart cities utilizing maintainable design. SAIS collects data from citizens and authorities for the security actions, using this information a dashboard smart phone application is produced which it helps each other to make better situation-awareness. Participatory mobile phone sensing systems can also be used for helping disaster relief [34]. Large-scale mobile phone can analyze the user data before and after earthquake movement behaviors, they construct a model and this model could predict community responses to future disasters. Similarly, in [35] Twitter could give near real-time report of earthquakes region by observing geo tagged user posts.

III. CROWDSENSING CHALLENGES

Crowdsensing has many challenges in addition to privacy and security challenges [36]. We focus on the social and technical challenges and we also outline general solutions. Some are as follows:

- Local analytics is key challenge in discovering searching and designing algorithms is to accomplish the imaginary function. Data mediation is one of the class of functions, for example clarifying of outliers, noise exclusion, or covering data gaps. For instance, GPS sample cannot be able to obtain correct or missing

(because of absence of observable pathway), in which occasion outliers ought to be eliminating or omitted specimens extrapolated.

- MCS applications depend on the examining data from accumulation of mobile devices, distinguishing spatial temporal designs. When a physical or social phenomenon is being observed these designs could helpful for constructing patterns. The challenge in recognizing designs from huge amounts of information is normally application- specific. It also contains data mining algorithms.
- Data storage in database is one of the conventional techniques of data mining. For the patterns detection different mining algorithms can be used for implementation against database.
- If sum of consistent data participation is excessively for storage, or application needs fast detection of patterns, stream data mining algorithms might be essential. Such algorithms take continuous data streams as an input and detect patterns, without the requirement of the first store of the data.
- The 3-tier system architecture [37] also have some challenges are as follows: Virtualization Overhead is the main challenge in system architecture.
- Configuration and performance is another challenge of inter-VM communication. Inter-VM communication performance is comparatively low when it is compare to inter-process communication.
- Migration-induced Reconfiguration is likewise challenges. With constraints of Non IP-based results , the Host Identify Protocol [38] are intended to scrape mind, still such protocols are essential for the evaluation of real networks.
- Standardization of Sensing Interfaces is a challenge.
- Different crowdsensing applications can construct similar sensor data, but use diverse system or model rate.
- Another challenge is how to provide valuable incentive mechanisms that allows honest contributions in mobile crowdsensing and computing becomes a critical challenge [6]. Recently, numerous game theory approaches [39-41] have been proposed for mobile crowdsensing and computing to encourage and reward truthful contributions. These game theory techniques are usually based on auction mechanisms, however slightly complex to apply in a fully distributed and time evolving system. Therefore, for a highly dynamic mobile crowdsensing and computing system, there is still need for new incentive and pricing mechanisms to attract, inspire, and reward truthful and high-quality sensing data contributors.
- Data delivery in transient network is a challenge in mobile crowdsensing, how to dispatch the sensed data from distributed participants to the backend server is another challenge because of an assortment of mobile crowdsensing and computing characteristics, for example the low bandwidth of wireless communication, recurrent network apportioning due to human mobility, and huge number of energy-constrained devices. Whereas this is also a well-known research challenge in both wireless sensor networks and general mobile systems.
- We have to consider an essential issue to mobile phone sensing (still no need of great tended): provided a block of focuses key steps or a focuses area, a set about mobile phones and a time limit, we discover a sensing schedule (which identifies sensing for each mobile phone) so aggregate energy utilization will minimize to subject to a coverage restriction [42]. Scheduling algorithms can solve this trouble and used sensing servers to arrange sensing events of mobile phones (an incentive mechanism use recruited). Note that opportunistic sensing applications will only use the scheduling algorithms; later on, in participatory sensing applications; mobile phone users control sensing task by manually.
- Since a more specialized point of view, one of the important difficulties is discovering a decent harmony between framework versatility and detecting accuracy for far reaching sending situations. In such another socio-specialized framework, the sorts of assets are overall different, crossing from figuring ones (system transmission capacity, memory, CPU, and so on.) to people (numeral individuals included, human consideration, specific aptitudes to contribute, and so on.), with these lines, it is difficult to entirely control them.
- Finally, the trade-offs are similar to the trade-offs that occur when using an ad hoc network instead of a fixed infrastructure network: it is easier to install and could be used in areas where establishing a fixed infrastructure is difficult but introduces the other complexities and challenges.
- The effect of scaling sensing applications from individual to resident's scale is ambiguous. Numerous problems identified with majority of the information exchanging, privacy, mining of data, and closing loop by given the meaningful feedback to a person, clusters, society, and people stay open. For constructing scalable sensing systems we just limited experience [6].
- Different sensing scale types are used, which ranges users that are actively includes in sensing system [43]. Author focus on two facts such as configuration space: participatory sensing, where individuals efficiently involves in data gathering events (i.e., the individuals physically decides how, what, and where to model) and opportunistic sensing, where information gathering phase is completely automatic and user has no participation.

- Opportunistic sensing has one key challenge which is phone context problem. For example, an application needs to simply take a sound pattern for a city extensive noise map whereas cell phone has not available in pocket or bag. Context problems can be solve by using different mobile phone sensors. These sensors such as accelerometer or light sensors determine whether mobile phone has not available in pocket.
- Participatory sensing is acquisition interest in cell phone sensing society, puts an extreme load or cost on individual; such as, manually select information to gather (e.g., least prices of petrol) and then model it (e.g., getting an image). Participatory sensing has one drawback, in which data quality is totally based on applicant interest to consistently gather sensing information and the similarity of an individual's flexibility pattern to the expected objectives of the application (e.g., get smog patterns in the region of schools).

IV. CROWDSENSING PRIVACY

Privacy is very important for everyone. No one wants to reveal his /her privacy in front of anyone. We can use different techniques to provide privacy to mobile devices or nodes. Here some overheads and risks are discussed. We also discuss privacy techniques, how these methods used in current sensing applications that address these issues. We also describe some solution of these overhead and risks. Data collection infrastructure layer is use to collect information from the chose sensor nodes. It gives information to data contributors along with privacy preserving techniques. Some component such as task allocation, sensor gateways, data anonymization, incentive mechanism and big data storage are used in this layer, which collected data from the selected nodes [6]. Author in [1] describes different privacy methods to protect our privacy, these methods are Anonymization, Encryption, and Data Perturbation.

1) Anonymization: Anonymization approach removes the identification information, which is collected through Crowdsensing applications. The anonymization of data will increase the privacy safeguard but is reduce data utility. Anonymization approach has two further techniques to preserve the privacy such as pseudonyms and connection anonymization. *Pseudonyms:* it is the simple technique that makes participants anonymous by replacing their identification information with an alias [44]. *Connection Anonymization:* Using this technique, we can avoid the network based tracking attacks using IP addresses. One such technique which is used in Crowdsensing applications [45] is onion routing [46].

2) Encryption: Encryption is a technique in which the illegal third parties not allowed to utilize the private data of mobile users. In encryption large volume of data required significant resources for encryption.

3) Data Perturbation: To preserve the privacy of individuals' data perturbation, immediately increase noise to the sensor data before distributing it with the group of people, that sensory data will be unidentifiable. However, such information empowers excellent process of Crowdsensing applications.

Providing privacy anonymous routing technique is used such as onion routing [47] in a decentralized mobile cloud. For example, exist in peer-to-peer domain [48-50]. However, there exist certain outflows and a risk of unreliable delivery connected with most anonymous peer-to-peer routing protocols [51]. As a solution, the degree of security and anonymity must be flexible and depend on the context. For example, the capability of malicious nodes is high and these nodes should have the high level of privacy but this would bring on higher transmission (i.e. longer path) and computation costs (i.e. cryptography processing overheads) [52]. Cryptography technique is used to transmute data to preserve the privacy. Another privacy preserving approach which is secure multiparty computation [53], in which cryptographic techniques are used to transfer data to preserve the security. Cryptographic approaches used for calculating intensively, not versatile. Cryptographic require maintenance and generation of numerous keys. Likewise prompts top vitality utilization which participant uploads their reports, k-anonymity technique can be used to provide location privacy. [54]. Basic idea behind the k-anonymity is to construct clusters of k applicants or reports. In this way they share common feature (e.g., k participants situated in the alike region), interpreting them indistinct to everyone. To build a group of k users we can use different methods to find the suitable and common attribute. So these methods categorized into two main sections such as generalization and perturbation [55] .

V. PRIVACY AND SECURITY CHALLENGES/THREATS

A sensor device may be used by the user to report a false data. There may be chances of location and time bias when the data is sensed. Additionally, the readiness and accessibility of setup is very important for these applications to be use full. There are some privacy and privacy challenges or risk where campaign administrator breaks the trust among participants and reveals the sensitive data about participants [44].

- *Time and location***:** HealthSense gather the information about time and place freely of their people environmental driven nature. So GPS receivers which is embedded in the smart phones provide vary accurate location of the user. So, within the absence of GPS, WiFi or cellular system depends mostly triangulation which utilized to get coarse-grained area data [57]. Through embedded sensors contextual information can be used to recognize a person location [58]. Moreover, the threats ensuing from time and location traces aren't confined to applications, wherever authentication is needed [59].
- *Sound samples:* Besides deriving personalities and preference form temporal and spatial data, the representation of participants clarified through completing this information by patterns of other detecting modalities. In a few of the previously stated applications, patterns if sound either recorded

purposefully by the users, or automatically caught through cell phones. However, participants simply secure their protection just recording non sensitive occasions in the previous instance; cell phones efficiently act as intelligent spies in situation of automatic recordings.

- *Picture and video:* The substance of distributed picture and recorded recordings is also probable to disclose personal information related to the members and their environments. While Diet- Sense [31] targets to take photos of consumed meal, no countermeasures are taken to cover the faces of person's share-out their meal with the participants. In entirely situations, in which the camera is arranged far from the participant, faces of other persons in the region are conceivably caught in the pictures, and consequences about the number and identity of the participant's social relations might be drawn. The publications of capturing images can lead to the alike conclusions as in online social networks, such as Facebook, where an instructor was suspended because of a photo demonstrating her holding glasses loaded with liquor [60], or a disheartened woman who lost benefits from her health insurance for images demonstrating her presence parties and relaxing on the beach [61]. Alike to sound recordings, the existing user context and their nearby environment could also be extracted from sensor data. For example, images displaying points of interest could easily found the participant's attendance at those locations.
- *Acceleration:* learning of raw accelerometer might show up fewer threatening in exposing personal data about the members. Yet this theory not every times correct also might regularly just help as an incorrect sense about safety. For example, if the mobile phone is carry on hip, data about the walk, in this way through conceivable indication about the user can inferred the identity of user [62]. Moreover, the study of action recognition also creates wide usage of accelerometer analyses [63]. The misuse of this data by pernicious clients might have undesirable outcomes. For instance, employers might need to confirm that their employees are really doing work throughout their working times. If employers discover any abnormalities. Could terminate the respective worker.
- *Environmental Data:* Recording gas and particles focuses or barometric burden might not be straightforwardly undermines protection of members without anyone else's input. Be that as it may, specific air pieces joined with optional data, for example, exact air temperature, may distinguish the area of the members at the phase of granularity as well as room levels inside structures, where area information can be wrong because of area administrations or non-availability of GPS.
- *Biometric Data:* Sensor data of biometric utilized for discovering the user's present physiological condition. Likewise, to medical staff, opponent may distinguish health irregularities or diseases built on the caught sensor information. Revealed medical data later used for health insurance corporations or employers to repudiate agreements, if diagnosis any damage of physiological states of the participants.

Another approach which produces a privacy threat to reveal the location information, forgetting frequently visited locations of individual's anonymized GPS sensor measurements is still used. It also used to get personal information of participants. PEIR [10] utilizes sensitive private information, and its schemes should planned to reduce information release from the user's control to prevent different security threats [64-67]. In different cases, getting features might be very sensitive than crude data: researchers discovered crucial places like home and offices. It also gives extremely valuable context information and can learn by inference methods, but discharging this information unnecessarily and should be preventable.

Applications such as PEIR should help the participants in linking sensing information to their everyday practices and long term aims. To this end, investigation, collection, and alerting specifically absolutely should be configurable as well as information and high level of conclusions navigable from various points of view. Support (1) comprehension of specific samples, (2) personal progression of research and substitutes, (3) setting aim, (4) comparison, (5) comprehension of marvels relationship [68]. In participatory detecting, the detecting operations needs some method for human impedance [69] e. g the assignment could request that the client take a photo of the menu when she visits an eatery or to remark on her view about the sustenance at the cafeteria or to gauge the gas costs when she passes a recording station [70]. The human components incorporate extra security assignments. As for security, the client may undertake additional data about his or her character by the way of his or her response.

VI. CONCLUSIONS

Mobile crowdsensing is an emerging sensing model based on participatory sensing paradigm. This paper describes different concepts of crowdsensing and how it is applied in different domains so far. Crowdsensing has the potential to produce interesting business models such as sensing as a service. This participatory sensing paradigm has many socio-technical challenges and major is a privacy. However, it requires innovative approaches to solve the socio-technical challenges.

REFERENCES

[1] Dimov, D., Crowdsensing: State of the Art and Privacy Aspects. InfoSec Institute, 2014. 29.

[2] Mardenfeld, S., et al. Gdc: Group discovery using co-location traces. in Social computing (SocialCom), 2010 IEEE second international conference on. 2010. IEEE.

[3] Lane, N.D., et al., A survey of mobile phone sensing. Communications Magazine, IEEE, 2010. 48(9): p. 140-150.

[4] Talasila, M., R. Curtmola, and C. Borcea, Mobile Crowd Sensing.

[5] Guo, B., et al. From participatory sensing to mobile crowd sensing. in Pervasive Computing and Communications Workshops (PERCOM Workshops), 2014 IEEE International Conference on. 2014. IEEE.

[6] Guo, B., et al., Mobile crowd sensing and computing: The review of an emerging human-powered sensing paradigm. ACM Computing Surveys (CSUR), 2015. 48(1): p. 7.

[7] Zhang, D., B. Guo, and Z. Yu, The emergence of social and community intelligence. Computer, 2011. 44(7): p. 21-28.

[8] Pournajaf, L., et al., A survey on privacy in mobile crowd sensing task management. 2014, Technical Report TR-2014-002, Department of Mathe-matics and Computer Science, Emory University.

[9] Sirsikar, S. and V. Powar, Mobile Crowd Sensing Using Voronoi Based Approach. International Journal of Computer Science And Applications, 2015. 8(1).

[10] Mun, M., et al. PEIR, the personal environmental impact report, as a platform for participatory sensing systems research. in Proceedings of the 7th international conference on Mobile systems, applications, and services. 2009. ACM.

[11] Stansfeld, S.A. and M.P. Matheson, Noise pollution: non-auditory effects on health. British medical bulletin, 2003. 68(1): p. 243-257.

[12] Maisonneuve, N., M. Stevens, and B. Ochab, Participatory noise pollution monitoring using mobile phones. Information Polity, 2010. 15(1, 2): p. 51-71.

[13] Rana, R.K., et al. Ear-phone: an end-to-end participatory urban noise mapping system. in Proceedings of the 9th ACM/IEEE International Conference on Information Processing in Sensor Networks. 2010. ACM.

[14] Directive, E., Directive 2002/49/EC of the European parliament and the Council of 25 June 2002 relating to the assessment and management of environmental noise. Official Journal of the European Communities, 2002. 189(12).

[15] Goldman, J., et al., Participatory Sensing: A citizen-powered approach to illuminating the patterns that shape our world. Foresight & Governance Project, White Paper, 2009: p. 1-15.

[16] Predic, B., et al. Exposuresense: Integrating daily activities with air quality using mobile participatory sensing. in Pervasive Computing and Communications Workshops (PERCOM Workshops), 2013 IEEE International Conference on. 2013. IEEE.

[17] Eisenman, S.B., et al., BikeNet: A mobile sensing system for cyclist experience mapping. ACM Transactions on Sensor Networks (TOSN), 2009. 6(1): p. 6.

[18] Thiagarajan, A., et al. VTrack: accurate, energy-aware road traffic delay estimation using mobile phones. in Proceedings of the 7th ACM Conference on Embedded Networked Sensor Systems. 2009. ACM.

[19] Mohan, P., V.N. Padmanabhan, and R. Ramjee. Nericell: rich monitoring of road and traffic conditions using mobile smartphones. in Proceedings of the 6th ACM conference on Embedded network sensor systems. 2008. ACM.

[20] Higuchi, T., H. Yamaguchi, and T. Higashino, Mobile devices as an infrastructure: A survey of opportunistic sensing technology. Journal of information processing, 2015. 23(2): p. 94-104.

[21] Eriksson, J., et al. The pothole patrol: using a mobile sensor network for road surface monitoring. in Proceedings of the 6th international conference on Mobile systems, applications, and services. 2008. ACM.

[22] White, J., et al., Wreckwatch: Automatic traffic accident detection and notification with smartphones. Mobile Networks and Applications, 2011. 16(3): p. 285-303.

[23] Ma, S., Y. Zheng, and O. Wolfson. T-share: A large-scale dynamic taxi ridesharing service. in Data Engineering (ICDE), 2013 IEEE 29th International Conference on. 2013. IEEE.

[24] Finucan, R., Mobile Social Networking. 2009, Google Patents.

[25] Khan, W.Z., et al., Mobile phone sensing systems: A survey. Communications Surveys & Tutorials, IEEE, 2013. 15(1): p. 402-427.

[26] Ziegler, C.-N., G. Lausen, and J.A. Konstan, On exploiting classification taxonomies in recommender systems. AI Communications, 2008. 21(2-3): p. 97-125.

[27] Rosi, A., et al. Social sensors and pervasive services: Approaches and perspectives. in Pervasive Computing and Communications Workshops (PERCOM Workshops), 2011 IEEE International Conference on. 2011. IEEE.

[28] Miluzzo, E., et al. Sensing meets mobile social networks: the design, implementation and evaluation of the cenceme application. in Proceedings of the 6th ACM conference on Embedded network sensor systems. 2008. ACM.

[29] Kanhere, S.S., Participatory sensing: Crowdsourcing data from mobile smartphones in urban spaces, in Distributed computing and internet technology. 2013, Springer. p. 19-26.

[30] Reddy, S., et al. Image browsing, processing, and clustering for participatory sensing: lessons from a DietSense prototype. in Proceedings of the 4th workshop on Embedded networked sensors. 2007. ACM.

[31] Gao, C., F. Kong, and J. Tan. Healthaware: Tackling obesity with health aware smart phone systems. in Robotics and Biomimetics (ROBIO), 2009 IEEE International Conference on. 2009. Ieee.

[32] Chen, W., et al., A survey and challenges in routing and data dissemination in vehicular ad hoc networks. Wireless Communications and Mobile Computing, 2011. 11(7): p. 787-795.

[33] Liao, C.-C., et al. Sais: Smartphone augmented infrastructure sensing for public safety and sustainability in smart cities. in Proceedings of the 1st International Workshop on Emerging Multimedia Applications and Services for Smart Cities. 2014. ACM.

[34] Consulting, V.W., mHealth for development: the opportunity of mobile technology for healthcare in the developing world. 2009.

[35] Sakaki, T., M. Okazaki, and Y. Matsuo. Earthquake shakes Twitter users: real-time event detection by social sensors. in Proceedings of the 19th international conference on World wide web. 2010. ACM.

[36] 36 Ganti, R.K., F. Ye, and H. Lei, Mobile crowdsensing: current state and future challenges. Communications Magazine, IEEE, 2011. 49(11): p. 32-39.

[37] Heggen, S., A. Adagale, and J. Payton, Lowering the barrier for crowdsensing application development, in Mobile Computing, Applications, and Services. 2013, Springer. p. 1-18.

[38] Nikander, P., A. Gurtov, and T.R. Henderson, Host identity protocol (HIP): Connectivity, mobility, multi-homing, security, and privacy over IPv4 and IPv6 networks. Communications Surveys & Tutorials, IEEE, 2010. 12(2): p. 186-204.

[39] Huangfu, S., et al. Using the model of markets with intermediaries as an incentive scheme for opportunistic social networks. in Ubiquitous Intelligence and Computing, 2013 IEEE 10th International Conference on and 10th International Conference on Autonomic and Trusted Computing (UIC/ATC). 2013. IEEE.

[40] Lee, J.-S. and B. Hoh. Sell your experiences: a market mechanism based incentive for participatory sensing. in Pervasive Computing and Communications (PerCom), 2010 IEEE International Conference on. 2010. IEEE.

[41] Yang, D., et al. Crowdsourcing to smartphones: incentive mechanism design for mobile phone sensing. in Proceedings of the 18th annual international conference on Mobile computing and networking. 2012. ACM.

[42] Sheng, X., et al., Sensing as a service: Challenges, solutions and future directions. Sensors Journal, IEEE, 2013. 13(10): p. 3733-3741.

[43] Smartphone Sensing Group. September 2013.

[44] Christin, D., et al., A survey on privacy in mobile participatory sensing applications. Journal of Systems and Software, 2011. 84(11): p. 1928-1946.

[45] Shin, M., et al., AnonySense: A system for anonymous opportunistic sensing. Pervasive and Mobile Computing, 2011. 7(1): p. 16-30.

[46] Dingledine, R., N. Mathewson, and P. Syverson, Tor: The second-generation onion router. 2004, DTIC Document.

[47] Syverson, P.F., D.M. Goldschlag, and M.G. Reed. Anonymous connections and onion routing. in Security and Privacy, 1997. Proceedings., 1997 IEEE Symposium on. 1997. IEEE.

[48] Han, J., et al. Provide privacy for mobile p2p systems. in Distributed Computing Systems Workshops, 2005. 25th IEEE International Conference on. 2005. IEEE.

[49] Han, J. and Y. Liu. Rumor riding: Anonymizing unstructured peer-to-peer systems. in Network Protocols, 2006. ICNP'06. Proceedings of the 2006 14th IEEE International Conference on. 2006. IEEE.

[50] Ghinita, G., P. Kalnis, and S. Skiadopoulos. PRIVE: anonymous location-based queries in distributed mobile systems. in Proceedings of the 16th international conference on World Wide Web. 2007. ACM.

[51] Han, J. and Y. Liu, Mutual anonymity for mobile p2p systems. Parallel and Distributed Systems, IEEE Transactions on, 2008. 19(8): p. 1009-1019.

[52] Fernando, N., S.W. Loke, and W. Rahayu, Mobile cloud computing: A survey. Future Generation Computer Systems, 2013. 29(1): p. 84-106.

[53] Yao, A.C. Protocols for secure computations. in Foundations of Computer Science, 1982. SFCS'08. 23rd Annual Symposium on. 1982. IEEE.

[54] Sweeney, L., k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2002. 10(05): p. 557-570.

[55] Huang, K.L., S.S. Kanhere, and W. Hu, Preserving privacy in participatory sensing systems. Computer Communications, 2010. 33(11): p. 1266-1280.

[56] Shi, J., et al. Prisense: privacy-preserving data aggregation in people-centric urban sensing systems. in INFOCOM, 2010 Proceedings IEEE. 2010. IEEE.

[57] LaMarca, A., et al., Place lab: Device positioning using radio beacons in the wild, in Pervasive computing. 2005, Springer. p. 116-133.

[58] Azizyan, M., I. Constandache, and R. Roy Choudhury. SurroundSense: mobile phone localization via ambience fingerprinting. in Proceedings of the 15th annual international conference on Mobile computing and networking. 2009. ACM.

[59] Shilton, K., Four billion little brothers?: Privacy, mobile phones, and ubiquitous data collection. Communications of the ACM, 2009. 52(11): p. 48-53.

[60] Did the Internet Kill Privacy? 2011.

[61] Depressed woman loses benefits over Facebook photos. 2009.

[62] Derawi, M.O., et al. Unobtrusive user-authentication on mobile phones using biometric gait recognition. in Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP), 2010 Sixth International Conference on. 2010. IEEE.

[63] Győrbíró, N., Á. Fábián, and G. Hományi, An activity recognition system for mobile phones. Mobile Networks and Applications, 2009. 14(1): p. 82-91.

[64] Gruteser, M. and D. Grunwald. Anonymous usage of location-based services through spatial and temporal cloaking. in Proceedings of the 1st international conference on Mobile systems, applications and services. 2003. ACM.

[65] Hoh, B., et al. Preserving privacy in gps traces via uncertainty-aware path cloaking. in Proceedings of the 14th ACM conference on Computer and communications security. 2007. ACM.

[66] Krumm, J., Inference attacks on location tracks, in Pervasive Computing. 2007, Springer. p. 127-143.

[67] Krumm, J., J. Letchner, and E. Horvitz. Map matching with travel time constraints. in SAE world congress. 2007.

[68] Agapie, E., et al. Seeing Our Signals: Combining location traces and web-based models for personal discovery. in Proceedings of the 9th workshop on Mobile computing systems and applications. 2008. ACM.

[69] Burke, J.A., et al., Participatory sensing. Center for Embedded Network Sensing, 2006.

[70] Bulusu, N., et al. Participatory sensing in commerce: Using mobile camera phones to track market price dispersion. in Proceedings of the International Workshop on Urban, Community, and Social Applications of Networked Sensing Systems (UrbanSense 2008). 2008.

Scalability and Performance of Selected Websites of Universities: An Analytical Study of Punjab (India)

Bhim Sain Singla
Department of Computer Engineering
Punjabi University
Patiala, Punjab, India

Dr. Himanshu Aggarwal
Department of Computer Engineering
Punjabi University
Patiala, Punjab, India

***Abstract*—Today, education has emerged as a major area of commercial activities. The access to various University websites through Internet has opened up new opportunities for the beneficiaries. The creation of these websites fully serves the purpose of educational institutions in advancing and achieving their goals. The varied information made available on these websites and the minimum transaction response time to address the queries of end-users can go a long way to influence their decision in selecting a particular course and an institution. The issue assumes greater significance especially in a developing country like India where a website development and deployment activity is primarily facing the shortage of formalized website design techniques and testing procedures. The performance of most of the University websites is reasonably well, but when accessed by only a few concurrent users. Thus, the aim of this study is to analyze and compare the scalability and performance of selected University websites of Punjab (India) by means of load testing. Simulation of realistic users' behavior is achieved through LoadRunner, a software tool for performance testing. Of all the University websites under study, on the basis of their scalability and performance, it has been found that the websites of Deemed and Central universities are the most and least efficient respectively. The findings of this study can be of great significance for the higher educational institutions in improving the performance quality of their websites resulting into their better ranking and satisfaction of the stakeholders. The paper also outlines the scope for further research in the area under study.**

Keywords*—Hits per second; scalability; performance; throughput; transaction response time; university websites

I. Introduction and Related Work

The advent of World Wide Web has greatly influenced every business in one way or the other. Frankly speaking, it has completely transformed our lives by providing access to a vast knowledge and information on every subject. Millions of Web applications serve billions of Web pages daily through software systems. Web applications have become interactive, dynamic and asynchronous through a number of revolutions [1], [2]. Web applications are an integral part of any website; and subsequently, existing websites have evolved from static information pages to dynamic and service-oriented applications. These are highly used for a broad range of activities on a daily basis in the health care, education, consumer business, banking and manufacturing sectors [3]. Academic institutions make their websites for a wide range of purposes which mainly include distribution of information to the public, delivering online learning facilities to students, promotion of their educational and research programmes and the like. Thus, through this medium, the universities are communicating with and disseminating information to various stakeholders. Students / prospective students, employees / prospective employees, parents, ranking organizations, and the media were identified as the regular users of academic websites [4]. If it is assumed that universities are the brands for education marketing, then websites emerge as a crucial part of this marketing process.

The website of an institution is a gateway to its information and services offered. As such, it should meet all the requirements of its stakeholders. A poor website of an organization can spoil its brand image, loose the potential customers, and weaken its organizational position. Thus, it is of utmost importance to explore those factors which highly influence the users' attitude towards a website which helps the organizations to chalk out a successful e-strategy for the purpose. Many research studies have established the relationship between Web quality factors and user acceptance [5]-[12]. This is due to the reason that Web quality factors can be controlled by an organization; and these can influence users' beliefs and their behavioral intention. A poorly performing website offering a product may fail in its objective. An effective website keeps a balance between end users' expectations and experience with the services being offered. Only those organizations can successfully achieve their goals which are able to lift end users' experience to a level that exceeds expectations. As the end users' expectations always keep on increasing, it is essential for organizations to improve their quality constantly. It raises the question: what should be improved to keep the end users satisfied? Lin [13] emphasized not only on the quality of information, but also the quality of system. System quality is technology-based and enables the users to get faster responses with more convenience and privacy [14]. The time to download a Web page is an important factor for most of the Internet users. The study undertaken by Hoffman and Novak [15] has established a positive correlation between website loading time and user satisfaction. Therefore, fast loading becomes essential for online transactions to be finalized. When loading time is below the expectations of users, they will either prefer to redirect the search engine to another site or give up their search [16]-[19]. Thus, it can be assumed that due to technology advances end users expect sites to be even quicker. The poor performance of a website may downgrade it in search engine results rankings.

Although a good number of websites exist in the academic domain today, yet only a small percentage of these websites satisfy their end users' requirements especially in developing countries like India. It has been observed that during the admission days, many of the University websites are not able to perform well on account of links opening slowly, some of the important links not opening at all, and lost payment transactions. However, such problems can be attributed to the limited use of formalized website design techniques, rapid advancement in Web technologies, limited experience and knowledge of individual designers and developers, short development and evaluation life cycle, lack of formalized website testing procedures, and less resource allocation for website design and development project [20]. Hite and Railsback [21], in their study, confirm that University websites have developed almost as rapidly as corporate websites. But a proper engineering approach for building a web system is not followed; and the engineering process itself is still to be engineered.

The whole scenario gives rise to the need for a thorough analysis of websites both during development and after deployment in order to ensure their conformance to high standards of quality especially in terms of performance. Website testing is considered from two distinct perspectives. The first perspective focuses on identifying the failures in functionality of a website, while the other perspective verifies the conformance of the site behavior with the specified non-functional requirements. The functionality of a website means what a system is supposed to do, while its non-functionality requirements means how a system is ought to be [22], [23]. The non-functional requirements of a system are often described as its "quality attributes". The non-functional requirements that a website is usually required to satisfy, either explicitly or implicitly, are shown in Figure 1.

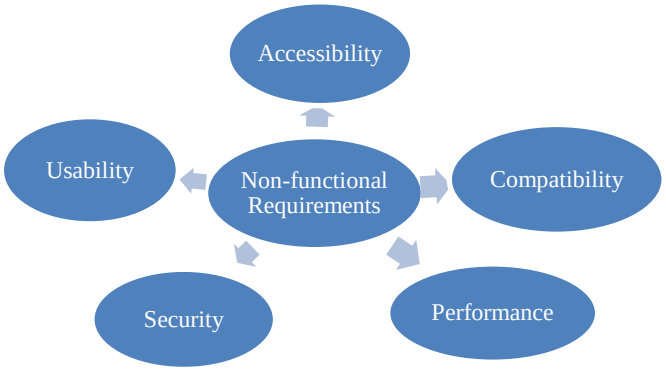


Fig. 1. Non-functional requirements of a website

Most of the methods and approaches followed to test the functional requirements of 'traditional' software can also be used for testing the websites. However, traditional testing theories and methods cannot be used as such to verify each non-functional requirement because of the peculiarities and complexities of websites. The Web application system is typically composed of a database (or the back-end) and Web pages (the front-end) with which users interact over a network through a browser. Garousi et al. [24] considered the website as a distributed system, with a client-server or multi-tier architecture. The other main characteristics are as follows:

- A wide number of users can access it concurrently.
- The use of different hardware, operating systems, Web browsers and network connections generates heterogeneous execution environments.
- A large variety of software components establish the heterogeneous nature of a system which it generally includes. These components can be constructed of different technologies / programming languages, for instance, Hypertext Markup Language (HTML), Cascading Style Sheets (CSS), JavaScript on the client-side and Hypertext Preprocessor (PHP), Ruby, Java on the server-side.
- The dynamic nature of a system makes the software components generate at run time as per user inputs and server status.

The characteristics as mentioned above pose a number of technical and non-technical challenges before the testers to effectively test the sites; and additional efforts are required in web testing [25]. Thus, specific testing activities need to be explored to test the non-functional requirements of a website.

Performance testing is a subset of performance engineering which strives to build performance into the implementation, design and architecture of a system. Liu [26] considers performance testing as a measure to find how fast an application can perform certain tasks, whereas scalability is used to measure performance trends over a period of time with increasing amounts of load. Performance testing consists of simulating multiple virtual users that send requests to the tested server concurrently for evaluating the application performance under a particular load. It is defined as the technical investigation carried out to identify the hurdles in a system, supports a performance tuning effort, determines compliance with performance goals & requirements, and/or collects other performance-related data enabling the stakeholders to make decisions related to the overall quality of the application/system being tested [27]. The performance of any system depends upon many parameters like response time, high throughput from the system, etc. [28]. In 2006, Google revealed that by reducing the size of web page "Google Maps" from 100KB to 80KB, their traffic shot up by 10% in the first week and then 25% in the following three weeks. The results produced by Amazon in 2007 were also the same. It was revealed that for every 100ms increase in load time of Amazon.com their sales decreased by 1% [29]. Thus, the performance of a website needs to be monitored regularly as it is an integral part of a Web design workflow and quality assurance programme.

Although website performance testing is of great significance, yet there has not been any significant study which examined the performance and scalability of University websites of Punjab (India). The higher education sector in the state of Punjab is highly vibrant, fast growing and highly competitive. This can be confirmed from the fact that 10 new universities have started their venture during the last five years.

Therefore, it is highly important for the universities not only to improve their academics and administrative procedures, but also the websites which is a common interface with their end users. Thus, the present study is the outcome of this research gap.

II. Research Objectives

The objectives of the study are as hereunder:

- To evaluate the scalability behavior of University websites by measuring how the average throughput and hits per second will increase with an increase in user load.
- To examine the performance of University websites by measuring average response time and the amount of data processed (throughput) for the same user load.

III. Research Methodology

The study focuses on certain selected websites of universities located in Punjab (India). Presently, there are twenty-five universities in Punjab which include central university (01), deemed universities (2), private universities (13), and state universities (09) (Appendix A). Therefore, stratified random sampling technique was applied to identify the University websites for the purpose of analysis. The university websites selected for this study are listed in Table 1.

TABLE. I. University Websites Considered for the Purpose of Analysis

University status	**University name**	**University web address**
Central Univ.	Central University of Punjab, Bathinda	http://www.cup.ac.in/
Deemed Univ.	Thapar University, Patiala	http://www.thapar.edu/
Private Univ.	Chandigarh University, Mohali	http://www.cuchd.in/
State Government Univ.	Guru Angad Dev Veterinary and Animal Sciences University, Ludhiana	http://www.gadvasu.in/

University websites are analyzed through LoadRunner software, a tool for performance testing. It largely suits this research because the measurement of website performance parameters is, generally, beyond the scope of other techniques such as heuristic evaluation, user evaluation, etc. Also, the evaluations exercised by human users are, usually, based on qualitative criteria which can be prone to error. For performance testing, throughput and response time should be absolutely measured in a concrete and verifiable manner for a set number of users. The framework of evaluation procedure has been shown in Figure 2.

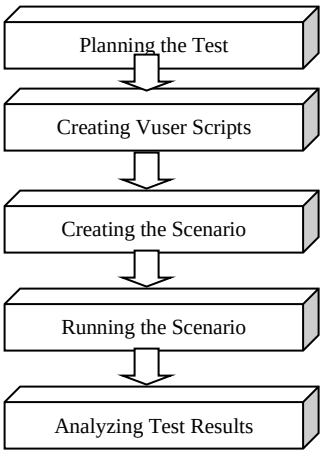


Fig. 2. Sequence of load testing activities performed

Planning the test was the first stage of experimental framework. After selecting the universities, the next step was to identify the web pages or links maximum explored by the end users. For this, a survey was conducted among the students who had recently taken admission in first year of the University programme. Their parents were also included in the survey. An enquiry was made from them to know which links were explored by them the most on the University websites at the time of seeking admission to various courses; and what sort of information they sought or expected under these links. The nine links identified for the purpose were "About Us", "Admissions", "Fee Structure", "Ph.D. / Research & Consultancy Cell", "Training & Placement Cell", "Hostel", "Downloads", "Contact Us" and "Scholarships / Fellowships / Financial aid". On the basis of these inputs, an attempt was made to find which pages/links of the University websites are suitable in the creation of script for performance testing. Uniform Resource Locator (URL) addresses for the identified pages/links of all the universities are listed in Table 2.

Further, scripts were created in Virtual User Generator (VuGen), a component of HP LoadRunner software. HP LoadRunner has 3 components, namely, Virtual User Generator (VuGen) which is used for script creation; Controller is used for executing the performance tests; and Analysis is used for analyzing the performance test results. The protocol used for script creation was Web (HTTP/HTML). The critical flows (Table 2) were first recorded for each university in a separate script file. The scripts were then optimized through parameterization, content checking, transaction naming and custom coding. The user inputs were handled through parameterization. The expected and actual response of the server request was matched through content checking. The transaction was considered passed only when there was a matching between the expected and actual response, otherwise, it was considered failed. Every user request was encapsulated within a transaction name. LoadRunner identified each

transaction through a name and reported its response time. Custom codes were added to determine the size of each visited web page. Scripting being a one-time activity, the same scripts were further reused for various test executions.

TABLE. II. URLS OF IDENTIFIED WEB PAGES

	Central University	Deemed University	Private University	State Government University
About Us	http://www.cup.ac.in/about_CUPB.php	http://www.thapar.edu/index.php/about-us/history	http://www.cuchd.in/about/	http://www.gadvasu.in/about-the-university.asp
Admissions	http://www.cup.ac.in/admission2016.php	http://www.thapar.edu/index.php/admissions/admissions-2016	http://www.cuchd.in/admissions/	http://gadvasu.in/noticedetails.asp?J=0&id=1449
Fee Structure	http://www.cup.ac.in/fee_structure.php	http://www.thapar.edu/images/fees/Fee%20Chart%20for%20ODD%20Semester%202016-17.pdf	http://www.cuchd.in/admissions/program-fee.php	http://www.gadvasu.in/matter.asp?MainCatID=2&SubCatID=70&ChildID=99&Catname=Detail+of+fees+for+UG+%26+PG+Programmes+under+College+of+Veterinary+Science
Ph.D. / Research & Consultancy Cell	http://www.cup.ac.in/research_initiatives_n.php	http://www.thapar.edu/index.php/admissions/phd-programme	http://www.cuchd.in/research/	http://www.gadvasu.in/matter.asp?MainCatID=3&SubCatID=17&Catname=List+of+Ongoing+Research+Projects
Training & Placement Cell	http://www.cup.ac.in/placement_cell.php	http://www.thapar.edu/index.php/students/placements/introduction	http://www.cuchd.in/placements/	http://www.gadvasu.in/PlacementCell.asp?Cat_ID=6
Hostel	http://www.cup.ac.in/campus_life.php	http://www.thapar.edu/index.php/students/hostels	http://www.cuchd.in/student-services/hostel-facility.php	http://www.gadvasu.in/matter.asp?MainCatID=6&SubCatID=35&Catname=Hostels+
Downloads	http://www.cup.ac.in/student_forms.php	http://www.thapar.edu/index.php/forms	http://www.cuchd.in/downloads/	http://www.gadvasu.in/downloads.asp
Contact Us	http://www.cup.ac.in/contact.php	http://www.thapar.edu/index.php/campus/contact-us	http://www.cuchd.in/contact/	http://www.gndu.ac.in/gndu2014/contact.asp
Scholarships / Fellowships / Financial aid	http://www.cup.ac.in/scholarships.php	http://www.thapar.edu/index.php/scholarships	http://www.cuchd.in/scholarship/	http://www.gadvasu.in/matter.asp?MainCatID=2&SubCatID=71&Catname=Awards+and+Medals

Once the script was created for each university, scenarios were created in Controller. A scenario is a file that defines the scripts to execute, the number of virtual users executing those scripts, and the machine that will host (load generators) the virtual users. Virtual users are not real users. Each virtual user works according to its script taken up for execution. For each University, scenario was created for 15, 30 and 45 users load test. Each scenario had a ramp-up phase (starting Vusers) during which the users were added gradually to the application till its predetermined limit was reached. After ramp-up, the script was executed for a fixed time span. During this period, users kept on doing their activities continuously. Later, during

the ramp-down phase, users started leaving the application gradually. The properties set for these actions in a particular scenario are depicted as below in Table 3.

TABLE. III. SCENARIOS SETTINGS

Load Tests (Concurrent User Load)	Action	Properties	Ramp-Up Phase Time	Approx. Ramp-Down Phase Start Time
15 Users	Start Vusers	Start all Vusers: 1 every 00:00:15 (HH:MM:SS)	Initial 3 min 30 sec	After 1 hr 3 min 30 sec
	Duration	Run for 01:00:00 (HH:MM:SS)		
	Stop Vusers	Stop all Vusers: 2 every 00:00:15 (HH:MM:SS)		
30 Users	Start Vusers	Start all Vusers: 1 every 00:00:15 (HH:MM:SS)	Initial 7 min 15 sec	After 1 hr 7 min 15 sec
	Duration	Run for 01:00:00 (HH:MM:SS)		
	Stop Vusers	Stop all Vusers: 2 every 00:00:15 (HH:MM:SS)		
45 Users	Start Vusers	Start all Vusers: 1 every 00:00:15 (HH:MM:SS)	Initial 11 min	After 1 hr 11 min
	Duration	Run for 01:00:00 (HH:MM:SS)		
	Stop Vusers	Stop all Vusers: 2 every 00:00:15 (HH:MM:SS)		

All the scenarios were executed after their creation; and the test results were taken up for analysis under the third component of HP LoadRunner software, namely, Analysis

IV. FINDINGS AND DISCUSSIONS

A. Scalability comparison of selected University websites

Scalability is a performance testing parameter that helps us to identify if an application (website) is capable of scaling with increase in number of concurrent users. Various attributes of an application can be used to gauge its scalability like response time, transactions per second, etc. However, the best possible ways to determine the scalability of any application are throughput and hits per second. Throughput is a performance testing metric which helps us to determine the amount of data (in bytes) an application is able to process. Typically, it is expressed as bytes/sec. Hits per second is a performance testing metric which helps us to determine the number of hits made on the Web server by users during each second of the load test [30]. The scaling of various websites with respect to throughput parameter with varying user loads is expressed through Table 4 and Figures 3 to 6.

TABLE. IV. THROUGHPUT STATISTICS REPRESENTING THE SCALABILITY BEHAVIOR OF UNIVERSITY WEBSITES

University Status (I)	Concurrent User Load (II)	Average Throughput (Bytes per second) (III)	Total (Bytes) (IV)	%age Increase in	
				User Load over previous run (V)	*Average Throughput over previous run* (VI)
Central University	15 Users	292,797.660	1,266,349,881	NA	NA
	30 Users	525,114.716	2,485,893,065	100	79.344
	45 Users	547,154.627	2,720,452,806	50	4.197
Deemed University	15 Users	292,949.293	1,315,635,274	NA	NA
	30 Users	565,405.674	2,663,060,724	100	93.005
	45 Users	839,690.288	4,224,481,841	50	48.511
Private University	15 Users	193,694.812	853,225,645	NA	NA
	30 Users	370,743.814	1,750,281,548	100	91.406
	45 Users	528,576.357	2,721,111,086	50	42.572
State Government University	15 Users	149,146.676	650,279,506	NA	NA
	30 Users	279,409.866	1,343,682,047	100	87.339
	45 Users	400,076.013	2,025,584,854	50	43.186

Table 4, column III determines the average throughput received; while column IV highlights the total data received from the web server during the entire test duration. Average throughput is calculated as total throughput divided by total test duration (in sec). Column V determines by what percentage the user load has increased as compared to previous runs. When user load is raised from 15 to 30, it shows 100% increase; and the increase is 50% when raised from 30 to 45. Column VI displays the percentage increase in average

throughput from the previous run. NA (not applicable) indicates that the current test has been taken as a baseline test.

It can be observed from the throughput graphs given below that for all the universities across various load tests, throughput has increased with the ramping up of users in ramp-up phase. The ramp-up phase period is constituted of initial 3 minutes 30 seconds, 7 minutes 15 seconds, and 11 minutes for the 15, 30 and 45 users' load test respectively. It was noticed during the ramp-down phase that the throughput decreased as the users came out of the application gradually; and finally, reached to zero. The ramp-down phase start time was 1 hour 3 minutes 30 seconds, 1 hour 7 minutes 15 seconds, and 1 hour 11 minutes for the 15, 30 and 45 users' load test respectively.

Figure 3 clearly depicts that with the increase in user load from 15 to 30 **(**100% rise) in the case of Central University of Punjab the throughput also increased by 79.344% (Table 4), though not proportionally. Later on, the user load was further increased from 30 to 45 users, but the throughput managed to increase by 4.197% only (Table 4). For this test, the throughput was neither stable, nor it showed any increase with the increase in user load (during a span of 12 to 35 minutes). Also, the throughput managed to become stable, but it does not scale up proportionally (during a span of 36 to 72 minutes). This indicates that the site under investigation faces the problem of scalability.

The curves shown in Fig. 4 represent the throughput achieved by Thapar University (a deemed University) for all the three tests. It shows that once the users' load stabilized, the throughput also got stabilized with ignorable fluctuations during a span of 15 to 65 minutes. With an increase in user load, the throughput also increased proportionally. With increase in user load from 15 to 30 (i.e., 100% rise) and 30 to 45 (i.e., 50% rise), the throughput increased by 93.005% and 48.511% respectively (Table 4), which evidently indicates that the website is scalable.

Fig. 5 highlights the throughput achieved by Chandigarh University (a private University) for all the three tests. It shows that once the users' load got stabilized, the throughput also got stabilized with least fluctuation for 15 and 30 users test during a span of 15 to 65 minutes which was not observed for 45 users test. In the case of this University, when user load was increased from 15 to 30 (i.e., 100% rise), the throughput also increased proportionately, i.e., 91.406%. However, when the test was conducted for 45 users the throughput was highly unstable; and it showed an increase of only 42.572% (Table 4). It indicates that the site under investigation faces the problem of scalability.

The throughput achieved by Guru Angad Dev Veterinary and Animal Sciences University (a state government University) under all the three tests is highlighted in Figure 6. The results showed that there was 87.339% increase in the throughput after raising the user load from 15 to 30 (Table 4). Thus, the proportional increase in throughput was considerably less than that of user load. Also, when the user load was increased from 30 to 45, the throughput with the increase of 43.186% was stable, but not proportional. Thus, the results indicate the University site under study needs to focus on certain scalability issues.

Hits per second determine the number of hits made on the web server by the users during the load test. Table 5 carries the data showing hits per second for all the University websites under study. The columns III and IV represent the number of hits per second and total hits received respectively by the web server during the entire test duration. It was observed that 100% increase in user load resulted into a maximum increase of 92.693% hits per second in the case of Thapar University (a deemed University), while it was the least, i.e., 79.016% in Central University of Punjab. However, when the user load was increased by 50%, the hits per second showed a maximum increase of 48.893% in the case of Thapar University, while it was as low as 1.367% in Central University of Punjab. Similar results were found after a comparison was made between the data provided in column VI of Tables 4 and 5.

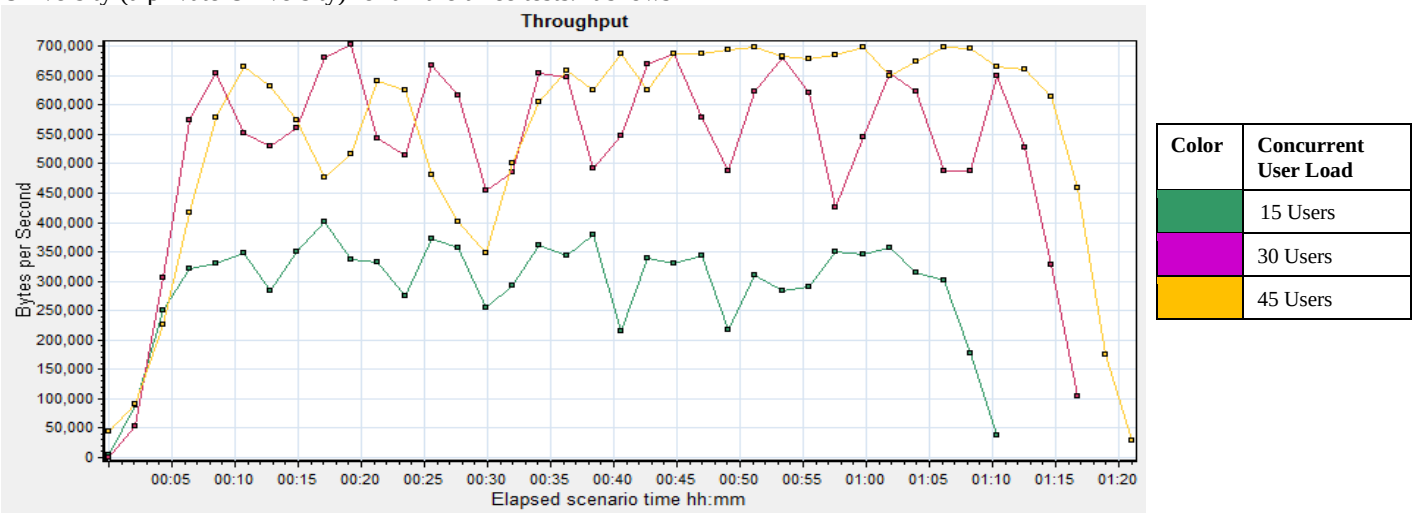
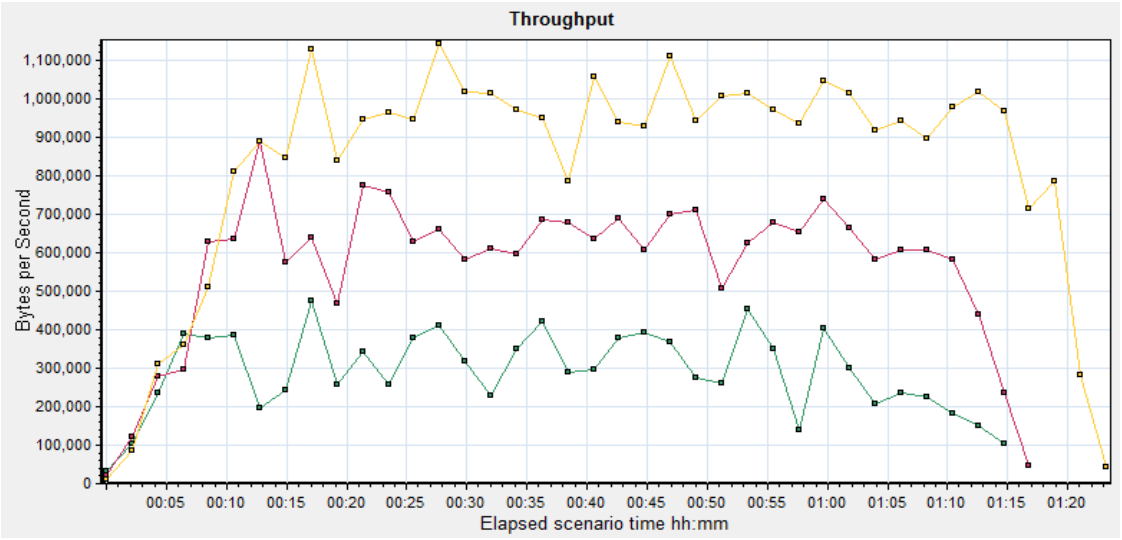
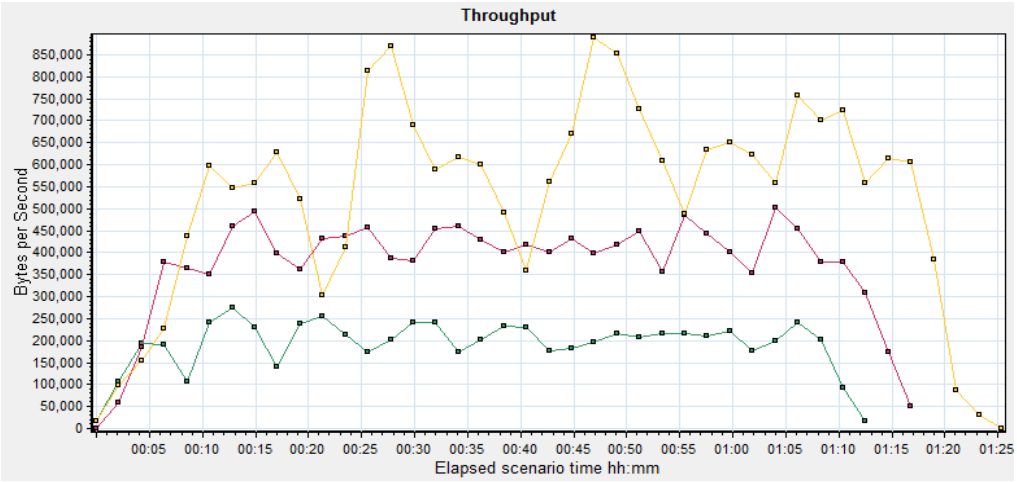


Fig. 3. Throughput behavior of Central University for 15, 30 and 45 users' load



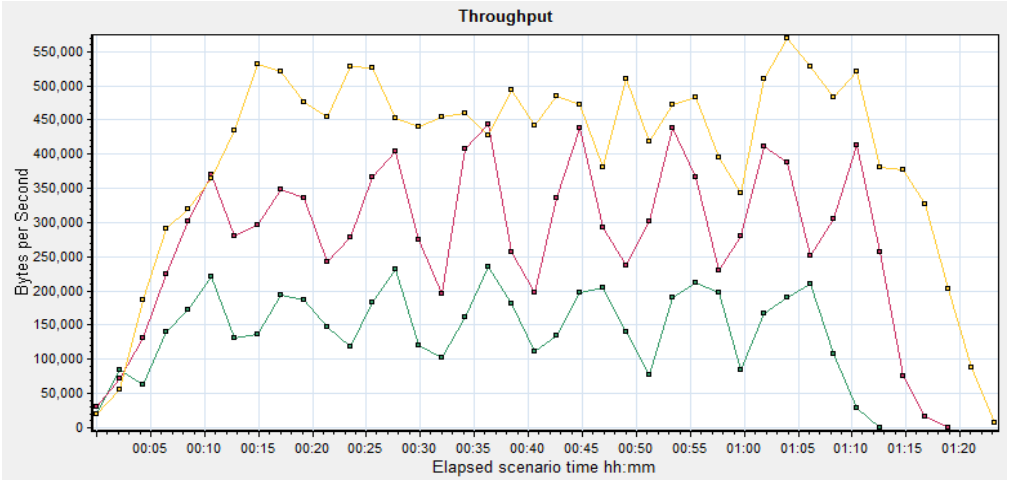
Color	Concurrent User Load
	15 Users
	30 Users
	45 Users

Fig. 4. Throughput behavior of Deemed University for 15, 30 and 45 users' load



Color	Concurrent User Load
	15 Users
	30 Users
	45 Users

Fig. 5. Throughput behavior of Private University for 15, 30 and 45 users' load



Color	Concurrent User Load
	15 Users
	30 Users
	45 Users

Fig. 6. Throughput behavior of State Government University for 15, 30 and 45 users' load

TABLE. V. HITS PER SECOND STATISTICS REPRESENTING THE SCALABILITY BEHAVIOUR OF UNIVERSITY WEBSITES

University Status (I)	Concurrent User Load (II)	Average (Hits per second) (III)	Total (Hits) (IV)	% Increase in	
				User Load over previous run (V)	*Average Hits per second over previous run* (VI)
Central University	15 Users	3.474	15,025	NA	NA
	30 Users	6.219	29,439	100	79.016
	45 Users	6.304	31,343	50	1.367
Deemed University	15 Users	8.622	38,723	NA	NA
	30 Users	16.614	78,251	100	92.693
	45 Users	24.737	124,453	50	48.893
Private University	15 Users	14.233	62,698	NA	NA
	30 Users	27.273	128,758	100	91.618
	45 Users	38.855	200,027	50	42.467
State Government University	15 Users	11.309	49,309	NA	NA
	30 Users	21.234	102,112	100	87.762
	45 Users	30.437	154,102	50	43.341

On the basis of results pertaining to throughput (Table 4) and hits per second (Table 5), it can be said that of all the four universities under study, the websites of Thapar University and Central University of Punjab are the most and least scalable respectively.

B. Performance comparison of selected university websites

The performance of University websites under study has been evaluated on the basis of time taken (response time) and data processed (throughput) for the same user load.

Transaction response time is the basic performance testing metric which helps us to determine the time taken by an application to process the user request. By definition, it is the time duration between users sending the request and receiving its complete response from the server, expressed in seconds. Tables 6 to 9 contain the data showing the transaction response time for all the universities across various user loads. Here, column I demonstrates all the links of the website that a virtual user hits in its iterations while running the script. Column II exhibits the user load for which the test was executed. Columns III to V display the minimum, average and maximum response time taken by that transaction for the respective user load. Column VI determines the number of times an expected response was received for that transaction. Similarly, column VII explains the number of times an expected response failed in its transaction. Most of the transactions failed due to step download timeout error. The error of step download timeout occurs when the user does not receive the whole response in a stipulated time as mentioned in load runner scripts. In this study, tests have a fixed value of 300 seconds. Column VIII demonstrates the page size of the respective link.

The data shown in Table 6 clearly reflects that the web page "Fee Structure" has taken the highest "average transaction response time" for all test runs, i.e., 10.754, 30.856 and 71.281 seconds for 15, 30 and 45 users load test respectively. The page size is of 4.337 MB which makes it the heaviest of this website. In the "Fee structure" page, a PDF file having all the fee details of various courses offered by the university has been downloaded from the server. The web page taking the least "average transaction response time" is not consistent across the different load tests, i.e., 1.423 secs by "Contact Us" page (15 users test), 4.176 secs by "Downloads" page (30 users test) and 10.247 secs by "Training & Placement Cell" page (45 users test). The page having the smallest size of 0.254 MB is "Downloads" for this university. It has been observed that the failure rate for the transactions is as high as 11.7% of the passed transactions for 45 users test run.

A glance at Table 7 provides that the web page "Hostel" has taken the highest "average transaction response time" for all the test runs, i.e., 12.132, 13.609 and 15.856 seconds for 15, 30 and 45 users load test respectively. The page size is of 6.988 MB which makes it the heaviest of this website. Hence, the average response time taken for a transaction is also the highest. However, the web page "Fee Structure" has shown the least "average transaction response time". The average transaction response time recorded after a load test of 15, 30 and 45 users is 1.618, 1.806 and 2.576 seconds respectively. Further, this page is of the smallest size, i.e., 0.326 MB.

As per the results shown in Table 8, in the case of Chandigarh University, the web page "Training & Placement Cell" is the heaviest of this website with a size of 1.241 MB. Thus, the "average transaction response time" is also the highest, i.e., 10.688, 11.075 and 27.731 seconds for all the test runs conducted under a load test of 15, 30 and 45 users respectively. However, the web page "Download" is the smallest with a size of 0.459 MB. Thus, the "average transaction response time" is also the minimum, i.e., 1.991, 2.041 and 7.884 seconds for a similar load test of 15, 30 and 45 users respectively.

TABLE. VI. TRANSACTION RESPONSE TIME STATISTICS OF CENTRAL UNIVERSITY OF PUNJAB FOR ALL LOAD TESTS

Transaction Name (I)	Load Tests (II)	Minimum Transaction Response Time (sec) (III)	Average Transaction Response Time (sec) (IV)	Maximum Transaction Response Time (sec) (V)	No. of Passed Transactions (VI)	No. of Failed Transactions (VII)	Page Size in MB (VIII)
About Us	15 Users	0.893	2.039	11.238	108	1	0.427
	30 Users	0.886	4.932	48.297	226	5	
	45 Users	0.958	13.122	125.254	292	37	
Admissions	15 Users	0.883	1.986	12.410	107	1	0.267
	30 Users	0.903	4.505	36.533	219	7	
	45 Users	0.926	11.211	125.043	269	23	
Fee Structure	15 Users	7.010	10.754	35.375	106	1	4.337
	30 Users	6.931	30.856	122.511	212	7	
	45 Users	7.212	71.281	227.626	244	25	
Ph.D. / Research & Consultancy Cell	15 Users	0.955	1.715	8.646	106	0	0.402
	30 Users	0.973	6.161	122.918	203	9	
	45 Users	1.012	12.316	135.353	215	29	
Training & Placement Cell	15 Users	0.855	1.625	10.340	105	1	0.258
	30 Users	0.886	4.657	122.75	199	4	
	45 Users	0.878	10.247	72.995	199	16	
Hostel	15 Users	6.851	10.124	21.974	104	1	4.233
	30 Users	7.047	29.566	95.418	195	4	
	45 Users	7.003	62.559	232.129	157	42	
Downloads	15 Users	0.868	3.275	122.009	102	2	0.254
	30 Users	0.878	4.176	37.404	190	5	
	45 Users	0.911	10.640	45.244	148	9	
Contact Us	15 Users	0.839	1.423	8.589	100	2	0.255
	30 Users	0.884	4.707	124.029	185	5	
	45 Users	0.926	10.423	69.498	136	12	
Scholarships / Fellowships / Financial aid	15 Users	1.824	7.204	122.849	98	2	1.019
	30 Users	1.830	8.816	48.033	176	9	
	45 Users	1.859	19.439	62.044	120	16	

TABLE. VII. TRANSACTION RESPONSE TIME STATISTICS OF THAPAR UNIVERSITY FOR ALL LOAD TESTS

Transaction Name (I)	Load Tests (II)	Minimum Transaction Response Time (sec) (III)	Average Transaction Response Time (sec) (IV)	Maximum Transaction Response Time (sec) (V)	No. of Passed Transactions (VI)	No. of Failed Transactions (VII)	Page Size in MB (VIII)
About Us	15 Users	2.476	6.292	140.435	108	3	0.570
	30 Users	1.133	4.469	11.817	225	3	
	45 Users	2.586	5.832	17.238	352	8	
Admissions	15 Users	2.833	8.980	213.856	108	0	0.890
	30 Users	2.636	5.442	22.659	222	3	
	45 Users	2.815	7.321	25.129	344	8	
Fee Structure	15 Users	0.597	1.618	4.589	107	1	0.326
	30 Users	0.618	1.806	13.783	220	2	
	45 Users	0.579	2.576	11.894	343	1	
Ph.D. / Research & Consultancy Cell	15 Users	3.210	6.859	128.859	107	0	0.890
	30 Users	2.886	5.479	16.560	220	0	
	45 Users	2.872	6.888	23.299	343	0	
Training & Placement Cell	15 Users	2.856	4.680	23.021	106	1	0.538
	30 Users	2.387	4.470	18.067	215	5	
	45 Users	2.497	5.620	15.592	340	3	
Hostel	15 Users	8.510	12.132	29.697	106	0	6.988
	30 Users	7.214	13.609	31.892	207	8	
	45 Users	7.929	15.856	36.673	338	2	
Downloads	15 Users	2.504	6.281	225.255	106	0	0.530
	30 Users	2.480	5.175	182.622	205	2	
	45 Users	2.347	5.376	18.389	332	6	
Contact Us	15 Users	3.339	10.648	80.455	102	4	0.533
	30 Users	3.077	9.555	22.391	203	2	
	45 Users	3.140	10.723	20.166	329	3	
Scholarships / Fellowships / Financial aid	15 Users	2.798	4.462	13.579	102	0	0.535
	30 Users	2.512	4.650	22.926	200	3	
	45 Users	2.458	5.596	14.367	326	3	

TABLE. VIII. TRANSACTION RESPONSE TIME STATISTICS OF CHANDIGARH UNIVERSITY FOR ALL LOAD TESTS

Transaction Name (I)	Load Tests (II)	Minimum Transaction Response Time (sec) (III)	Average Transaction Response Time (sec) (IV)	Maximum Transaction Response Time (sec) (V)	No. of Passed Transactions (VI)	No. of Failed Transactions (VII)	Page Size in MB (VIII)
About Us	15 Users	1.875	2.319	4.861	112	0	0.508
	30 Users	1.912	2.457	30.732	230	0	
	45 Users	1.979	10.851	75.738	358	0	
Admissions	15 Users	3.026	5.325	30.457	112	0	0.767
	30 Users	3.011	4.913	30.956	230	0	
	45 Users	3.266	15.537	109.318	358	0	
Fee Structure	15 Users	3.348	4.951	31.335	112	0	0.919
	30 Users	3.353	4.669	33.155	230	0	
	45 Users	3.590	16.848	97.816	358	0	
Ph.D. / Research & Consultancy Cell	15 Users	2.234	3.040	30.383	112	0	0.529
	30 Users	2.170	2.890	30.326	230	0	
	45 Users	2.185	10.519	81.621	358	0	
Training & Placement Cell	15 Users	8.871	10.688	37.424	112	0	1.241
	30 Users	8.688	11.075	38.762	230	0	
	45 Users	8.946	27.731	136.325	357	1	
Hostel	15 Users	2.101	2.733	30.293	112	0	0.601
	30 Users	2.078	3.066	31.579	230	0	
	45 Users	2.183	10.231	84.570	357	0	
Downloads	15 Users	1.643	1.991	3.044	112	0	0.459
	30 Users	1.591	2.041	9.726	230	0	
	45 Users	1.687	7.884	66.541	357	0	
Contact Us	15 Users	3.219	4.476	31.049	112	0	0.792
	30 Users	3.202	4.284	34.371	230	0	
	45 Users	3.248	14.321	93.436	357	0	
Scholarships / Fellowships / Financial aid	15 Users	3.510	4.847	30.380	112	0	0.843
	30 Users	3.508	4.452	31.449	230	0	
	45 Users	3.543	15.305	112.491	357	0	

It is clear from Table 9 that in the case of Guru Angad Dev Veterinary and Animal Sciences University, the web page "Fee Structure" is the heaviest of this website with a size of 1.461 MB. Thus, the "average transaction response time" is also the highest for all the test runs, i.e., 6.440, 7.440 and 17.403 seconds for 15, 30 and 45 users load test respectively. Of all the web pages, the best "average transaction response time" for the web page "Training & Placement Cell" is 3.203, 3.716 and 8.039 seconds for 15, 30 and 45 users load test respectively. Further, the web page "Hostel" is of the smallest size, i.e., 0.492 MB.

While analyzing the results of all the University websites under study, it has been found that more the size of web page, higher would be the response time and vice versa. The composition of each web page, i.e., the type of files it integrates has also been analyzed. However, the composition of only those web pages having a size of over one MB in the case of all the University websites is shown in Table 10.

TABLE. IX. TRANSACTION RESPONSE TIME STATISTICS OF GURU ANGAD DEV VETERINARY AND ANIMAL SCIENCES UNIVERSITY FOR ALL LOAD TESTS

Transaction Name (I)	Load Tests (II)	Minimum Transaction Response Time (sec) (III)	Average Transaction Response Time (sec) (IV)	Maximum Transaction Response Time (sec) (V)	No. of Passed Transactions (VI)	No. of Failed Transactions (VII)	Page Size in MB (VIII)
About Us	15 Users	4.327	5.126	40.642	111	0	0.576
	30 Users	4.585	5.256	22.074	232	0	
	45 Users	4.574	11.354	57.102	350	6	
Admissions	15 Users	3.132	3.347	4.554	111	0	0.590
	30 Users	3.389	3.893	8.342	231	1	
	45 Users	3.333	9.959	80.649	350	0	
Fee Structure	15 Users	5.944	6.440	12.886	111	0	1.461
	30 Users	6.308	7.440	21.295	228	3	
	45 Users	6.222	17.403	104.814	343	7	
Ph.D. / Research & Consultancy Cell	15 Users	2.931	3.383	23.521	110	1	0.516
	30 Users	3.228	3.750	8.216	220	8	
	45 Users	3.239	8.403	55.385	338	5	
Training & Placement Cell	15 Users	2.891	3.203	18.317	109	1	0.496
	30 Users	3.189	3.716	5.583	217	3	
	45 Users	3.211	8.039	56.334	333	5	
Hostel	15 Users	2.893	3.208	19.856	107	2	0.492
	30 Users	3.225	3.755	5.203	213	4	
	45 Users	3.187	8.482	96.564	324	9	
Downloads	15 Users	3.168	3.637	20.711	105	2	0.540
	30 Users	3.408	3.942	5.682	210	3	
	45 Users	3.274	8.455	54.565	317	7	
Contact Us	15 Users	3.095	3.477	23.416	105	0	0.506
	30 Users	3.286	3.878	5.249	209	1	
	45 Users	3.285	9.384	62.160	304	13	
Scholarships / Fellowships / Financial aid	15 Users	2.919	3.400	22.784	103	2	0.498
	30 Users	3.254	3.870	20.018	203	6	

TABLE. X. CONTENT BREAKDOWN OF WEB PAGES (IN KB) OF SIZE AT LEAST ONE MB FOR ALL THE UNIVERSITY WEBSITES

University Status	Webpage Name	Images			CSS	Java Script	PHP	PDF	Others*
		JPG	PNG	GIF					
Central University	Fee Structure	131.047	0	10.389	16.552	71.456	26.614	4181.482	3.131
	Hostel	4195.42	0	10.389	16.552	71.456	37.752	0	3.151
	Scholarships / Fellowships / Financial aid	131.047	0	10.389	16.552	71.456	27.742	783.222	3.042
Deemed University	Hostel	247.689	6723.082	0	48.11	115.543	0	0	20.891
Private University	Training & Placement Cell	637.162	267.011	6.03	34.314	310.484	0	0	15.285
State Government University	Fee Structure	1272.369	60.806	8.095	25.462	67.542	0	0	62.147

* Validation scripts, fonts, URLs, directory overheads, etc

Table 10 reveals that University sites under study have used various types of images like PNG, JPEG and GIF. Images constitute 97.420% of page size for the web page "Hostel" of the Deemed University website (having URL: *http://www.thapar.edu/index.php/students/hostels*) which adversely affects the transaction response time. Similarly, in the case of Central University website (having URL: *http://www.cup.ac.in/campus_life.php*), images for the web page "Hostel" constitute 97.026% of the overall page size which lead to increase the transaction response time. Further, the website of Deemed University has mainly used PNG images which are heavier in size as compared to other types of images such as JPG and GIF used by other universities for their websites. However, in the case of Central University website, no PNG image has been used.

Table 11 demonstrates the data explaining average response time with respect to all user load tests undertaken for all the four selected University websites. The average response time has been calculated by taking mean of "average transaction response times". The results, also presented graphically in Figure 7, explain the performance of all the University websites more clearly and effectively. Finally, the overall average response time results calculated for each university across all load tests are presented at the bottom of this table.

A good website is always capable of having a stable response time and maximum throughput under different user load tests. Here, Central University's website has the highest overall average response time of 13.324 seconds across all user load tests (Table 11). As is evident from Fig. 7, there is a great variation in the average response time of selected University websites under different user load tests. This variation is the highest in the case of Central University's website. Thus, it can be said that the website of this University is the least performing among the four University websites. Further, although the overall average response time for the State University website is 6.098 seconds which is less than that of Deemed University (6.755 seconds), yet the throughput is comparatively more than double in the case of Deemed University under all the load tests (see Table 11). Furthermore, Deemed University's website has shown the most stable average response time under different user loads. Thus, it can be said that Deemed University's website is the best performing site of all the four selected University websites.

V. CONCLUSIONS

The main conclusions drawn from this research work are as follows:

- From the scalability point of view, the website of Deemed University under study has been found to be the best as an increase in user load has resulted into an increase in the throughput and hits per second both in tandem and proportionally. Whereas the Central University's website is the least scalable among the sites studied in this research as there has not been a proportional increase in throughput and hits per second against the user load.
- As far as the performance of all the University websites under study is concerned, in terms of overall average response time along with the amount of data processed and stability of average response time relative to varying user load tests, the websites of Deemed and Central universities have been found to be the most and least performing sites respectively.

All the considered Universities have their own underlying hardware, server configurations and technology architecture for their websites. Therefore their performance testing results are bound to be different despite the fact that all the other test parameters such as test users, test scenarios, internet speed etc. are the same for all of these Universities. However each of the universities has a common goal of serving to all the stakeholders' viz. the existing & aspirant students, the faculty members and providing them a good user experience. Hence, this study is an attempt to highlight which university website is designed for providing better performance and scalability

relatively and suggests how other Universities may perform better.

C. Recommendations

The recommendations made on the basis of findings of this study are as follows:

- Larger the size of a web page, higher would be the value of average transaction response time. Thus, the average transaction response time can be improved by way of optimizing the size of web page.
- The size of images to be posted on a website should be reduced to the minimum without data loss and compromising on image' visual quality.
- Web page developers should use file compression utility, e.g., GZIP to minimize the amount of data being downloaded by the end users. It would lead to improve the "average transaction response time".

TABLE. XI. PERFORMANCE STATISTICS OF ALL UNIVERSITY WEBSITES

Load Tests (I)	University Status (II)	Average Response Time (sec) (III)	Total Throughput (Bytes) (IV)
15 Users	Central University	4.460	1,266,349,881
	Deemed University	6.883	1,315,635,274
	Private University	4.485	853,225,645
	State Government University	3.913	650,279,506
30 Users	Central University	10.930	2,485,893,065
	Deemed University	6.072	2,663,060,724
	Private University	4.427	1,750,281,548
	State Government University	4.388	1,343,682,047
45 Users	Central University	24.582	2,720,452,806
	Deemed University	7.309	4,224,481,841
	Private University	14.358	2,721,111,086
	State Government University	9.992	2,025,584,854
Overall average response time of Central University across all the load tests			Average of (4.460, 10.930, 24.582) = 13.324 sec
Overall average response time of Deemed University across all the load tests			Average of (6.883, 6.072, 7.309) = 6.755 sec
Overall average response time of Private University across all the load tests			Average of (4.485, 4.427, 14.358) = 7.757 sec
Overall average response time of State University across all the load tests			Average of (3.913, 4.388, 9.992) = 6.098 sec

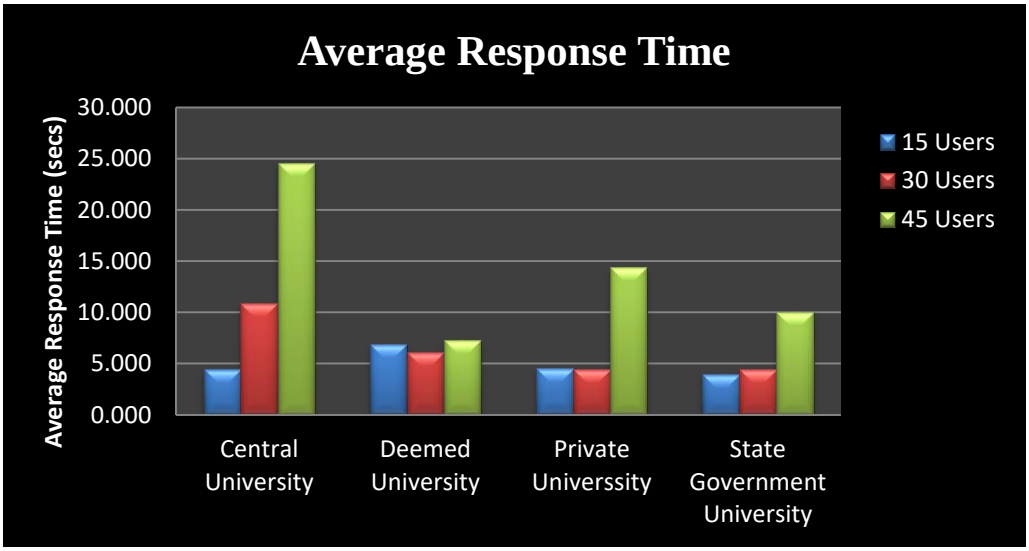


Fig. 7. Average response time of selected University websites under different load tests

D. Scope for further research

- The present study is confined to the websites of selected universities located in the state of Punjab (India) only. However, to corroborate and extend the outcomes of this study, an extensive research is required to be carried out with a larger sample of universities covering diverse regions of the country and world as well.
- The dynamic nature of websites suggests that a longitudinal approach can be followed to examine the changes in performance level of university websites.

ACKNOWLEDGEMENT

The authors acknowledge the Editor-in-Chief Dr. Kohei Arai and editorial committee of the *IJACSA* and in particular to the anonymous reviewers for their valuable comments and suggestions that helped to improve the quality of this manuscript. The authors also feel indebted to Mr. Subhash Rahi of Punjabi University, Patiala for his valuable contribution in improving the manuscript of this research paper.

REFERENCES

[1] Y.-F. Li, P. K. Das and D. L. Dowe, "Two decades of web application testing – A survey of recent advances," Information Systems, vol. 43, pp. 20-54, July 2014.

[2] M. S. Chaves, C. C. S. D. Araújo, L. R. Teixeira, D. V. Rosa, I. G. Júnior and C. D. Nogueira, "A new approach to managing Lessons Learned in PMBoK process groups: the Ballistic 2.0 Model," International Journal of Information Systems and Project Management, vol. 4, no. 1, pp. 27-45, January 2016.

[3] M. Haydar, A. Petrenko, S. Boroday and H. Sahraoui, "A formal approach for run-time verification of web applications using Scope-Extended LTL," Information and Software Technology, vol. 55, no. 12, pp. 2191-2208, December 2013.

[4] S. M. McAllister and M. Taylor, "Community college web sites as tools for fostering dialogue," Public Relations Review, vol. 33, no. 2, pp. 230-232, June 2007.

[5] L.-A. Ho, T.-H. Kuo and B. Lin, "The mediating effect of website quality on Internet searching behavior," Computers in Human Behavior, vol. 28, no. 3, pp. 840-848, May 2012.

[6] M. M. Al-Debei, D. Jalal and E. Al-Lozi, "Measuring web portals success: A respecification and validation of the DeLone and McLean information systems success model," International Journal of Business Information Systems, vol. 14, no. 1, pp. 96-133, January 2013.

[7] M. M. Al-Debei, "The Quality and Acceptance of Websites: An empirical investigation in the context of higher education," International Journal of Business Information Systems, vol. 15, no. 2, pp. 170-188, February 2014.

[8] D. J. Kanagaraj and J. C. Sudhahar, "Website quality: Imperatives for effective industrial marketing through websites' usage intensity augmentation," International Journal of Electronic Customer Relationship Management, vol. 9, no. 4, pp. 203-219, January 2015.

[9] G. Sahi and S. Madan, "STEP model for linking website Usability dimensions and website success measures in B2C e-Commerce setting," International Journal of Business Information Systems, vol. 20, no. 2, pp. 219-237, August 2015.

[10] M. Butkiewicz, H. V. Madhyastha and V. Sekar, "Understanding website complexity: Measurements, metrics, and implications," in *Proceedings of the 2011 ACM SIGCOMM Conference on Internet measurement*, Berlin, Germany, 2-4 November 2011, pp. 313-328.

[11] L. Mich, "Evaluating website quality by addressing quality gaps: A modular process," in *Proceedings of the IEEE International Conference on Software Science, Technology and Engineering (SWSTE)*, Ramat Gan, Israel, 11-12 June 2014, pp. 42-49.

[12] K.-C. Chang, M.-C. Chen, C.-L. Hsu and N.-T. Kuo, "Integrating loss aversion into a technology acceptance model to access the relationship between website quality and website user's behavioural intentions," Total Quality Management & Business Excellence, vol. 23, no. 7-8, pp. 913-930, January 2012.

[13] H.-F. Lin, "An investigation into the effects of IS quality and top management support on ERP system usage," Total Quality Management & Business Excellence, vol. 21, no. 3, pp. 335-349, April 2010.

[14] T. Ahn, S. Ryu and I. Han, "The impact of web quality and playfulness on user acceptance of online retailing," Information & Management, vol. 44, no. 3, pp. 263-275, April 2007.

[15] D. L. Hoffman and T. P. Novak, "Marketing in hypermedia computer-mediated environments: Conceptual foundations," Journal of Marketing, vol. 60, no. 3, pp. 50-68, July 1996.

[16] B. D. Weinberg, "Don't keep your Internet customers waiting too long at the (Virtual) front door," Journal of Interactive Marketing, vol. 14, no. 1, pp. 30-39, December 2000.

[17] S. Y. Kim and Y. J. Lim, "Consumers' perceived importance of and satisfaction with Internet shopping," Electronic Markets, vol. 11, no. 3, pp. 148-154, July 2001.

[18] M. Cao, Q. Zhang and J. Seydel, "B2C e-Commerce website quality: An empirical examination," Industrial Management & Data Systems, vol. 105, no. 5, pp. 645-661, June 2005.

[19] B. Yen, P. J.-H. Hu and M. Wang, "Toward an analytical approach for effective web site design: A framework for modeling, evaluation and enhancement," Electronic Commerce Research and Applications, vol. 6, no. 2, pp. 159-170, June 2007.

[20] M. J. Taylor, J. McWilliam, H. Forsyth and S. Wade, "Methodologies and website development: A survey of practice," Information and Software Technology, vol. 44, no. 6, pp. 381-391, April 2002.

[21] N. G. Hite and B. Railsback, "Analysis of the content and characteristics of University websites with implications for web designers and educators," Journal of Computer Information Systems, vol. 51, no. 1, pp. 107-113, September 2010.

[22] B. Johansson and M. Lahtinen, "Getting the balance right between functional and non-functional requirements: the case of requirement specification in IT procurement," International Journal of Information Systems and Project Management, vol. 1, no. 1, pp. 5-16, January 2013.

[23] A. Stellman and J. Greene, Applied Software Project Management, 1st ed., O'Reilly Media: Sebastopol, USA, 2006.

[24] V. Garousi, A. Mesbah, A. Betin-Can and S. Mirshokraie, "A systematic mapping study of web application testing," Information and Software Technology, vol. 55, no. 8, pp. 1374-1396, August 2013.

[25] F. A. Torkey, A. Keshk, T. Hamza and A. Ibrahim, "A new methodology for web testing," in *Proceedings of the 5th International Conference on Information and Communications Technology*, Cairo, Egypt, 16-18 December 2007, pp. 77-83.

[26] H. H. Liu, Software Performance and Scalability: A Quantitative Approach, 1st ed., John Wiley & Sons: Hoboken, Hudson, 2009.

[27] J. D. Meier, C. Farre, P. Bansode, S. Barber and D. Rea, Performance Testing Guidance for Web Applications: Patterns & Practices, 1st ed., Microsoft Press: Washington, USA, 2007.

[28] S. Garg, K. Modi and S. Chaudhary, "A QoS-aware approach for runtime discovery, selection and composition of semantic web services," International Journal of Web Information Systems, vol. 12, no. 2, pp. 177-200, June 2016.

[29] A. Lumsden, "*Best practices for increasing website performance",* https://webdesign.tutsplus.com/articles/best-practices-for-increasing-website-performance--webdesign-9109

[30] R. Khan and M. Amjad, "Performance testing (load) of web applications based on test case management," Perspectives in Science, vol. 8, pp. 355-357, September 2016.

APPENDIX A. LIST OF UNIVERSITIES IN PUNJAB (INDIA)

TABLE. XII. CATEGORIZATION OF UNIVERSITIES ACCORDING TO THEIR STATUS

University status	University name	University web address	Total number of universities
Central	Central University of Punjab	http://www.cup.ac.in/	01
Deemed	Sant Longowal Institute of Engineering and Technology	http://sliet.ac.in/	02
	Thapar University	http://www.thapar.edu/	
Private	Adesh University	http://adeshuniversity.ac.in/	13
	Akal University	http://auts.ac.in/	
	Chandigarh University	http://www.cuchd.in/	
	Chitkara University	http://www.chitkara.edu.in/	
	DAV University	http://www.davuniversity.org/	
	Desh Bhagat University	http://www.deshbhagatuniversity.in/	
	GNA University	http://gnauniversity.edu.in/	
	Guru Kashi University	http://www.gurukashiuniversity.in/#/	
	Lovely Professional University	http://www.lpu.in/	
	Rayat-Bahra University	http://www.rayatbahrauniversity.edu.in/	
	RIMT University	http://www.rimt.ac.in/	
	Sant Baba Bhag Singh University	http://www.sbbsuniversity.in/	
	Sri Guru Granth Sahib World University	http://sggswu.edu.in/	
State Government	Baba Farid University of Health Sciences	http://bfuhs.ac.in/	09
	Guru Angad Dev Veterinary and Animal Sciences University	http://www.gadvasu.in/	
	Guru Nanak Dev University	http://www.gndu.ac.in/	
	Guru Ravidas Ayurved University	http://www.graupunjab.org/	
	I. K. Gujral Punjab Technical University	https://www.ptu.ac.in/	
	Maharaja Ranjit Singh State Technical University	http://www.mrsstu.ac.in/	
	Punjab Agricultural University	http://web.pau.edu/	
	Punjabi University	http://www.punjabiuniversity.ac.in/	
	Rajiv Gandhi National University of Law	https://www.rgnul.ac.in/	

Source: University Grants Commission's website (http://www.ugc.ac.in/, 12/05/2016)

A Proposed Framework to Investigate the User Acceptance of Personal Health Records in Malaysia using UTAUT2 and PMT

[1]Ali Mamra
Faculty of Information and Communication Technology
Universiti Teknikal Malaysia Melaka
Melaka, Malaysia

[3]Gede Pramudya Ananta
Faculty of Information and Communication Technology
Universiti Teknikal Malaysia Melaka
Melaka, Malaysia

[5]Yasir Hamad Ahmed
Faculty of Information and Communication Technology
Universiti Teknikal Malaysia Melaka
Melaka, Malaysia

[2]Abdul Samad Sibghatullah
Faculty of Information and Communication Technology
Universiti Teknikal Malaysia Melaka
Melaka, Malaysia

[4]Malik Bader Alazzam
Faculty of Information and Communication Technology
Universiti Teknikal Malaysia Melaka
Melaka, Malaysia

[6]Mohamed Doheir
Faculty of Information and Communication Technology
Universiti Teknikal Malaysia Melaka
Melaka, Malaysia

***Abstract*—Personal Health Records (PHRs) can be considered as one of the most important health technologies. PHRs enroll the patients directly to their health decision making through giving them the authority to control and share their health information. Testing the user acceptance toward new technology is a vital process. Over the previous decades many models have been used and the latest one was UTAUT2. UTAUT2 have been widely used in e-business and gaming user acceptance researches, whereas, it has been rarely used in health field. This study proposes a combination of UTAUT2 and PMT in order to investigate the user acceptance of PHRs. Relevant factors may be added at the literature review stage. The final model will be used as a framework to investigate the user acceptance of PHRs in Malaysia.**

***Keywords*—PHRs; User Acceptance; UTAUT2; PMT; Malaysia**

I. INTRODUCTION

Health sector is the one of the biggest sectors in many nation[1]. Governments everywhere focus on this sector as the most necessary sector for their citizens. This concern has resulted in huge investments on health infrastructures, high monitoring on health services, and rapid implementation on technologies related to this sector [2][1]. Information technology on the other hand has been rapidly growing over the past years. E-Health as a shared part between information technology and health sector has its part of this evolution. E-Health is a global need in nowadays communities which has led many governments to announce their vision of the implementation of e-Health at each country [2][3].

II. HIT AND MALAYSIAN VISION

Healthcare Information Technology (HIT) can be considered as one of the major components of the modern e-health structure. E-health involves various HIS that deals with the health information of patients such as: Electronic Health Record (also known as EHR), Medical Health Record (Also known as EMR), and Personal Health Record (also known as PHR)[4]. Electronic Health Record is an e-health system used by any health service provider to store and provide detailed health information for a certain patient [5][6]. Electronic Medical Record refers to the records which can be accessed from various sources of information; usually involve the patient's lab tests, allergies, and diagnoses [7]. Personal Health Record is a record which can be accessed from the patient himself in order to manage his health record, save new information, share information among more than one health provider, and emergency cases such as traveling or disasters[8]. The importance of the HIT motivates many governments to announce their vision related to the adoption of HIT which can be considered as the backbone of the health informatics. It also reflects the striving toward establishing new departments and issuing new rules and acts that could help in controlling the health information obtained from patients[9].

According to Malaysia's Health report [10], The health vision of the Malaysian Ministry of Health says that:

"Malaysia is to be a nation of healthy individuals, families and communities, through a health system that is equitable, affordable, efficient, technologically appropriate, environmentally-adaptable and consumer-friendly, with emphasis on quality, innovation, health promotion and respect of human dignity and which promotes individual responsibility and community participation towards an enhanced quality of life".

Malaysia has its own vision to adopt the PHRs in the near future; according to the [11] Malaysian Healthcare is going to adopt a USB-based PHR within the next five years, which reflects the intention to adopt electronic PHRs in the future. Regarding to the use of the PHRs in Malaysia in the

future,[11] had a vision to the PHRs to be used by a wide segment of users regardless to their being sick or healthy. Moreover, in June 2010 The Personal Data Protection Act (PDPA) has been issued by the Malaysia government as a step in the way to adopt the HIS in Malaysia; the issued act determine a group of responsibilities of the people who works at the health care sector in dealing with the medical information of the patients. It has been published for the first time in the Federal Gazette, and the act has been effective since November 2013 with new added obligations [12]. The new Act involves various obligations related to variant HISs; many issues has been listed such as privacy, security, accessibility, and commercial issues[13]. Certain actions can be taken by any patient related to their personal data protection against their healthcare providers; the patients have the right to access, update, modify, and share their health information, whereas, healthcare provider can only use these information for the therapeutic purposes[12]. These rules, act, and vision reflect the intention to adopt the HIT by the Malaysian government.

PHRs are one of many other technologies that can be listed as a consumer health informatics technology. Consumer health informatics technologies have been adopted in order to be used by the patient at home in order to control his health information using online applications such as web-based applications. These significant capabilities enrolled the patients more into making decisions related to their health status. The variant platforms used in this technology relay on the good communications between the patients and the health care provider[14]. Therefore, the internet penetration in any nation can be considered as a significant and vital aspect in describing the future of patient to physician communications. The internet penetration in Malaysia is above average in the region, almost 70 % of Malaysians are able to access the Internet using various technologies such as homeliness, broadband, and mobile data. Compared to other neighbouring countries such as: Thailand, Philippine, Indonesia, Viet Nam, and except Singapore, Malaysia has the highest Internet penetration in the region. The usage of the Internet by gender in 2014 was almost equal; 51.4 per cent were males, and 48.6 per cent were females, which reflects the wide spread of the technology among citizens from both genders. Laptops were the most preferred way to access the Internet by a computer in 2014; 21.8 % used personal computer to access the Internet, 34.1 % used tablets, whereas 52.8 used laptops Internet usage by age vary from year to another, young people has the heist percentage, whereas, elderly people penetration has increased from 5.3 % in 2013 to 7.3 % in 2014 [15]. This is another significant aspect that should help to facilitate the adoption of PHRs in Malaysia.

III. PERSONAL HEALTH RECORDS (PHRS)

PHRs are electronic medical records that involve information about the patient such as personal information, lab test, diet information, medical history, and other medical information which can be accessed and maintained by the patients themselves with the ability to share this information with their health care providers in a private, convenient, and secure manner[16][17], [18], [19], and [9]. A common well-known and unified definition of PHRs in not yet decided; many different organizations gave various definitions of PHR depending on the functions and the role of the PHR decided by these organizations[20] . Over the previous years, many institutions such as the Markle Foundation, the American Health Information Management Association (AHIMA), and the National Alliance for Health Information Technology (NAHIT) have addressed the importance of the PHRs and the wide adoption of the technology in many nations[21].

According to Markle Foundation PHRs defined as:

"The Personal Health Record (PHR) is an Internet-based set of tools that allows people to access and coordinate their lifelong health information and make appropriate parts of it available to those who need it"[22].

According to the American Health Information Management Association (AHIMA), PHRs can be defined as:

"An electronic, universally available, lifelong resource of health information needed by individuals to make health decisions, individuals own and manage the information in the PHR, which comes from health care providers and the individual. The PHR is maintained in a secure and private environment, with the individual determining rights of access. The PHR is separate from and does not replace the legal record of any provider" [23].

According to the National Alliance for Health Information Technology (NAHIT), PHRs can be defined as:

"An electronic record of health-related information on an individual that conforms to nationally recognized interoperability standards and that can be drawn from multiple sources while being managed, shared, and controlled by the individual" [24].

PHR is one of the newest services in the e-health field which enrol the patient directly to make decision related to his health information and status. One of the most important advantages of the PHR is that; patients are able to access, update, and share their medical information such as blood pressure, diabetes, spirometry, or any other info with their health care providers or others [4]. These features save a lot of patients' time and efforts since they do not have to visit their health service provider every time in order to modify their health record. Therefore, PHR can be considered as one of the most important health services which have not been applied in many countries yet [25]. On the other hand, the normal meetings between patients and professionals may not involve enough information about the patient; patient may forget or hide some information for many reasons, some patient may feel shy from reviling some information. In addition to this, the limited time of the meeting may not be enough for the information exchange between the professional and his patient. Therefore, it is important for both the patient and the professional to start the information exchange among them using specific tools[26]. Thus, PHR is needed to be adopted in any nation since it has the ability to enhance the information exchange between the patient and the healthcare provider, which will reflect positively on the patient to physician relationship.

The important characteristics of PHR; the ability to access medical information, the ability to modify medical information, and the information sharing have been reviling many concerns regarding to the ease of use, privacy, and user attitude toward this technology. The ease of use is a significant issue especially to those elderly people who have not good experience with technology. Privacy on the other hand has a significant influence on patient decision since they might not have a clear vision about the level of security applied in such technologies, and the laws that guarantee their rights to be secured and to their information to be private [8]. The user acceptance of the technology is a key to success; investigating the user acceptance and behavioural intention toward new technology can be considered as a critical issue[27]. Testing the user acceptance of PHR in certain society may revile the future of the interaction with this technology and prevent certain problems that the technology may face in the future.

IV. User Acceptance Theory

According to [28]; investigating the user acceptance toward a new technology, goods, or services plays a significant role in deciding the future of the interaction between the user and the technology, goods, or services in the future. Along the previous decades, many theories and models have been introduced in order to describe and investigate the factors that may affect the user acceptance toward new technology. Technology Acceptance Model (TAM) has been introduced by [29], Theory of Reasoned Action (TRA) (Fishbein et al. 1975), Theory of Planned Behaviour (TPB) which can be considered as one of the first theories that describe the behavioural intention toward technology [30], Innovation Diffusion Theory (IDT) (Rogers 1995), and many other models.

A. Unified theory of acceptance and use of technology (UTAUT)

UTAUT has been introduced by [31] as a unification of eight models have been widely used in previous studies. The UTAUT involves a detailed analysis that describes the factors affecting the user acceptance of a certain item. UTAUT explained about 70 percent of the variance in behavioural intention to use a technology and about 50 percent of the variance in technology use [27]. Since 2003, UTAUT alone, a combination of UTAUT and other models, or an extended version of UTAUT has been widely applied in many fields and various studies that aim to investigate the user acceptance toward new technology[32]. UTAUT is generally consisting of four major parts; Performance Expectancy (PE) which describes the degree of benefit gained by an individual using certain technology, Effort Expectancy (EE) describes the ease of use of a certain technology by an individual, Social Influence (SI) which describes the effective motivation by others (family, friends, or colleagues) to an individual intended to use certain technology, and Facilitating Conditions (FC) which describes the facilities available which may support the use of the new technology by an individual (Venkatesh et al. 2003).

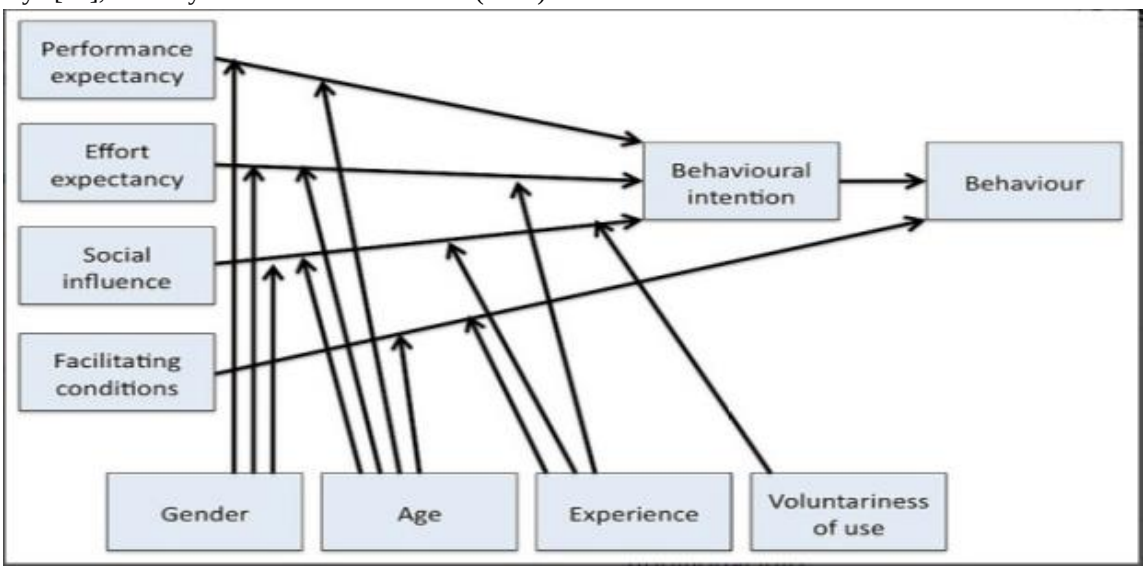


Fig. 1. UTAUT

As shown in (Figure 1), UTAUT relationships represented by variant individual variables such as: gender, age, experience, and voluntariness of use.

B. UTAUT2

Despite the huge success of the UTAUT, [27] have introduces UTAUT2 as an enhanced version of UTAUT that involves new three elements which are: Price Value (PV), Hedonic Motivation (HM), and Habit (HT), each new element has a significant influence on the behavioural intention of users. Hedonic Motivation (HM) describes the level of enjoyment of the technology used by an individual; this factor has been used in previous models such as TAM as intrinsic motivation. Price Value (PV) describes the price influence on choosing a new technology; people generally looking for the price to performance aspect in order to make a decision on using a new technology. Habit (HT) describes the habitual behavior of the user of the new technology; being used to use similar technologies will motivate an individual to use the new technology. HT can be considered as one of the most

significant factor in predicting the use of the technology in the near futureSince its introduction in 2012 UTAUT2 has been widely applied in many fields except a noticeable lack in the health sector[5]. However, UTAUT2 has a higher accuracy that may enhance its ability to investigate the user acceptance and the behavioural intention toward new technology, and it has been recommended by its author to be applied in different fields, studies, and countries in order to come out with the highest accurate results[27].

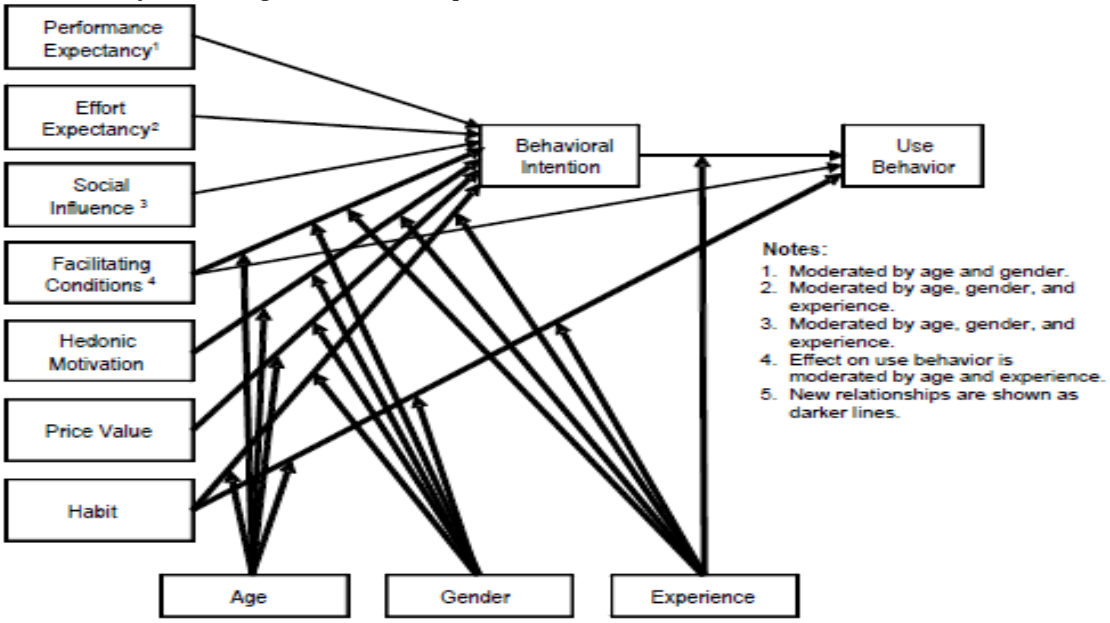


Fig. 2. UTAUT2

C. Proposed model

UTAUT and UTAUT 2 have been widely applied in many fields such as e-business, e-learning, gaming, and mobile technologies[33], and [32]. Few studies discussing the user acceptance of health care services and HIT have been done using UTAUT2[33]. UTAUT2 has higher accuracy in investigating the user acceptance and behavioural intention of a user toward new technology[27]. On the other hand, investigating the behavioural intention toward healthcare technologies has been done over the past years using various models. The most widely applied model in this field is the Protection Motivation Theory (PMT). PMT discusses the individual's vision regarding to the vulnerability and severity of certain threat, and what is the capability of these users to deal with such threat which will reflect the behavioural intention of these users toward the new technology[34]. According to [35], adding threat appraisal endogenous variable from PMT on his unified model have found to be significantly affecting the user acceptance and the behavioural intention toward new health technology. Thus, and based on the results appeared on previous studies, this research proposes an enhancement of UTAUT2 by adding the Threat Appraisal endogenous variable from PMT which may have a significant influence on the user acceptance and behavioural intention toward e-health technologies (See Table 1).

TABLE. I. (UTAUT2 AND PMT COMPARISON)

Endogenous Variables	***UTAUT 2***	***PMT***
Performance Expectancy	Available	Available
Effort Expectancy	Available	Not-Available
Social Influence	Available	Not-Available
Facilitate Conditions	Available	Available
Price Value	Available	Not-Available
Hedonic Motivation	Available	Not-Available
Habit	Available	Not-Available
Threat Appraisal	Not-Available	Available

Other factors such as security, privacy, and trust might be added after reviewing related studies.

V. PROBLEM STATEMENT

The rapid development in ICT and HIT has reviled new needs such as the use of Personal Health Records as a part of the e-health. The authority of accessing and modifying the health records by patients has great advantages in term

of saving time and efforts. On the other hand, accessing these services by patients may generate doubts about the performance, use, and the privacy issues.

Testing the user acceptance of new technology in certain society may revile the future of the interaction between the user and the new technology. UTAUT is one of the most desirable models in order to be used to test the user acceptance of a certain technology. UTAUT has been used and applied in few studies related to PHRs. On the other hand, UTAUT2 have not been yet applied in the PHRs research area despite its enhanced structure. Therefore, this research proposes UTAUT2 to be used as the main model to investigate the user acceptance of PHRs in Malaysia. UTAUT and UTAUT2 have been introduced as a business model; therefore, some features must be added in order to fit the healthcare field. PMT on the other hand has been widely used in investigating the behavioural intention to use HIT. The most significant component related to the HIT in this model is the "Threat Appraisal" variable.

This research will focus on those features that affect the user acceptance of Personal Health Record in Malaysia by enhancing the UTAUT2 in order to be suitable for healthcare sector. A combination of UTAUT2 and PMT will be proposed as a result of this research.

VI. Research Objectives

The main objectives of this research are:-

1) Deciding the factors affecting the user acceptance of PHR in certain society.
2) Analyzing the impact of these factors.
3) Developing an enhanced model in order to analyze these factors.

VII. Expected Outcomes

The expected outcomes of this research are:-

1) A list of factors that affect the PHR acceptance in Malaysia.
2) The impact of these factors on the user acceptance of PHR in Malaysia.
3) An enhanced user acceptance testing model.

VIII. Scope of Research

The scope of this research revolves around the factors affecting technology acceptance of Personal Health Record. The factors involved in this research are the technology acceptance factors adopted from the UTAUT2 and the factors added by this research and how they affect the user acceptance and the behavioural intention toward the Personal Health Record.

IX. Limitations

The main limitations will be the lack of resources related to the PHRs acceptance research field which has been found to be few in comparison with other fields such as EHR, and EMR. On the other hand, the resources describing the history of e-health in Malaysia are insufficient.

X. Contribution to the Knowledge

This research contribution will be:-

1) Illustrating the Obstacles facing the adoption of PHRs in Malaysia.
2) The concerns related to the use of PHRs in Malaysia.
3) The current capabilities of adopting the PHRs in Malaysia.
4) The cultural differences that may affect the use of PHRs in Malaysia.
5) The desired components of UTAUT2 in healthcare field.
6) A unique combination represented in an enhanced model of user acceptance toward new technology to be used in healthcare field.

XI. Research Methodology

A. Introduction

Over the previous decades, researchers from all over the world have tried to define the users' needs which reflect their acceptance of any product. Several theories and models have been introduced in order to investigate the user acceptance of a specific product. The results of these researches were group of a different models reflects the vision of their founders. Many models and theories such as:-

1) Theory of reason Action (TRA)
2) Technology acceptance model (TAM)
3) Motivational model (MM)
4) Theory of planned behaviour (TPB)
5) Combined TAM and TPB (C-TAM-TPB)
6) Model of PC Utilization MPCU
7) Innovation Diffusion Theory (IDT)
8) Social Cognitive Theory (SCT)

All the above models and theories have been combined and unified in order to have a unified model with effective tools that can describe the user acceptance in the optimum way. The unification of these models and theories has led to the introduction of UTAUT "Unified Theory of acceptance and Use of Technology"[31].

Figure 1 and 2 illustrate the relationships that exist in the UTAUT and UTAUT2 model. The UTAUT model has four components, PE which refers to Performance expectancy, EE, which refers to effort expectancy, SI which refers to social influence, and FC which refers to facilitating conditions. Whereas, UTAUT2 has three extra important components which are: Hedonic Motivation, Price Value, and Habit beside the above mentioned components in UTAUT. The components of these models (also known as endogenous variables) are the technology intention to use and behaviour. There are other four moderators namely age, experience, gender and voluntariness. [31] and [27].

UTAUT and UTAUT2 have been introduced as an e-business user acceptance model; therefore, adding some features to this model will enhance it in order to be used in PHRs research.

B. Methodology

Methodology is the "a way of thinking and studying social reality" as defined by [36], they also suggested that quantitative methods might be used appropriately in any research paradigm, which is defined as the basic belief system or worldview that guides the investigator. The issue of appropriate method is only secondary to the choice of the appropriate paradigm [36]. In this research (See Figure 3) the information gathered from the literature review will guide us to find out the issues that affects the user acceptance of Personal Health Records in Malaysia. These new factors or issues will be added to the desired model in order to fit the e-health field.

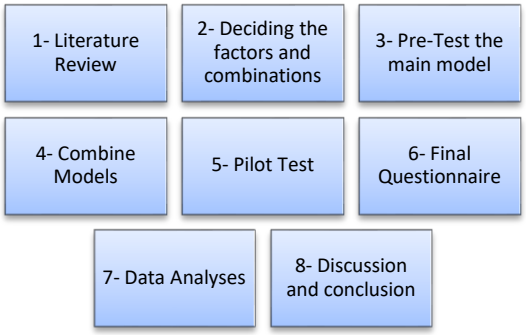


Fig. 3. Research Methodology

Choosing between qualitative or quantitative study still a concern, researchers from both parties have their own reasons to prefer any. This study focuses on examining the user acceptance of PHRs, and since it is targeting a wide segment of users, the quantitative study is preferred.

The obtained data from the literature review regarding to the factors affecting the acceptance of PHRs in related works will be analysed in order to decide the primitive questionnaire. After finishing the review analyses, the new model creating will start by three major phases which are: Pre-Test, Pilot-Test, and the Final Questionnaire.

First of all, a pre-test process will be done to examine the undesired components in UTAUT2 in order to be ignored at the pilot test stage; pre-test will be held in order to find out the undesired components which do not have any influence on the user acceptance toward new technology in PHRs research field.

The second stage will involve a pilot test of the combination in order to find out the suitability of the questions that will be used in the final stage which is the final questionnaire; pilot test process is a questionnaire done using small sample size in order to find out any problems or missing issues in the final questionnaire.

Finally, a quantitative study represented in a modified questionnaire will be held as a result of the pilot test process. The modified questionnaire will be distributed to a sample of expected PHR users in order to be analysed using specific tool (e.g. SPSS) and provide the user acceptance details of PHR. The findings of the final questionnaire results analysis will reveal the user acceptance of PHR in Malaysia, the capabilities to adopt this technology, and the obstacles may face the adoption of PHRs in Malaysia.

XII. Conclusion

This study aims to investigate the user acceptance toward PHRs in Malaysia. Using suitable testing model may increase the accuracy of the obtained results. UTAUT2 is one of the most accurate user acceptance testing models. Adding some features to increase the suitability of this model to the health field may result in higher accurate results. As a beginning, threat appraisal will be added to the UTAUT2, some other factors such as privacy and trust may be added after reviewing a number of related studies.

Acknowledgment

This work was supported by Centre of Research and Innovation Management (CRIM) and Faculty of ICT, Universiti Teknikal Malaysia Melaka, Hang Tuah Jaya, 76100 Durian Tunggal, Melaka, Malaysia.

References

[1] E. Liikanen, "'The Importance of eHealth in Europe,'" World Health, no. December, 2003.

[2] L. M. Hernandez, "Health Literacy, e-Health, and Communication: Putting the Consumer First: Workshop Summary." 2009.

[3] J. M. McGrath, N. H. Arar, and J. A. Pugh, "The influence of electronic medical record usage on nonverbal communication in the medical interview.," Health Informatics J., vol. 13, no. 2, pp. 105–118, 2007.

[4] L. Kumar, K. Sharanie, and D. Jaspaljeet, "iMedPub Journals Barriers to Adoption of Consumer Health Informatics Applications for Health Self-Management Abstract," Heal. Sci. J. ISSN 1791-809X, vol. Vol. 9 No., pp. 1–7, 2015.

[5] M. B. Alazzam et al., "Ehrs Acceptance in Jordan Hospitals By Utaut2 Model: Preliminary Result," J. Theor. Appl. Inf. Technol., vol. 3178, no. 3, pp. 473–482, 2015.

[6] A. I. Muhamad, M. R. M. Rosman, M. I. Ramzi, and M. I. M. Salleh, "Conceptualizing Medical Application Software for Managing Electronic Health Records (EHR) and Cash Flow Management in Private Clinics," Int. J. Innov. Manag. Technol., vol. 3, no. 2, pp. 151–155, 2012.

[7] K. Noraziani et al., "An overview of electronic medical record implementation in healthcare system: Lesson to learn," World Appl. Sci. J., vol. 25, no. 2, pp. 323–332, 2013.

[8] HealthInsight Utah, H. Mexico, and Others, "PHR Ignite Environmental Scan : The Personal Health Record Landscape of Utah and New Mexico," no. 212050, 2013.

[9] G. Demiris, "New era for the consumer health informatics research agenda," vol. 1, no. 1, pp. 13–16, 2012.

[10] Moh, "Dealing With Adolescents Using the Headss Assessment," Malaysia's Heal., pp. 89–96, 2008.

[11] S. Ponnudurai, "Mobile Health Records in Malaysia," Asia Biotech, vol. 14, no. 5&6, pp. 32–33, 2010.

[12] Norton Rose Fulbright, "Global Data Privacy Directory," Glob. data Priv. Dir., vol. July, 2014.

[13] T. J. Kobus III and G. S. Zeballos, "2014 International Compendium of Data Privacy Laws," 2015.

[14] L. Calvin et al., "Factors affecting home care patients ' acceptance of a web-based interactive self-management technology," pp. 51–59.

[15] MCMC, "Communications and Multimedia Pocket Book of Statistics," J. Chem. Inf. Model., vol. 53, no. 9, pp. 1689–1699, 2015.

[16] E. Agrawal, "ACCEPTANCE AND USE OF PERSONAL HEALTH RECORD: FACTORS AFFECTING PHYSICIANS' PERSPECTIVE," Indiana Univ., vol. December, p. 113, 2010.

[17] J. Barlow, W. Crawford, and D. Lansky, The value of personal health records. A joint position statement for consumers of health care., vol. 137. 2008.

[18] J. Cruickshank, "Putting patients in control ? Personal Health Records," no. September, 2012.

[19] Daglish E, "a Matter of Trust Electronic Personal Health Records : a Matter of Trust," McMaster Univ., p. 149, 2013.

[20] Miller H.D, Yasnoff W.A, and B. H.A, "Personal health records: the essential missing element in 21st century healthcare.," Healthc. Inf. Manag. Syst. Soc., 2009.

[21] K. Jeongeun, "The Personal Health Record," Heal. Informatics Res., vol. 107, no. 9, p. 27, 2011.

[22] Markle Foundation, "The Personal Health Working Group Final Report," Markle Found., p. 58, 2003.

[23] L. A. Wiedemann, "Practice brief. The role of the personal health record in the EHR," J Ahima, vol. 76, no. 7, p. 64A–64D, 2005.

[24] NAHIT, "Defining Key Health Information Technology HIT Terms," Natl. Coord. Heal. Inf. Technol., 2008.

[25] C. J. Gearon, "Perspectives on the future of personal health records," no. June, pp. 1–29, 2007.

[26] M. Bliemel and K. Hassanein, "Consumer satisfaction with online health information retrieval: a model and empirical study," E-Service J., vol. 5, no. 2, pp. 53–83, 2006.

[27] V. Venkatesh, J. Thong, and X. Xu, "CONSUMER ACCEPTANCE AND USE OF INFORMATION TECHNOLOGY : EXTENDING THE UNIFIED THEORY," MIS Q., vol. 36, no. 1, pp. 157–178, 2012.

[28] V. Venkatesh, F. D. Davis, and M. G. Morris, "Dead Or Alive? The Development, Trajectory And Future Of Technology Adoption Research," J. Assoc. Inf. Syst., vol. 8, no. 4, pp. 267–286, 2007.

[29] F. Davis, R. Bagozzi, and P. Warshaw, "User acceptance of computer technology: a comparison of two theoretical models," Management science, vol. 35, no. 8. pp. 982–1003, 1989.

[30] I. Ajzen, "The theory of planned behavior," Orgnizational Behav. Hum. Decis. Process., vol. 50, pp. 179–211, 1991.

[31] V. Venkatesh, M. Morris, G. Davis, and Fred Davis., "User acceptance of information technology: Toward a unified view," MIS Q., vol. 27, no. 3, pp. 425–478, 2003.

[32] A. Chang, "UTAUT AND UTAUT 2 : A REVIEW AND AGENDA FOR FUTURE RESEARCH," J. WINNERS , Vol., vol. 13, no. 9, pp. 106–114, 2012.

[33] M. B. Alazzam, A. S. H. Basari, A. S. Sibghatullah, and M. Doheir, "Review of Studies with UTAUT2 as Conceptual Framework," MAGNT Res. Rep., vol. 3, no. 3, pp. 620–629, 2015.

[34] R. W. Rogers, "Cognitive and physiological processes in attitude change: A revised theory of protection motivation," Soc. Psychophysiol., no. July, pp. 153–176, 1983.

[35] Y. Sun, N. Wang, X. Guo, and Z. Peng, "Understanding the Acceptance of Mobile Health Services: a Comparison and Integration of Alternative Models," J. Electron. Commer. Res., vol. 14, no. 2, pp. 183–200, 2013.

[36] A. Strauss and J. Corbin, Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory, vol. 3. 2008.

A Multi-Level Process Mining Framework for Correlating and Clustering of Biomedical Activities using Event Logs

[1]Muhammad Rashid Naeem
School of Computer Science and Technology,
Harbin Institute of Technology, Harbin, China
School of Software Engineering
Chongqing University, Chongqing, China

[2]Hamad Naeem, [3]Muhammad Aamir
School of Computer Science and Technology
Sichuan University, Chengdu, China

[4]Waqar Ali
School of Computer Science and Technology,
Nanjing University of Science and Technology
Nanjing, China

[5]Waheed Ahmed Abro
School of Computer Science and Engineering
Southeast University, Nanjing, China

***Abstract*—Cost, time and resources are major factors affecting the quality of hospitals business processes. Bio-medical processes are twisted, unstructured and based on time series making it difficult to do proper process modeling for them. On other hand, Process mining can be used to provide an accurate view of biomedical processes and their execution. Extracting process models from biomedical code sequenced data logs is a big challenge for process mining as it doesn't provide business entities for workflow modeling. This paper explores application of process mining in biomedical domain through real-time case study of hepatitis patients. To generate event logs from big datasets, preprocessing techniques and LOG Generator tool is designed. To reduce complexity of generated process model, a multilevel process mining framework including text similarity clustering algorithm based on Levenshtein Distance is proposed for event logs to eliminate spaghetti processes. Social network models and four distinct types of sub workflow models are evaluated using specific process mining algorithms.**

***Keywords*—biomedical event data; business process modeling; Levenshtein similarity clustering; multilevel process mining; spaghetti process models**

I. Introduction

Information retrieval is a big challenge in IT as the data is rapidly increasing day by day. The growth of data and technology are incredibly high resulting in business process management as a major problem within organizational entities. Today Business processes are more twisted and timely changing compared to old school of thoughts. It is need of any organization to identify, monitor and ensure their business processes are running accordingly to their workflow structure to prevent future losses [1]. Organizations are focusing more on business process improvement to increase concerns and success factors within business [2]. Event driven business process management has become one of emerging trends as it can be applied on businesses processes for compliance monitoring to analyze past business flaws and identify future risks [3]. Thus, process management has become a great importance to organizations and also a need of time in any business environment [4].

Hospitals have greater importance to process management as compared to any other organization as they have problems related to resources, cost and time management. Hospital systems are facing many challenges towards business process management because failure modes are intolerable in hospitals as it can put patients' life at stake. Patients' safety is also a critical factor directly linked to hospital business processes. In past, there are many incidents in medical history due to surgical mistakes or wrong treatment taken on patients possibly due to poor management or work pressure over resources. An abstract view of hospital system can be described using an enterprise architecture. Ahsan et al describes importance of healthcare enterprise architecture as it has more potential to facilitate healthcare units and business processes as a strategy to reduce critical factors and improve business processes [5]. Financial problems are also becoming major concerns in hospital systems. Freund describes a survey conducted by American College of Healthcare Executives and reviews taken from 338 executives of hospitals about hospital business concerns. Report highlighted financial concerns as one of the top concern in hospital managements which could result in resource and technology management problems [6]. A proper resource management can play a significant role to improve hospital services quality. Technology can also play important role to visualize hospital problems. Using latest ERP tools, fraction of business process can be utilized which are helpful in making future business decisions.

Biomedical processes are one of the most ignored parts in hospital business process management as they are mostly based on blood examinations taken on patient subjects. Biomedical processes are unstructured, twisted and based on time series, making it impossible for business analysts to do process modeling for them due lack of internal knowledge and code sequences. Another problem in biomedical process management is that there are thousands of biomedical experiments taken on patients on regular basis. Therefore process modeling for them can only be possible through an automatic process generation system.

II. Background and Motivation

Process mining is a new emerging trend in business process management. It provides different ways to extract business processes using knowledge management techniques without need of any background understanding of subject organization. For further discussion, this section is divided into two parts. At First, we provide a basic overview of process mining and its applications to business process management domain. Secondly, we evaluate business management problems in healthcare and importance of process mining to solve these problems.

A. Process Mining Overview

Process mining is the experimental, interdisciplinary scientific domain that provides popular algorithms to plot process models from event logs. Event logs consist of events which can be extracted from historical data i.e. ERP systems, distributed databases or SAP systems etc. [7]. As data mining provides prediction analysis from big data, likewise process mining prediction provides business process analysis from big event logs. Nowadays, Big data has become an essential ingredient of business process management to improve organization business entities but it has itself complexity issues. Therefore, enhanced business process management techniques such as process mining are required to provide proper validation and verification of business entities. Currently, three different types of process mining is being used in IT industry as shown in figure 1.

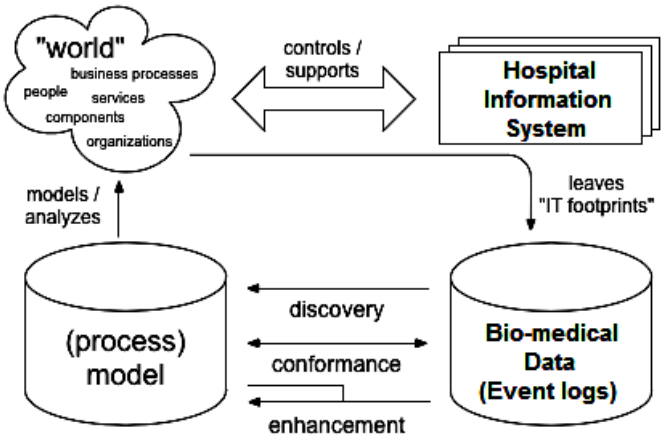


Fig. 1. Process mining as Discovery, Conformance and Enhancement

The diagram shows interaction of real world with healthcare information system. The real world always leaves footprints in IT systems' databases. For example, an employee has done "register" task which has been saved in database of IT system could be a footprint of that employee to "register" process for case "patient ID". We can further analyze these footprints to model the real world behavior using event logs as discovery, conformance and enhancement. In discovery, we generate process model from event log while conformance checking algorithms are used to replay that event log over generated process model to analyze compliance of process model with event log. Lastly, Enhancement algorithms are applied to enhance previously generated process models with respect to event log.

B. Process Mining Significance for Biomedical Processes

Biomedical data can also be referred as big data due to machine generated codes while stored data is consisting of textual information without any standardized terminology. There are difficulties and challenges of understanding and extracting useful information from biomedical data. Structure complexity issues also imposed limitations on biomedical data [8]. Extracting biomedical processes from biomedical data can only be possible using automatic process generation techniques i.e. process mining and process prediction etc. Many useful methods are discovered to analyze biomedical processes such as: McNames proposes use of biomedical filter to estimate event rates and extract point processes within in biomedical signals [9]. Augen proposes using of bioinformatics and information technology together has possibility to do process discovery in drugs [10]. Bose and Aalst describes process mining techniques can be used to extract non-trivial process related knowledge and analyze interesting insights to biomedical data which can later be used for performance analysis and other mathematical operations [11]. Petri net diagrams (also known as place/transition net) provide a graphical view to analyze business processes which are usually generated using process mining algorithms. Process mining i.e. discovery, conformance and enhancement could provide automatic analysis to biomedical process. Chaouiya describes petri nets emerged as promising tool to analyze biological networks efficiently [12]. Ferreira et al propose sequence clustering technique for bioinformatics to extract sequence behaviors using process mining [13]. Xing et al also propose an algorithmic approach to mine distributed bioinformatics workflows which can be applied within hospital systems to handle concurrence and recurrence of restricted bioinformatics workflow processes [14].

Hospital business processes are critical and changing timely, therefore process improvement is becoming requirement of time and one of main concerns of hospitals especially for healthcare technologies and informatics [15][16]. From previous observations, it can be concluded using process mining techniques, it is possible to make proper business process models from hospital data logs to visualize concerns especially in biomedical domain. On other hand, extracting biomedical processes is also a critical requirement of this century. Due to increase in biomedical equipment usage, biomedical record is rapidly converting into big data making it difficult to do estimations and performance analysis at business process level. Extracting useful information is now possible using number of open source and commercial information processing software's available but most of information gained is related to analytical estimations only. Extracting business processes from bioinformatics is a big challenge in biomedical domain which is main contribution of this paper.

III. Preprocessing of Data using *"LOG Generator"* Tool

In this section, we provide an overview of case study and techniques to convert non-event biomedical data into events by querying among multiple dataset tables. Secondly, we provide

"LOG Generator" tool algorithm which is used generate event log for from preprocessed event data.

A. Case study Overview

For biomedical case study, Hepatitis Patients' data has been selected for process mining which is taken from ECML/PKDD discovery challenge website [17]. The data set contains examinations of Hepatitis B and C on patients admitted to academic hospital. The dataset contains huge amount of time series experimental data. All experiments are taken from different laboratories and medical facilities between year 1978 and 2001.

Dataset consists of seven tables. First table "pt_ e030704" contains information related to patient's identity i.e. IDs, birth dates etc. "bio_e030704" table contains information about biopsy of patients. "ifn_e030704" table contains interferon therapy information performed on patients. "hemat_e030704" contains information about hematology experiments while "ilab_e030704" and "olab_ e030704" tables contain different experiments taken inside and outside hospital laboratories. Lastly, "labn_ e030704" table contains measurement units for in-hospital laboratories.

In-hospital data contains results of 230 distinct examinations while out-hospital data consists of 753 distinct examinations performed on 771 hepatitis patients admitted to academic hospital.

B. Preprocess Event Data

In preprocessing, we extract real time events from non-event data sources i.e. data table as we discussed in section 3.1 case study overview. There are four required elements of an event needed to be extracted to create a business event are "Case ID", "Activity", "Timestamp" and "Resource". Almost every distributed information system has fractions of events i.e. hospital patient's database, transaction log etc. Let's take example of hepatitis patients' data to illustrate an events extraction. A subject patient is a "Case ID", time is used as "Timestamp" and experiment name is an "Activity" while facility can be used as a "Resource" for that activity. There are five tables in hepatitis patients database from which events can be extracted by querying with patient info table. For this paper, we are using biopsy table to show events extraction practice from biomedical data logs. In figure 2, an event extraction technique for biopsy processes is presented.

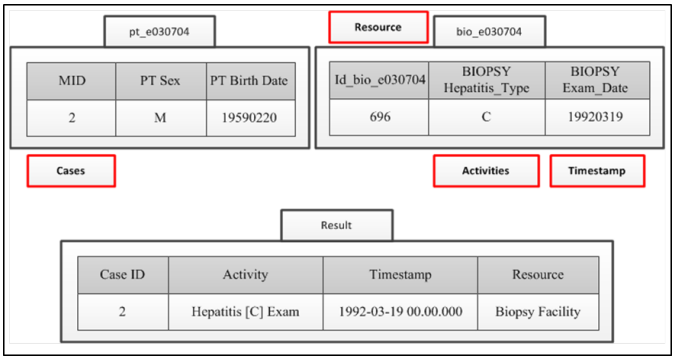


Fig. 2. A resultant event extracted by querying patient info and biopsy table

"MID" field of patient info table is treated as "Case ID" from "pt_e030704" table while "Biopsy Hepatitis_ Type" is used as "Activity" and "Biopsy Exam_ Date" is used as timestamp from "bio_e030704" table. There is no specific resource field provided in biopsy table, therefore "Biopsy Facility" is used as a resource for respective event. The Date is converted to timestamp format while "C" is name of blood experiment which become "Hepatitis [C] Exam" process. All remaining events are extracted from other tables using similar extraction technique.

After extraction, all event data is exported to sing *CSV* file collectively having more than 1.6 million events with 727 cases and 1067 distinct activities. The smallest case consists of 3 events while the longest case has 13257 events.

C. Event Log Generator Tool

Hepatitis patient's event data is huge in size and needed to be converted into XES (eXtensible Event Stream) event log format to apply process mining. XES is successor of old MXML (Mining eXtensible Markup Language) event log. Contrary to MXML, XES is simple without any restrictions of classifying data attributes. XES user manual and latest XES standard definitions are presented in [18]. There are also many XES event log generator tools available today and widely used for process mining. The most popular and open source XES event log generator tools are "openXES" and "XESame". There are some complexities of usage issues in these log generator tools i.e. java jdbc connections and query setting are required to build a successful event log which is too difficult for newbies and required deep understanding of tools to use them properly. The user manual guide for "penXES" and "XESame" log generator tools are presented in [19] and [20] respectively. Due to large size of data, generating an event log is not an easy task as it causes computer memory and application heap size problems in tools as well querying takes lot of time for generating event log which are not handled properly in mentioned tools. To tackle such problems a "LOG Generator" tool has been developed which uses same data setting of CSV event data file as described in 3.3 preprocess event data section. Our tool uses an openCSV api to read biomedical event data ensuring faster event log generation as it is designed to handlle large CSV datasets for read and write purposes. To elaborate internal structure of "LOG Generator" tool a pseudo code of EventBuilder algorithm is presented in 3.3.1.

"LOG Generator" is a user friendly tool and doesn't require any prior usability knowledge. *PrintWriter* and *PrintReader* utility classes are used for reading and writing data in XES event log format. Final output of tool is *Eventlog.xes* file which is an XES format event log file have compatibility to any popular process mining tool to apply process mining algorithms. The time format of biomedical data is usally in *YYYYMMDD* format which is not a standard in any of widely used programming languages. For this purpose a *TIMESTAMP-CONVERTER* procedure is designed to convert date into *yyyy-mm-ddT hh:mm:ss.sss+GMT* timestamp format which is used as a standard in Process mining. The meta structure of event log generated using "LOG Generator" tool is shown in table 1. The average time tool takes to generate 1.6 million events log is 24 secs.

Algorithm 3.3.1 : EVENTBUILDER (csv)

procedure $TIMESTAMPCONVERTER\,(n,\ D)$
$f \leftarrow DATEFORMATER\,(`yyyy - mm - dd'\mathrm{T}'hh : mm : ss.sss + GMT')$
$t \leftarrow \{\}$
if $n \geq 8$
then $\begin{cases} y \leftarrow \{D_0,\ ...,\ D_3\} \\ m \leftarrow \{D_4,\ D_5\} \\ d \leftarrow \{D_6,\ D_7\} \\ t \leftarrow CONCAT(y, '-', m, '-', d) \end{cases}$
$timestamp = SETFORMAT(f, t)$
return $(timestamp)$
main
$DT \leftarrow READALL(csv)$
$SORT(DT)$
$size \leftarrow LENGTH(DT)$
$output \leftarrow CONCAT(output, '<LOG>')$
$x \leftarrow 0$
while $x \leq size$
do $\begin{cases} e \leftarrow DT[x] \\ \textbf{if } x = 0 \\ \quad \textbf{then } \{\ output \leftarrow CONCAT(ouput, '<trace>') \\ e_{time} \leftarrow TIMESTAMPCONVERTER\,(LENGTH\,(e_{time}), e_{time}) \\ event \leftarrow CONCAT('<event>', '<e_{caseID}>', '<e_{activity}>', '<e_{time}>', '<e_{resource}>', '</event>') \\ \textbf{if } x + 1 > x \textbf{ and } x \neq \{0 \textbf{ or } size\} \\ \quad \textbf{then } \begin{cases} output \leftarrow CONCAT(output, event) \\ output \leftarrow CONCAT(output, '</trace><trace>') \end{cases} \\ \textbf{else} \\ \quad \textbf{then } \{\ output \leftarrow CONCAT(output, event) \\ \textbf{if } x = size \\ \quad \textbf{then } \{\ output \leftarrow CONCAT(ouput, '</trace>') \\ x \leftarrow x + 1 \end{cases}$
$output \leftarrow CONCAT(output, '</LOG>')$
$WRITEFILE(ouput\ as\ \textbf{'EventLog.xes'})$

IV. Complexity Analysis and Proposed Framework

For complexity analysis and process model generation, event log is imported to Prom6 open source tool. We use heuristic miner algorithm to generate process model as it works better with larger size event logs. Heuristic miner uses frequencies and sequences of events by ignoring infrequent paths. It uses casual dependencies and AND/XOR split joins to construct process models [21]. To mine dependencies using heuristic miner, following equation is presented.

$$a \Rightarrow_W b = \left(\frac{|a >_W b| - |b >_W a|}{|a >_W b| + |b >_W b| + 1} \right) \tag{1}$$

Where $a \Rightarrow_W b \in \{-1, 1\}$ and $a, b \in T$ while "W" is an event log over trace "T"

For all non-observable tasks, depending relation equation is used in heuristic miner as follows:

TABLE I. Meta Format of XES Event Log Generated using LOG Generator Tool

```
<log xes.version= "1.0" xmlns= "http://www.xes-standard.
org" xes.creator= "Log Generator Tool">
<!– Extensions –>
<global scope= "trace">
<string key= "concept:name" value= "name"/>
</global>
<global scope= "event">
<string key= "concept:name" value= "name"/>
<string key= "org:resource" value= "resource"/>
<string key= "lifecycle:transition" value= "transition"/>
<date key= "time:timestamp" value= "2001-04-14T05:40:
17.017+8:00"/>
<string key= "Activity" value= "string"/>
<string key= "Resource" value= "string"/>
</global>
<trace>
<string key= "concept:name" value= "1"/>
<string key= "creator" value= "LOG Generator Tool"/>
<!– Events –>
<event>
<string key= "concept:name" value= "Hepatitis [C] Exam"
/>
<string key= "org:resource" value= "Biopsy Facility"/>
</event>
<!– Traces –>
</trace>
</log>
```

$$a \Rightarrow_W b \wedge c = \left(\frac{|b >_W c| - |c >_W b|}{|a >_W b| + |a >_W c| + 1} \right) \tag{2}$$

Where $a \Rightarrow_W b \in \{-1, 1\}$ and $a, b, c \in T$, "W" is an event log over trace "T" while "b" & "c" are in depending relationship with "a"

Correctness/Fitness of generated model is measured using Continuous Parsing Measure (CPM) and equation of CPM is given as:

$$CPM = \frac{1}{2}\frac{(e - m)}{e} + \frac{1}{2}\frac{(e - r)}{e} \tag{3}$$

Where e = events, m = total no. of missing activated inputs and r = no. of remaining activated outputs

One of the main advantages of heuristic miner is it deals with noise and exceptions efficiently and focuses on main process workflow instead of mapping every possible path resulting in reduced spaghetti processes. Testing conformance of event log with process model i.e. replaying all events over process model is also memory and time consuming while heuristic miner uses dependency alignments for each dependency to calculate fitness of generated model. Process model generated using heuristic miner having fitness of 0.78 is shown in figure 3.

Fig. 3. Spaghetti process model generated using heuristic miner

The processes and nodes in figure 3 are twisted and too many giving it a spaghetti look. Therefore current process model can't be used as business process management solution. To resolve this issue, a clustering strategy based multilevel processing framework for event logs is proposed to eliminate these spaghetti processes.

A. Multi-Level Process Mining Framework

To eliminate spaghetti processes and based on structure of biomedical activities, we propose multi-level process mining framework for event logs to generate multi process models shown in figure 4. Clustering of similar and less-frequent processes with removal of non-frequent activates is applied to reduce complexity of main workflow process model. Workflow models are further divided into multi sub models based on their complexities. The main workflow process model will be "Level 0" in framework while corresponding models are "1" ,"2" and so on.

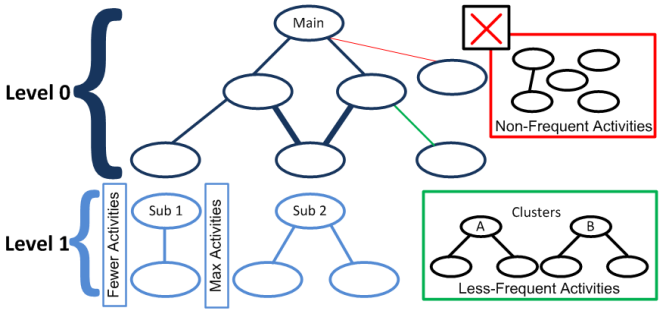


Fig. 4. A multi-level process mining framework to correlate & cluster event logs

Sub process models are further classified into four distinct groups of categories based on frequency of activities. First type of process model consists of fewer activities which mean resulting model will be sound. Second type of process model has extreme number of activities where total number of activities is also huge. The complexity of such process models is same as of main workflow model therefore they are further simplified using "Level 0" techniques. Other types of activities are those having fewer activities as well as non-frequent activities. We further elaborate different properties of framework in next sub sections.

1) Removal of non-frequent activities: The non-frequent activities will be removed from event log as they cant be used as a business activity. For example in hepatitis patients' event log there are more than 100 activities whose occurrence is less than 0.00006% of the total event log. In general, they are automatically ignored by process mining algorithms when putted to noise threshold but it will also affect accuracy and fitness of process models. Therefore to prevent this, all those activities having frequency of occurrence less than equal to 100 are removed from event log.

2) Grouping of Less-Frequent Activities: Second strategy in framework is to reduce spaghetti processes by grouping of those activities having lesser frequency of occurrence in event log as one. For instance, we have created two groups named as "Cluster A" and "Cluster B" where "Cluster A" consists of activities whose frequency ranges between 101 to 1000 while "Cluster B" consists of activities whose frequency ranges between 1001 and 10000 collectively covering 17000 and 102000 events respectively.

3) Similar-Frequent Word Clustering Algorithm: To cluster set of similar and frequent activities within event log, we propose word similarity clustering algorithm for frequent activities. Firstly, all distinct activities in event log are extracted in separate CSV file. Then we use three strategical measures in our algorithm as follows: To extract all possible words within file using strings division, a "Term Frequency" formula is presented which calculate ratio of each term occurrence within CSV file.

$$tf(t,d) = \frac{f_d(t)}{\max\limits_{\omega \in d} f_d(\omega)} \tag{4}$$

Where "d" = document i.e. CSV file and $f_d(t)$ = frequency of term "t" in document "d"

To check similarity among extracted words, Levenshtein Distance wording matching technique is applied through cross multiplication matching between distinct activities in CSV file and terms extracted form sub division of CSV data fields. The documentation for Levenshtien Distance is presented in [22] and formula for calculation of Levenshtien Distance is given as follows:

$$lev_{a,b}(i,j) = \begin{cases} \max(i,j) & if\ \min(i,j) = 0 \\ \min \begin{cases} lev_{a,b}(i-1,j)+1 \\ lev_{a,b}(i,j-1)+1 \\ lev_{a,b}(i-1,j-1)+1_{(a_i \neq b_j)} \end{cases} & else \end{cases} \tag{5}$$

Where "a" and "b" are two strings passed into the loops for similarity distance matching

After calculating distance of all elements in loops, clusters will be made of those having less distance among matching string while duplicate values will be removed by using sorting and distinct value selection techniques as given below:

$$f_{(a,b)}(x) = \left\{ \begin{array}{l} SORT(x \Rightarrow x[a]) \\ DISTINCT(x \Rightarrow x[a]) \end{array} \right\} \tag{6}$$

Where "a" and "b" are cells in multi valued array "f" while sorting and distinct selection is made using "a" column of array.

The detailed pseudo code for clustering and correlating events for event log is described in algorithm 4.4.1.

Algorithm 4.1.1 : FREQCLUSTERBUILDER (csv)

procedure $TERMEXTRACTOR\,(t)$
$t \leftarrow VALIDATEREGEXP\,(t, `[\hat{}A-Za-z0-9]')$
$tf_{1,...,n} \leftarrow SPLITER(t)$
return (tf)
main
$a \leftarrow READALL(csv)$ **where** $csv = \{a_1, a_2, a_3, ..., a_n\}$
$size \leftarrow LENGTH(DT)$
$b \leftarrow \{\}$
$x, y \leftarrow 0$
while $x \leq size$

do
- $t \leftarrow TERMEXTRACTOR\,(DT[x])$
- **while** $y \leq LENGTH(t)$
  - **do**
    - $ADD(a,\ t[y])$
    - $y \leftarrow y + 1$
- $x \leftarrow x + 1$

$a = DISTINCT(a)$ **(i)**
$b = DISTINCT(b)$ **(ii)**

Comment: Now applying cross multiplication among *"a"* and *"b"* arrays where a = {all distinct extracted terms} & b = {all distinct extracted sub terms}

$x, y \leftarrow 0$
$output \leftarrow \{\}$
while $x \leq LENGTH(a)$

do
- **while** $y \leq LENGTH(b)$
  - **do**
    - **if** $b[y].FINDIN(a[x])$
      - **then**
        - $d = \begin{cases} \text{LEVDISTANCE}(a[x], \\ LENGTH(a[x],), \\ b[y],\ LENGTH(b[y])) \end{cases}$
        - $ADD(output, \{a[x],\ b[y],\ d\})$
    - $y \leftarrow y + 1$
- $x \leftarrow x + 1$

Comment: Now sorting output by *"a"* then *"d"* and selecting first row based on *"a"* from *"output"* array.

$SORT(output,\ a \Leftarrow a[0],\ a[2])$
$output \leftarrow DISTINCT(output,\ a \Leftarrow FIRST(a))$
$WRITEFILE(output\ as\ \textbf{csv}_{\textbf{(a,b,d)}})$

After executing algorithm 4.1.1, a CSV file is returned containing three columns "a", "b" & "d", where "a" contains biomedical activities while "b" contains clusters of these biomedical activities in "a". "c" contains distance measured using Lev. distance methodology. Clusters are applied to 1.6 million events using database queries and sample of resultant event data is shown in table 2.

TABLE II. RESULT OF CLUSTER ALIGNMENTS OF EVENTS BY USING DATABASE QUERYING

Case ID	Activity	Cluster	Timestamp	Resource
72	HCV-AB	HCV	1993-08-09T...	ILab Facility
90	B-LIP	Cluster B	1994-06-01T...	ILab Facility
100	ZTT	ZTT	1994-08-24T...	OLab Facility
701	Hepatitis C	Cluster A	1995-08-01T...	Biopsy Facility
752	U-K	U-	1994-10-24T...	ILAB Facility
757	DNA-II	DNA	1997-04-07T...	OLab Facility
.....n	.....n	.....n	.....n	.....n

To visualize graphical view of events' variations due to framework approaches, *x-axis* and *y-axis* are randomly assigned to nearby cluster points by treating each cluster as centroid on rapid miner scatter graph visualizer. Non-frequent, less-frequent and clusters are visualized separately in figure 5 a, b and c.

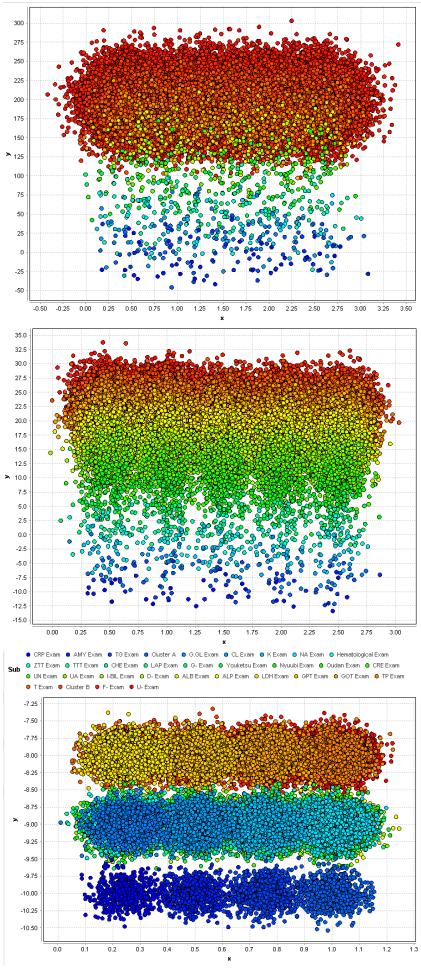


Fig. 5. **(a)**Initial: Event data having both non-frequent and less-frequent activities **(b)** Event data after removal of non-frequent and grouping of less-frequent activities **(c)** Event data after applying similar-frequent clustering algorithm having fewer groups of activities

V. CASE STUDY ANALYSIS

There are five different types of workflow process models generated using multi model process mining framework approaches. At "Level 0", main workflow model is derived while at "Level 1", four distinct types of sub workflow models are generated. They are briefly discussed in coming sections.

A. Main workflow process model

To generate workflow process model, we use same heuristic miner algorithm as described in section 4 complexity analysis part. The generated workflow model after using framework approaches is shown in figure 6.

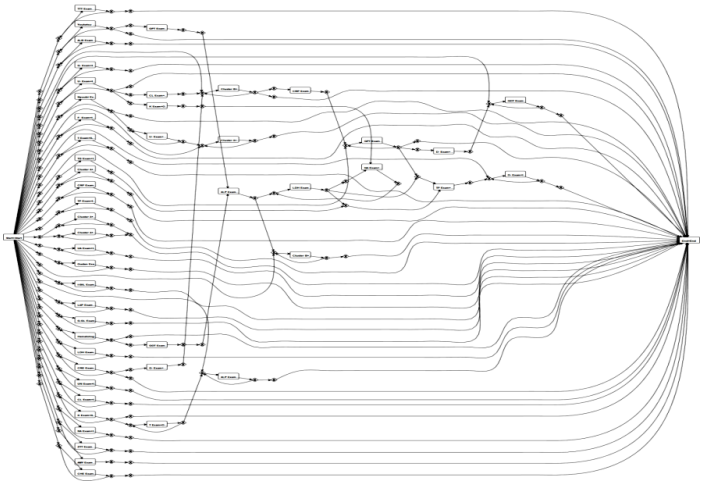


Fig. 6. Main BPMN workflow model generated using heuristic miner through framework approaches

Difference between two diagrams is clearly seen in both figure 3 and 6. Before applying process mining framework, figure 3 illustrates spaghetti and cohesive processes. Conversely, Figure 6 is generated through framework approaches on event log uttering clear workflow nodes with visible spits and joins making it an effective model for biomedical BPM solution. Fitness is another major factor associated with usability and trustworthiness of process models. The fitness of this model is also increased to 0.954 compared to process model in figure 3 which is 0.78 on a scale of 1.

B. Sub workflow process models

Sub workflow process models consist of pertinent spaghetti processes eliminated from main workflow process model and fall in "Level 1" of multi-level process mining framework. Based on hepatitis patients data, several sub workflow models can be derived from event logs where some of them have identical criticality as of main workflow process model. Some of them are comprising of fewer activities and doesnt required any additional process modeling. Based on nature of event logs, workflow models are further divided into four distinct groups as follows:

1) events with fewer event classes workflow model: The first group of event logs has small number of events with fewer event classes. We take "DNA Exam" event log for instance. Event log consists of 206 cases. The timestamp falls between years 1984 to 2000. Event log have 1161 events with 18 event classes and one originator (resource). The process model shown in figure 7 is derived using "Mine Petri Net using Visual Inductive Miner" algorithm provided in Prom6 tool. Visual inductive miner ensures number of input token within petri net are equal to no of output tokens hence ensuring maximum fitness [23]. Noise threshold is added to 0 to map all possible transitions resulting in fitness of derived model to 1.

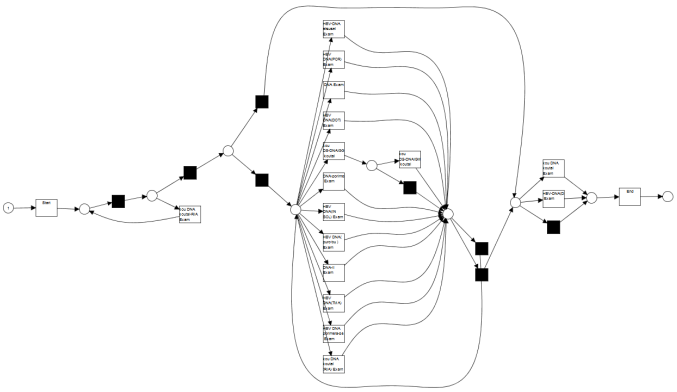


Fig. 7. Petri net sub workflow process model derived from DNA event log using inductive miner

2) Fewer events with extreme event classes workflow model: The second group of event logs has less number of events with extreme number of distant event classes. "Cluster A" event log have fewer events and extreme event classes as it consists of business activities having event frequency rate less than 1000. We apply same "Mine Petri Net using Visual Inductive Miner" algorithm in Prom6 and add noise threshold to 60% to avoid spaghetti processes. Due to 60% noise threshold, the generated model has too many deviations and fitness of derived model is not as much to be used as business process management solution.

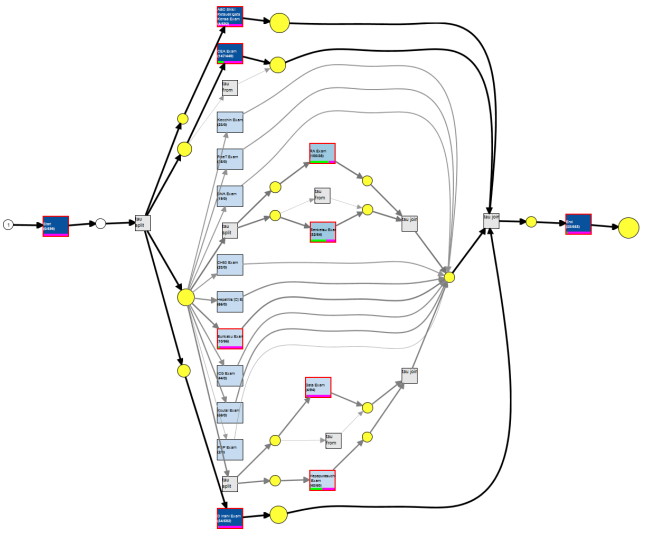


Fig. 8. Performance analysis of petri net sub workflow process model derived from "Cluster A" event log using "Align Log and Model Repair Model for Repair (global Cost)" algorithm

To estimate fitness and fitness cost for each missing token in petri net workflow process model, we use "Align Log and Model Repair for Repair (global Cost" algorithm for performance analysis as shown in figure 8. The repaired workflow model has conformance for only 714 traces while log fitness is 0.57. The calculated global repair cost for the model is 1.74 which is based on number of nonaligned tokens.

Using global cost repair algorithm, It is also possible to trace individual case or individual activity to provide operational support towards trace completion using petri nets.

3) Extreme events with fewer event classes workflow model: Third group of event logs consist of extreme number of events with fewer event classes. For example "F-Exam" event log has 150932 events and 8 distinct event classes. As there are few event classes, therefore generated model will be simple and clear. Noise threshold is also unnecessary due to less number of activities. Therefore, resulting model fitness will be 1 as it covers all paths and activities with in event log similar to figure 7.

4) Extreme events with extreme event classes workflow model: The fourth and last group of event logs consist of extreme number of events with extreme distinct event classes. As the complexity of such event logs is same as of main event log for which multi-level process mining framework is proposed. Therefore it can be further customized using any of three complexity analysis techniques proposed in section 4. Fuzzy miner algorithm can also be useful for such event logs as it makes mini clusters within generated fuzzy model to enhance model detailed view [24]. For instance, we use "U- Exam" event log covering 704 cases 33 event classes and 164095 events. It is also the largest cluster in hepatitis patient's event log. "Mine for fuzzy model" algorithm is used by applying fuzzy inputs to balance fitness and detailed view. Generated fuzzy model is shown in figure 9 a with detailed view of one cluster in figure 9 b. "Model Detail" is 92.68% and "Model Conformance" is 87.26%. Based on proposed framework the cluster falls in "Level 2" of multi-level process mining framework.

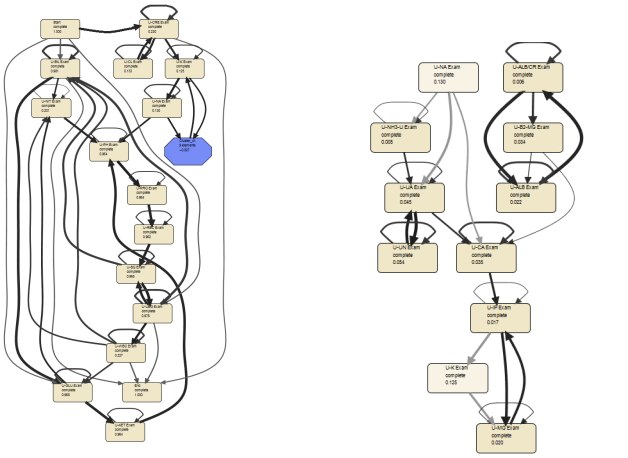


Fig. 9. **(a)** Fuzzy workflow process model generated from "U-" Event log **(b)** Detailed view of fuzzy cluster (blue color business activity) shown in figure 15(a)

C. Social Network Analysis

Organizational perspective plays an important role in applying proper business process management within any organization. Social network models can be used to visualize social behavior and work distribution among organizational resources. Using social network algorithms provided in process mining tools, it is possible to visualize social behavior. Aalst et al terms mining social networks as a virtuous and cost-effective business process analysis technique if combined workflow model concepts with social network analysis to build social network based on hand over work from one performer to another [25].

In biomedical business process environments social networks can be useful to analyze workloads and interaction within hospital resources. As hospital resources strongly dependent on each other i.e. a biopsy cannot be performed without biomedical experimental results, therefore a proper resource management is required to make any decision on resources workload. To visualize organizational perspective, a social network based workflow model is generated using DISCO process mining tool and presented in figure 10.

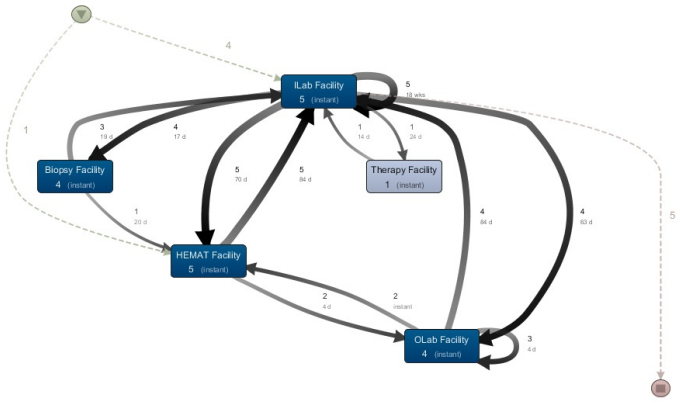


Fig. 10. Visualization of resources executing business activities and their responsibilities using social network mining

VI. Conclusion

Current research focused on implication of process mining in biomedical domain. For case study, an unstructured and time series hepatitis patients' data of academic hospital is utilized. Preprocessing techniques are proposed for events extraction and event logs are generated using "LOG Generator" tool designed to handle huge datasets. To reduce complexity and spaghetti processes in workflow process model, a multi-level process mining framework is envisioned with complexity reduction techniques and a clustering algorithm. Through framework, distant activities are clustered while less-frequent are removed from event logs. The resultant model has shown comprehensible detailed view with greater fitness. Besides, four distinct groups of event logs are elaborated using process mining algorithms to generate sub process models with optimal soundness. Social network model is presented to illustrate organizational behavior and resources. The methods and techniques proposed in research work are helpful for scientific community to apply automatic process modeling from complex and huge bio-medical datasets. In future work, we will extend and apply these methods to solve process modeling problems for other complex and similar series domains.

References

[1] W. X. Mu, F. Benaben, and H. Pingaud, "A Methodology Proposal for Collaborative Business Process Elaboration Using a Model-Driven Approach", *Enterprise Information Systems*, 2015, Vol. 9 No. 4, pp. 349-383

[2] Y. Liu and R. Aron, "Organizational Control, Incentive Contracts, and Knowledge Transfer in Offshore Business Process Outsourcing", *Information Systems Research*, 2015, Vol. 26 No. 1, pp. 81-99

[3] F. Koetter and M. Kochanowski, "A Model-Driven Approach for Event-Based Business Process Monitoring", *Information Systems and E-Business Management*, 2015, Vol. 13 No. 1, pp. 5-36

[4] V. Rajarathinam, S. Chellappa, and A. Nagarajan, "Conceptual Framework for The Mapping of Management Process with Information Technology in a Business Process", *Scientific World Journal*, 2015, No. l983832

[5] K. Ahsan, H. Shah, and P. Kingston, "Patients' Processes in Healthcare: An Abstract View through Enterprise Architecture", *In: Proc. Int. Conf. on Information Management and Evaluation*, Cape Town 2010, pp.459-466

[6] L.M. Freund, "American College of Healthcare Executives Announces Top Issues Confronting Hospitals: 2013", retrieved on 07/03/2015 from http://www.ache.org/pubs/Releases/2014/top-issues-confronting-hospitals-2013.cfm

[7] W.M.P van der Aalst, "Process Mining: Discovery, Conformance and Enhancement of Business Process", *Springer Verlag*, Eindhoven 2011

[8] R.R. Brinkman, M. Courtot, and D. Derom et al, "Modeling Biomedical Experimental Processes with OBI", *In Proc. Of the Bio-Ontologies: Knowledge in Biology*, Stockholm 2009, Vol. 1 Supp. 1, pp. 1-11

[9] J. McNames, "Optimal Rate Filters for Biomedical Point Processes", *In Conf. Proc. of Engineering in Medicine and Biology Society EMBS*, Shanghai 2006, pp. 145-148

[10] N.S. Buchan, D.K. Rajpal, Y. Webster, and C. Alatorre, "The role of translational bioinformatics in drug discovery", *Drug Discovery Today*, 2011, Vol. 16 No. 9, pp. 426-434

[11] R.P.J.C. Bose and W.M.P. van der Aalst, "When Process Mining Meets Bioinformatics; IS Olympics", *Information Systems in a Diverse World*, 2012, Vol. 107, 202-217

[12] C. Chaouiya, "Petri Net Modeling of Biological Networks", *Briefings in Bioinformatics*, 2007, Vol. 8 No. 4, 210-219

[13] D. Ferreira, M. Zacarias, M. Malheiros, and P. Ferreira, "Approaching Process Mining with Sequence Clustering: Experiments and Findings", *Lecture Notes in Computer Science*, 2007, Vol. 4714, 360-374

[14] J. Xing, Z. Li, Y. Cheng, and F. Yin, "Mining Process Models from Event Logs in Distributed Bioinformatics Workflows, *In Proc. of the 1st Int. Symposium on Data, Privacy and E-Commerce*, Chengdu 2007, pp. 8-12

[15] M.H. Yarmohammadian, H. Ebrahimipour, and F. Doosty, "Improvement of Hospital Processes through BPM in Qaem Teaching Hospital: A Work in Progress", *Journal of Education and Health Promotion*, 2014, Vol. 3 No. 111, pp. 1-10

[16] F. Ruiz, F. Garcia, L. Calahorra, and C. Llorente, et al, "Business Process Modeling in Healthcare", *Studies in Health Technology and Informatics*, 2012, Vol. 179, pp. 75-87

[17] S. Hirano and S. Tsumoto, "Guide to Hepatitis Data for ECML/PKDD 2005 Discovery Challenge", retrieved on 05/02/2015 from http://lisp.vse.cz /challenge/index.html

[18] C.W. Gunther, H.M.W. Verbeek, "XES-Standard Definition. v. 2.0", 2014

[19] C.W. Gunther, "penXES Developers Guide. 1.0 RC5", 2009

[20] H.M.W. Verbeek, J.C.A.M. Buijs, B.F. van Dongen, and W.M.P van der Aaist, "XES, XESame and Prom 6", *Lecture Notes in Business Information Processing*, 2010, Vol. 72, pp. 60-75

[21] A.J.M.M Weijters, W.M.P. van der Aalst, A.K. Alves de Medeiros, "Process Mining with the Heuristics Miner-Algorithm", *Technology University Eindhoven Tech. Rep.*, Eindhoven 2006, WP 166, pp. 1-34

[22] V. Pieterse and P.E. Black, "Levenshtein distance.: Algorithms and Theory of Computation Handbook", retrieved on 05/03/2015 from http://www.nist.gov/dads/HTML/Levenshtein.html

[23] J.J.L Sander, D. Fahland, W.M.P. van der Aalst, "Process and Deviation Exploration with Inductive Visual Miner", BPM Demos 2014

[24] C.W. Gunther and W.M.P. Van der Aalst, Fuzzy Mining Adaptive Process Simplification Based on Multi-Perspective Metrics, *Lecture Notes in Computer Science*, 2007, Vol. 4714, 328-343

[25] W.M.P. van der Aalst, H. Reijers, and M. Song, "Discovering Social Networks from Event Logs", *Computer Supported Cooperative Work*, 2007, Vol. 14 No.6, pp. 549-593

Area and Energy Efficient Viterbi Accelerator for Embedded Processor Datapaths

Abdul Rehman Buzdar*, Liguo Sun*, Muhammad Waqar Azhar‡, Muhammad Imran Khan†§, Rao Kashif†
*Department of Electronic Engineering and Information Science
†Micro/Nano Electronic System Integration R & D Center (MESIC)
University of Science and Technology of China (USTC), Hefei, China
‡Department of Computer Science and Engineering, Chalmers University of Technology, Gothenburg, Sweden
§Department of Electronics Engineering, University of Engineering and Technology Taxila, Pakistan

***Abstract*—Viterbi algorithm is widely used in communication systems to efficiently decode the convolutional codes. This algorithm is used in many applications including cellular and satellite communication systems. Moreover, Serializer-deserializers (SERDESs) having critical latency constraint also use viterbi algorithm for hardware implementation. We present the integration of a mixed hardware/software viterbi accelerator unit with an embedded processor datapath to enhance the processor performance in terms of execution time and energy efficiency. Later we investigate the performance of viterbi accelerated embedded processor datapath in terms of execution time and energy efficiency. Our evaluation shows that the viterbi accelerated Microblaze soft-core embedded processor datapath is three times more cycle and energy efficient than a datapath lacking a viterbi accelerator unit. This acceleration is achieved at the cost of some area overhead.**

***Keywords*—Viterbi decoder; Codesign; FPGA; MicroBlaze; Embedded Processor**

I. Introduction

Channel coding is used in wireless communication systems for reliable data transfer over noise prone communication channels. Various forward error correction (FEC) schemes e.g. Low-density parity-check (LDPC), Reed Solomon, Viterbi and Turbo codes are used to meet the growing need to improve the spectrum efficiency [1], [2], [3], [4], [5]. In FEC schemes the encoding of data is done using convolutional encoding and at the receiver end the decoding process is done by viterbi or turbo decoders [21-31]. The viterbi decoder is suitable in wireless communication systems in which the transmitted signals are corrupted by additive white Gaussian noise [6].

The decoding process in FEC schemes is computationally intensive and power hungry. The hand held devices are battery powered, so they must be energy efficient. The customized hardware implementation of these FEC decoders are performance and power efficient but lacks flexibility. As the wireless standards evolve with time, so the hardware needs to be flexible. The viterbi decoder can be implemented in software and executed on an embedded processor but it will require a lot of clock cycles. The viterbi decoder can be implemented more efficiently in dedicated hardware which will require few clock cycles at the cost of flexibility. The high speed communication systems today requires fast data rates which can only be delivered using dedicated hardware solutions.

Different hardware modules like USB, Ethernet, TCP/IP, CRC and CAN protocol are included in modern embedded processors [7], [8], [9] to speedup certain parts of application in areas like signal processing, communication and control systems. This provides effective use of viterbi accelerator in programming systems where a series of viterbi decoding is required to be computed.

II. Convolutional Encoding and Viterbi Decoding

Convolutional encoding of data is implemented with a shift register having $K-1$ memory elements and cascaded network of exclusive-or gates. Here K is the constraint length and having 2^{K+1} encoder states. The shift register is a chain of flip-flops and the output of nth flip-flop goes as input into the $(n+1)th$ flip-flop. The data in the registers is shifted to the next register and the value in the last register gets discarded. The combinational logic consisting of exclusive-or gates is used to perform modulo-2 addition. The encoder outputs n symbols using generator polynomials and values in the shift register. Fig. 1 shows a convolutional encoder for $K = 3$, $R = 1/2$ and generator polynomials $G1 = (1, 1, 1)$ and $G2 = (1, 0, 1)$. The code rate is the ratio of the number of input bits to the number of output bits $(R = m/n)$. The reason for the convolutional codes being efficient compared to block codes is the fact that every input bit has an impact on K successive output symbols [10]. The value of K is directly proportional to the code complexity and error correction capability. The decoder complexity and memory requirements increases with increasing K.

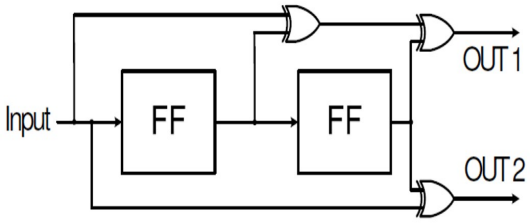


Figure 1: Convolutional Encoder general architecture.

Trellis diagram is used to visualize the state transitions of an encoder, as shown in Fig. 2. The black lines represent input bit 0 and the dotted lines represent input bit 1. The trellis path of input sequence is represented by the red lines. The basic concept is that the valid path through trellis diagram is

generated by the sequence of input bits from left to right. The viterbi decoder is able to find the valid path on trellis which is closest match when some transmission error occurs [11]. In start the reset state of encoder is "00". If the input is 0 the encoder state will become 00 as shown by the black line. The encoder will transmit 00 as output. The viterbi decoder reconstructs the valid input bit. If the input bit is 1 the decoder goes to state "10" and "11" is transmitted.

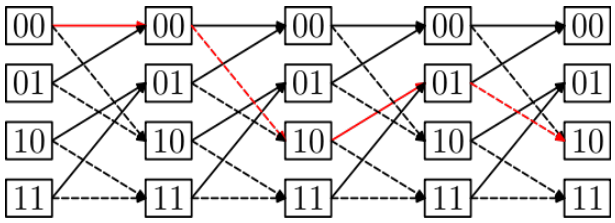


Figure 2: Trellis diagram.

The main concept of viterbi decoder is mapping received symbols to most likely valid sequence. The decoding process consist of following steps.

1) Branch Metric Unit: In this step difference is calculated between received symbol and every possible encoder output combinations. In hard decision decoder the difference is the Hamming distance and in case of soft decision decoder the Euclidean distance is used. There can be 2^K output combinations for an encoder with 2^{K+1} states and 0 or 1 as input bit.

2) Path Metric Unit: This step is very computationally intensive. It performs add compare select (ACS) operation on branch metric which comes from previous step to calculate path metric which is accumulated distance. The branch having biggest accumulated distance gets discarded.

3) Trace Back Unit: In trace back unit accumulated column error metrics are traced back beginning from the last smallest metric value. The next step in trace back unit is finding previous two possible states and the state with smallest entry is picked. They are stored in survivor state table. These steps are continued until metric table's first column is reached. The survivor table state transitions are used to recreate original message in the last step of the viterbi decoding process.

The decoder output table size is $2^{K+1}.2^m.n$ bits. Where as the size of metric table is $2^{k-1}.b.(5k+1)$ bits. Here b represent number of bits of every entry in metric table. Implementing viterbi decoder in a memory efficient way is a challenging task. For every symbol output table's each entry is accessed once. During decoding process the output table is accessed 2^k times and every entry of metric table is accessed two times for each symbol. The calculation of one entry of the metric table requires two distance calculations and one ACS operation. To calculate one metric table column for each received symbol 2^{k-1} ACS operations and 2^k distance calculations are needed. This process is done for every symbol.

III. Viterbi Accelerator Unit

The aim of this paper is to design and integrate a viterbi accelerator unit with Microblaze soft-core processor datapath. To enhance the processor performance in terms of execution time and energy efficiency. The integration of accelerator will have an impact on the performance of processor. So the accelerator unit should be area, timing and power efficient.

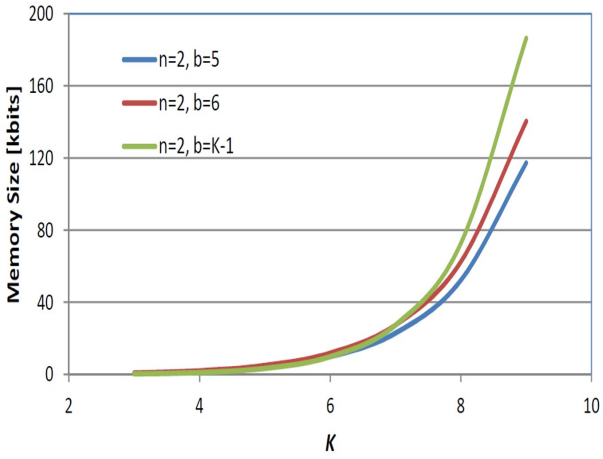


Figure 3: Metric Table memory size variation.

A. Initial Viterbi Decoder

The initial viterbi decoder consist of one ACS unit and one hamming distance calculation unit. For every ACS operation hamming distance unit is used twice sequentially. The control unit is implemented as a state machine. The initial viterbi decoder is shown in Fig. 4. Fig. 3 shows the impact of increasing constraint length K. Here b represents each metric table entry bits and n represent number of output bits. This initial implementation of viterbi decoder with different constraint lengths was synthesized on 7vx485tffg1157-3 Virtex-7 FPGA device which is based on a 28nm technology. Fig. 7 shows the area of implementation for three different constraint lengths. As can be seen the area of decoder increases exponentially with increasing constraint length K. It is observed that major portion of decoder area is consumed by metric table.

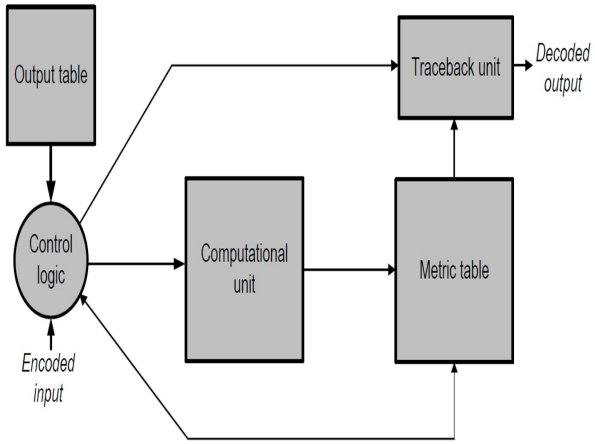


Figure 4: Initial Viterbi decoder architecture.

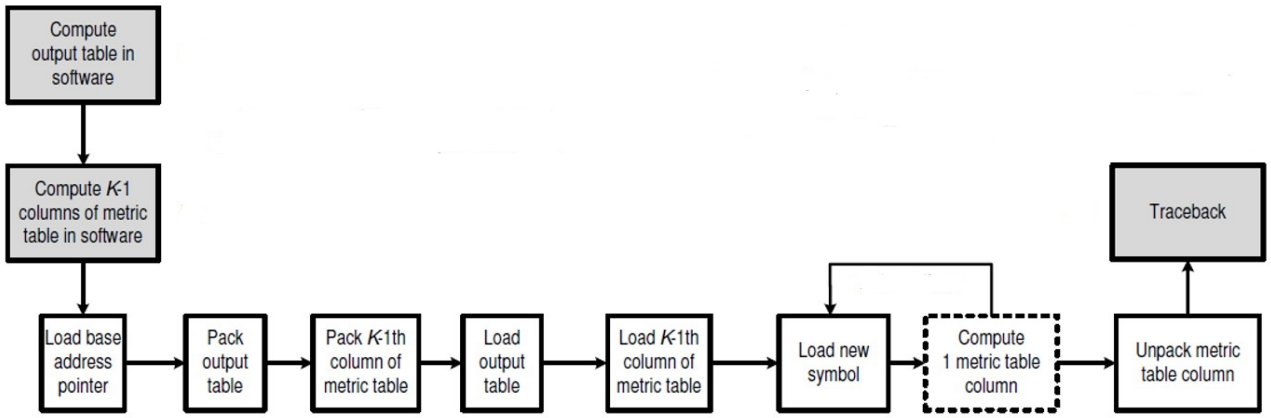


Figure 5: Flow chart for Viterbi Decoder in Full mode.

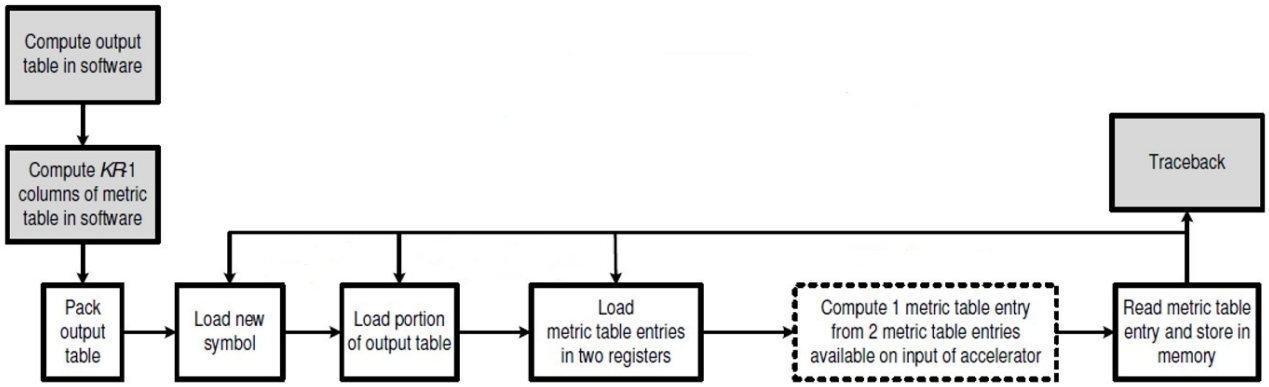


Figure 6: Flow chart for Viterbi Decoder in Sub-State mode.

B. Mixed HW/SW Viterbi Accelerator

The mixed hardware/software approach helps to achieve a good balance between flexibility and performance. We intend to implement the portion of viterbi code in part of accelerator which is computational and memory intensive. The remaining portion of the code which is not frequently executed will be handled by the processor. The decision to define a suitable boundary between hardware and software in designing accelerator-centric heterogeneous systems is a challenging task. Based on the analysis done in the previous section we have made the following conclusions:

- The branch metric and path metric are computational intensive calculations and repeating steps.
- Output table is initialized once.
- After the computation of complete metric table, Trace-back is needed once.
- Output table and previous column of metric table is needed for the computation of new column.

Based on these observations we intend to perform branch metric and path metric calculations in hardware part of mixed hardware/software viterbi accelerator. The output table and traceback computations are done in software and executed on the Microblaze soft-core processor. As the metric table is very computation-intensive and its parallelism can be exploited in hardware implementation. The last metric table column is important for the computation of next column that is why they are stored in the local memory of hardware accelerator. The full metric table is kept in main memory of processor which is needed for trace-back operation. Fig. 8 shows the mixed Hardware/Software viterbi decoder having four computational blocks. Every computational block has one ACS unit and two Euclidian distance computation units for increasing the throughput. This viterbi accelerator is capable to support any

Table I: Synthesis Results of Viterbi Accelerator

Power	241mW
Max Freq	179.808MHz
Latency	5.562ns
Slice Registers	1087
Slice LUTs	3079
Occupied Slices	1038

constraint lenght K. The full metric table computations are performed in hardware when applications constraint length is less than or equal to viterbi accelerator constraint length. Fig. 5 shows flow chart for viterbi decoding in Full mode. Whereas in situations in which the constraint length of applications is greater than viterbi hardware accelerator constraint length and accelerator memory is not enough then Sub-State mode is used. In this mode metric table value is received from the MicroBlaze processor register file. Fig. 6 shows the flow chart for the steps performed in Sub-State mode by the mixed Hardware/Software viterbi decoder. The gray boxes in Fig. 5 and Fig. 6 represent part of the code that is executed in software, whereas the transparent boxes show the steps done in hardware accelerator.

IV. Integration of Viterbi Accelerator Unit with MicroBlaze Processor

We have implemented the viterbi accelerator unit in VHDL hardware description language and verified it using Xilinx ISE design suit [12]. We used Xilinx Spartan-6 FPGA SP605 Evaluation Kit [14] and Xilinx Embedded Development Kit(EDK) [12] for the implementation. Xilinx Microblaze soft-core processor [13] was used to run the software implementation of viterbi decoder. The Hardware/Software co-design is a well established technique which improves the performance of the system [16-20]. There are two ways to integrate a hardware accelerator core into a MicroBlaze-based embedded soft processor system. One way is to connect the accelerator through the Processor Local Bus (PLB). The second way is to connect it using MicroBlaze dedicated Fast Simplex Link (FSL) bus system [15]. First PLB was tried but it was taking a lot of cycles. Because it is a traditional memory mapped transaction bus. Then it was decided to integrate our viterbi accelerator unit using a dedicated FIFO style FSL Bus with the MicroBlaze processor system, shown in Fig. 9.

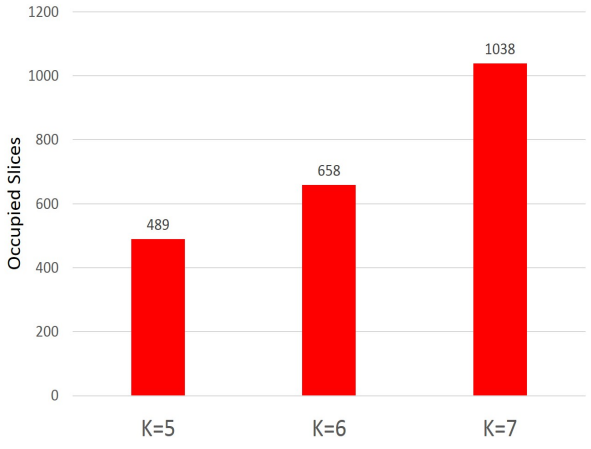


Figure 7: Total area for different constraint lengths.

The software only C code for viterbi decoder was implemented and verified. Later this C code was executed on the MicroBlaze processor using Xilinx Software Development Kit (SDK) [12]. The cycle count for the complete software implementation of viterbi was measured using the XPS hardware timer block, shown in Table II. Fig. 10 and 11 shows the cycle

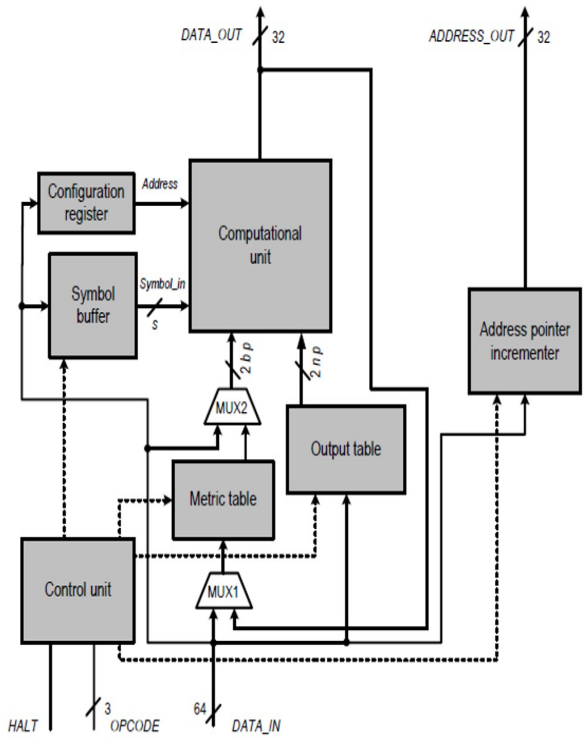


Figure 8: Mixed HW/SW Viterbi decoder architecture.

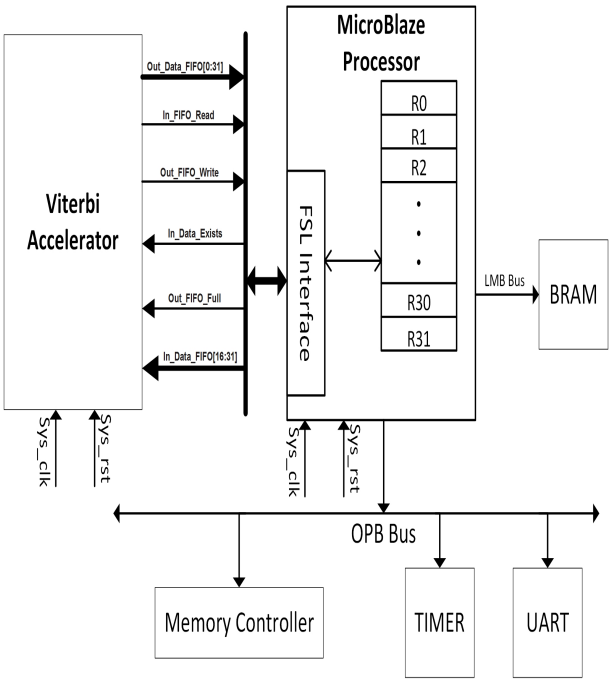


Figure 9: Viterbi Accelerator Unit with MicroBlaze Processor System

count and energy dissipation of two Viterbi implementations, respectively.

The viterbi accelerator unit was attached with the Mi-

Table II: Cycle Count and Energy Dissipation at Clock Period 20ns

Architecture	#Cycles	Power (mW)	Energy* (μJ)
Software Only	8312	178	29.590
Accelerated	2518	185	9.3166

*: Energy = #cycles $\times$ clock period $\times$ power.

croblaze processor system via FSL bus using Xilinx Platform Studio (XPS) [12]. The software part of viterbi accelerator unit was implemented in C programming with Xilinx SDK. The predefined C functions of SDK were used to communicate with hardware part of viterbi accelerator unit via FSL bus. Our evaluation shows that an accelerated MicroBlaze processor datapath is three times more cycle and energy efficient than a datapath lacking viterbi accelerator. This acceleration is achieved at some area overhead.

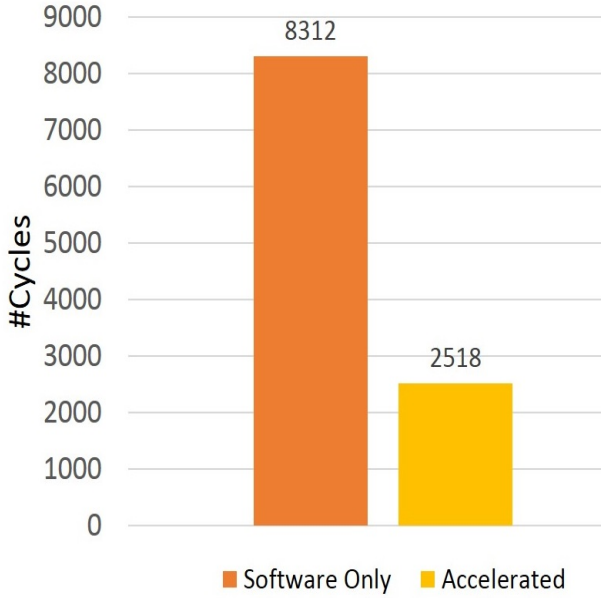


Figure 10: Cycle count of two Viterbi implementations.

V. Conclusion

In this paper, we have designed a mixed hardware/software viterbi accelerator unit using VHDL. We have integrated the viterbi accelerator unit with the Microblaze soft-core processor system using FSL Bus to enhance the processor performance in terms execution time and energy efficiency. We used Xilinx Spartan-6 FPGA Evaluation Kit and Xilinx Embedded Development Kit (EDK) for the implementation. We have shown that a viterbi accelerated Microblaze embedded processor datapath is three times more cycle and energy efficient that a datapath lacking a viterbi accelerator. This acceleration is achieved at the cost of some area overhead.

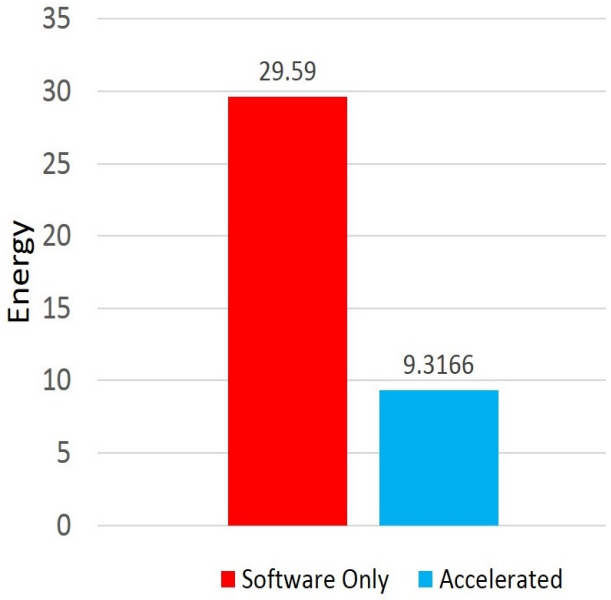


Figure 11: Energy dissipation of two Viterbi implementations.

Acknowledgment

This work is partially supported by the Chinese Academy of Sciences and The World Academy of Sciences CAS-TWAS President's Fellowship 2013-2017.

References

[1] M. F. Brejza, L. Li, R. G. Maunder, B. Al-Hashimi, C. Berrou, L. Hanzo, "20 years of turbo coding and energy-aware design guidelines for energy-constrained wireless applications", IEEE Commun. Surveys Tuts., vol. 18, no. 1, pp. 8-28, 1st Quart. 2016.

[2] Mehran Mozaffari Kermani, Vineeta Singh, Reza Azarderakhsh, "Reliable Low-Latency Viterbi Algorithm Architectures Benchmarked on ASIC and FPGA," IEEE Transactions on Circuits and Systems I: Regular Papers, vol. 64, no. 1, pp. 208-216, 2017.

[3] Linjia Chang, Avhishek Chatterjee, Lav R. Varshney, "Performance of LDPC Decoders With Missing Connections," IEEE Transactions on Communications, vol. 65, no. 2, pp. 511-524, 2017.

[4] Salvatore Pontarelli, Pedro Reviriego, Marco Ottavi, Juan Antonio Maestro, "Low Delay Single Symbol Error Correction Codes Based on Reed Solomon Codes," IEEE Transactions on Computers, vol. 64, no. 5, pp. 1497-1501, 2015.

[5] G. Krishnaiah, N. Engin, and S. Sawitzki, "Scalable Reconfigurable Channel Decoder Architecture for Future Wireless Handsets," in IEEE Design, Automation Test in Europe Conference, Apr. 2007, pp. 1-6.

[6] R. Johannesson and K. S. Zigangirov, "Fundamentals of Convolutional Coding". Wiley-IEEE Press, 1999.

[7] Atmel, "Secure microcontroller for smart cards." [Online]. Available: http://www.atmel.com

[8] Freescale, "MAPLE hardware accelerator and SC3850 DSP core." [Online]. Available: http://www.freescale.com

[9] Microchip, "PIC32mx775f512l datasheet." [Online]. Available: http://www.microchip.com

[10] O. O. Khalifa, T. Al-Maznaee, M. Munjid, and A.-H. A. Hashim, "Convolution Coder Software Implementation Using Viterbi Decoding Algorithm," J. Computer Science, vol. 4, no. 10, pp. 847-856, 2008.

[11] G. D. Forney, Jr., "The Viterbi Algorithm," Proceedings of the IEEE, vol. 61, no. 3, pp. 268278, Mar. 1973.

[12] Xilinx Inc. FPGA Design Tools. Silicon Devices. [Online]. Available: http://www.xilinx.com

[13] Xilinx MicroBlaze [Online] www.xilinx.com/tools/microblaze.htm

[14] Xilinx Spartan-6 FPGA SP605 Evaluation Kit. [Online] Available: www.xilinx.com/products/boards-and-kits/ek-s6-sp605-g.html

[15] Xilinx Fast Simplex Link (FSL). [Online] Available: http://www.xilinx.com/products/intellectual-property/fsl.html

[16] Abdul Rehman Buzdar, Liguo Sun, Azhar Latif and Abdullah Buzdar, "Distance and Speed Measurements using FPGA and ASIC on a high data rate system" International Journal of Advanced Computer Science and Applications(IJACSA), 6(10), 2015, pp.273-282.

[17] Abdul Rehman Buzdar, Liguo Sun, Azhar Latif and Abdullah Buzdar, "Instruction Decompressor Design for a VLIW Processor", Informacije MIDEM-Journal of Microelectronics, Electronic Components and Materials Vol. 45, No. 4 (2015), pp.225-236.

[18] Abdul Rehman Buzdar, Azhar Latif, Liguo Sun and Abdullah Buzdar, "FPGA Prototype Implementation of Digital Hearing Aid from Software to Complete Hardware Design" International Journal of Advanced Computer Science and Applications(IJACSA), 7(1), 2016, pp.649-658.

[19] Abdul Rehman Buzdar, Liguo Sun, Shoab Ahmed Khan, Abdullah Buzdar, "Area and Energy efficient CORDIC Accelerator for Embedded Processor Datapaths" Informacije MIDEM-Journal of Microelectronics, Electronic Components and Materials Vol. 46, No. 4(2016), pp.197-208

[20] Abdul Rehman Buzdar, Liguo Sun, Rao Kashif, Muhammad Waqar Azhar, Muhammad Imran Khan, "Cyclic Redundancy Checking (CRC) Accelerator for Embedded Processor Datapaths" International Journal of Advanced Computer Science and Applications(IJACSA), 8(2), 2017, pp.321-325.

[21] Muhammad Waqar Azhar, Magnus Sjlander, Hasan Ali, Akshay Vijayashekar, Tung Thanh Hoang, K. K. Ansari, and Per Larsson-Edefors, "Viterbi Accelerator for Embedded Processor Datapaths," in Proc. of IEEE Int. Conf. on Applicationspecific Systems, Architectures and Processors, 2012.

[22] J. Heller and I. Jacobs, Viterbi Decoding for Satellite and Space Communication, IEEE Trans. Communication Technology, vol. 19, no. 5, pp. 835848, Oct. 1971.

[23] T. Gemmeke, M. Gansen, and T. G. Noll, "Implementation of Scalable Power and Area Efficient High-Throughput Viterbi Decoders," IEEE J. Solid-State Circuits, vol. 37, no. 7, pp. 941-948, Jul. 2002.

[24] M. Kawokgy and C. A. T. Salama, "A Low-Power CSCD Asynchronous Viterbi Decoder for Wireless Applications," in Proc. Int. Symp. Low Power Electronics and Design, 2007, pp. 363-366.

[25] M. Kamuf, V. wall, and J. B. Anderson, "Optimization and Implementation of a Viterbi Decoder Under Flexibility Constraints," IEEE Trans. Circuits and Systems I: Regular Papers, vol. 55, no. 8, pp. 2411-2422, Sep. 2008.

[26] M. A. Anders, S. K. Mathew, S. K. Hsu, R. K. Krishnamurthy, and S. Borkar, "A 1.9 Gb/s 358 mW 16-256 State Reconfigurable Viterbi Accelerator in 90 nm CMOS," IEEE J. Solid-State Circuits, vol. 43, no. 1, pp. 214-222, Jan. 2008.

[27] C.-C. Lin, Y.-H. Shih, H.-C. Chang, and C.-Y. Lee, "A Low Power Turbo/Viterbi Decoder for 3GPP2 Applications," IEEE Trans. Very Large Scale Integration (VLSI) Systems, vol. 14, no. 4, pp. 426-430, Apr. 2006.

[28] M. A. Bickerstaff et al., "A Unified Turbo/Viterbi Channel Decoder for 3GPP Mobile Wireless in 0.18-m CMOS," IEEE J. Solid-State Circuits, vol. 37, no. 11, pp. 1555-1564, Nov. 2002.

[29] J. R. Cavallaro and M. Vaya, "Viturbo: A Reconfigurable Architecture for Viterbi and Turbo Decoding," in Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing, vol. 2, Apr. 2003, pp. 497-500.

[30] D. E. Hocevar and A. Gatherer, "Achieving Flexibility in a Viterbi Decoder DSP Coprocessor," in Proc. 52nd IEEE Vehicular Technology Conf., vol. 5, 2000, pp. 2257-2264, vol.5.

[31] A. Niktash, H. Parizi, and N. Bagherzadeh, "A Reconfigurable Processor for Forward Error Correction," in Proc. Int. Conf. on Architecture of Computing Systems, 2007, pp. 1-13.

Comparison of Localization Free Routing Protocols in Underwater Wireless Sensor Networks

Muhammad Khalid
Institute of Management Sciences
Peshawar Pakistan, 25000

Zahid Ullah
Institute of Management Sciences
Peshawar Pakistan, 25000

Naveed Ahmad
Department of Computer Science
University of Peshawar, Pakistan 25000

Awais Adnan
Institute of Management Sciences
Peshawar Pakistan, 25000

Waqar Khalid
Institute of Management Sciences
Peshawar Pakistan, 25000

Ahsan Ashfaq
Institute of Management Sciences
Peshawar Pakistan, 25000

***Abstract*—Underwater Wireless Sensor Network (UWSN) is newly developed branch of Wireless Sensor network (WSN). UWSN is used for exploration of underwater resources, oceanographic data collection, flood or disaster prevention, tactical surveillance system and unmanned underwater vehicles. UWSN uses sensors of small size with a limited energy, memory and allows limited range for communication. Due to multiple differences from terrestrial sensor network, radio waves cannot be used over here. Acoustic channel are used for communication in deep water, which has many limitations like low bandwidth, high end to end delay and path loss. With the above limitations while using acoustic waves, it is very important to develop energy efficient and reliable protocols. Energy efficient communication in underwater networks has become uttermost need of UWSN technology. The main aim nowadays is to operate sensor with smaller battery for a longer time. This paper will analyse various routing protocols in the area of UWSN through simulation. This paper will analyse Depth Based Routing (DBR), Energy Efficient Depth Based Routing (EEDBR) and Hop by Hop Dynamic Addressing Based (H2-DAB) protocol through simulation. This comparison is carried out on the basis of total consumed energy, end to end delay, path loss and data delivery ratio.**

***Keywords*—Underwater Networks; Sensor; Wireless Communication; Survey; Localization Based; Routing; Protocols**

I. Introduction

UWSN is a newly emerging wireless technology which is providing the most promising mechanism used for discovering acoustic environment very efficiently for many scenarios like military [1], emergency and commercial purposes [2]. Autonomous Underwater and unmanned Vehicles which are equipped with sensors that are specially designed for underwater communication, which are mostly used in those areas where exploration for natural resources which lies underwater is needed. These unmanned vehicle gather those data and send it back to off shore sinks which is forwarded to other stations for further processing[3]. Radio waves cannot be used in underwater communication. Therefore, communication is made through acoustic channels. Once data packet reaches sink then it is forwarded through radio waves to other sinks and stations [4]. Underwater wireless sensor environment is much different from that of terrestrial network where no such ambiguities are found which we face in underwater communication while using radio communication [5]. Normally the problems we faces during communication in underwater communication are dense salty water, electromagnetic as well as optical signal does not work here [6]. Due to high attenuation and absorption effect, signals cannot travel long distances. Hence to overcome these problems acoustic communication, is used[7]. It can overcome these problem and provides a better transfer rate in underwater environment[8]. Using acoustic communication propagation speed lowered down from speed of light to that of sound speed which is 1500m/sec. Due to lower speed there is usually long propagation delay and higher end to end time [9]. In acoustic communication bandwidth is very limited which is less than 100KHz. In underwater scenarios, sensor nodes are usually considered static but it is also considered that they may move from 1 to 3 meter/second due to flow of water[10]. Sensor nodes used in underwater network are battery operated and it is almost impossible to replace its batteries. In underwater applications a multi-hop or multipath network is required and data is forwarded by passing all nodes towards sink. Once data is received at any of the sink then data is forwarded to concerned node through radio transmission[11]. While using those routing protocols which requires higher bandwidth, usually has higher delay at the nodes end[12]. As we know that acoustic communication does not support higher bandwidth so using routing protocols that are used in terrestrial network will not perform good due to it higher delay and high energy consumption. Using underwater network, topology does not remains the same as node moves due to flow of water[13]. In localization based protocol, geographical network information is necessary so it possess more control messages than localization free protocol, in which no prior network information is necessary. Usually ocean are vast and covers around one hundred and forty millions square miles, which is more than 70 percent of Earth total surface, Not only it has been considered to be major source of the nourishment, but with span of time its taking a good role in transportation stuffs, defence as well as adventurous purposes and natural resources presence[14]. All its importance towards humanity, it is very strange that we know a very little of about Earth water bodies. Less than ten percent of whole ocean volume is investigated, while a large amount of area has still not been explored[6]. The increase in roles of the oceans in the lives of humans, importance of these largely unexplored area has got a lot of importance. If we see, on one hand the traditional approaches for underwater monitoring have got several disadvantages while on the other side human presence is not considered

to be feasible for underwater environment[15]. We face very unique challenges as compared to other networks. Protocols suites that are used in other networks cannot be directly applied to underwater networks. Till date, many protocols has been proposed for underwater sensor networks[16]. These are mainly divided into two types which are localization based and localization free protocols. Localization free protocols does not require any prior geographic or network information. Most of these protocols are used in underwater networks.

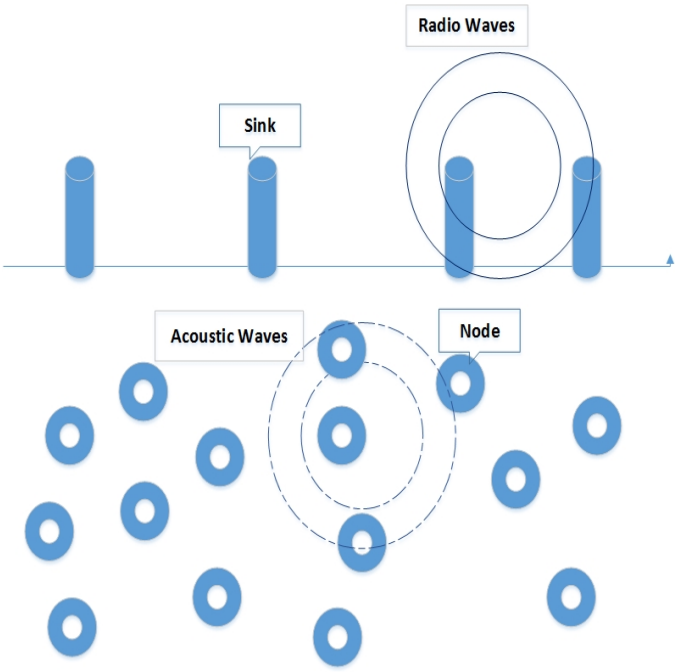


Fig. 1: UWSN Architecture

The rest of the paper is organized as follow. In section 2 related work has been discussed. In section 3 terminologies regarding routing are defined. Section 4 has discussed location free routing protocol. Evaluation and result has been discussed in section 5 and finally conclusion is drawn in section 6.

II. Constraints in Underwater Wireless Sensor Network

UWSN carries multiple differences in comparison with terrestrial area network[17]. In which nodes are stable or move in a specified direction while in underwater networks they usually displaces their positions with the flow of water[18]. Acoustic communication is used for underwater transmission which minimizes the bandwidth for data transferring[19].

- Limited Bandwidth Acoustic channels offer very limited amount of bandwidth, as radio transmission cannot be used for underwater communication [4]. Acoustic communication requires more energy to send a small amount of data due to its lower bandwidth[20].
- Propagation Delay Due to use of acoustic communication, propagation speed becomes five times slower than that of radio frequency i.e. 1500m/sec [5].which obviously results in high propagation delays in the network.
- Limited Energy Nodes that are used in underwater communication are larger in size [4], hence they require larger amount of energy for communication. Furthermore, acoustic channels also required more energy for communication than terrestrial network[21],[22]. Batteries in UWSN cannot be recharged or replaced therefore use of energy efficient communication is always required to provide network with higher life time[14].
- Limited memory In UWSN nodes are small in size and therefore they have a limited amount of storage and processing capacity [6],[23].
- Variable Topology UWSN does not have a specific or static topology as flow of water make it difficult for node to remain static in one place, therefore node moves randomly[24].

III. Related Work

In this section, relevant routing techniques in litearature are discussed. In Energy Efficient Dynamic Address Based routing (EE-DAB) [13] every node is assigned node id, s-hop id and c-hop id. Node id show the physical address of node, s-hop id consist of two digits which show how many hops away one or two sinks are. Left hop is considered as highest priority and is selected as primary route. The C-hop id also consist of 2 digits which show that how many hop the receiving nodes are away from courier nodes. acoustic communication uses more energy than that of radio communication. As wireless sensor nodes are battery operated and higher energy consumption lead towards a serious problem. Thus energy efficiency has become a major problem in underwater wireless sensor networks. In [24], a delay tolerant protocol is proposed which is called delay-tolerant data dolphin scheme. This proposed scheme is designed for delay tolerant systems and applications. In this protocols all the sensing node stay static and data sensed by static nodes are passed on to data dolphin which acts a courier nodes. So in this methodology high energy consumed hop by hop communication is avoided. Data dolphins which acts a courier nodes are provided with continuous energy. In the architecture all the static nodes are deployed in the sea bed. These static sensor goes into sleep mode if there is no data to sense and it periodically wakes up when it sense some data. After sensing some kind of desired data it simply forward this data to courier nodes which are also called data dolphins. These data dolphins take this data and deliver it to base station or sink. The number of dolphin nodes depend upon the kind of network and its application and the number of nodes deployed in the network. In [22], a virtual sink architecture is proposed where sinks are connected with each other through radio communication. In this scheme, each and every sink broadcast a hello packet which is also known as hop count update packet. After receiving hello packet by nodes, a hop count value is assigned to every sensor. These hop counts are used for selection of forwarding nodes while sending data packet from one node to another. However the proposed scheme has a few limitations which includes redundant transmission i.e. transmission of a same packet multiple times. Routing protocols which needs prior network information before send any data over the network are called localization based routing protocols. These protocol usually

need geographical information of all node in the network as well as information about sink location. These protocols are considered to be less energy efficient most of energy is wasted in collecting their geographical information. These records are updated dynamically after fixed interval of time as nodes position may changes due to water flow. Routing protocols basically need the assumption of sensor nodes in underwater sensor networks [2]. In localization based routing protocols a node need the information of all the network nodes as well as of sink like in this scenario prior network information is needed for a node [8], [24], [21]. In [25], Focused Beam routing protocol requires geographical information of itself and as of destination. It uses Ready To Send and Clear To Send mechanism to forward data. Sender protocol transmit the RTS and receiver of the packet send back CTS. In Vector Based Forwarding [23], a source node develop a vector based routing pipe starting from sender node towards sink. Various times it is hard to find an available node in the routing pipe for data forwarding. SBR-DLP [15], also known as sector base routing, with destination location prediction is a localization based routing algorithm where node is not needed to have information of its neighbor nodes. It only need to carry its own information and pre-planned movement of sink although it decreases the flexibility of the network and it will only move around in a scheduled manner. Those routing protocols which does not require any geographical information of the network are called localization free routing protocols. These protocols perform their operation without having location information of other nodes. In these kind of routing protocols, a sensor node does not require any prior network information of other network nodes [22], [14]. Most of the localization protocols work on flooding phenomenon and are considered to have fast packet delivery ratio and low end to end delay [14], [20]. In [21], Depth base routing does not need any pre network information. It just take the depth of sensor nodes into account and forward a packet. It actually compares the depth of sending node with that of receiving node so if depth of sender node is higher than that of receiver node then it will forward the data otherwise it will ignore that node. Similarly in [11], Energy Efficient DBR, it take into account the depth information as well as residual energy of the node at the time of sending data.

IV. Node Architecture

A general architecture of underwater wireless sensor node is composed of five main elements. Which are energy management unit, data sensing unit, depth measuring unit, communication unit and central processing unit [21]. As show in 2.

Processing unit is responsible for all kind of data processing which energy management unit has the responsibility to manage the remaining energy of the node and consumption of energy in run time [3]. Data sensing unit is used to sense data. It always remains active even when node is in sleep mode [26]. Communication unit is responsible for all kind of data communication whereas depth measuring unit is used for measuring depth of nodes when it is deployed in sea [9].

V. Localization Free Routing Protocols

In UWSN, many routing protocols had been proposed [27],[10]. Each protocol has its good and bad aspects. These

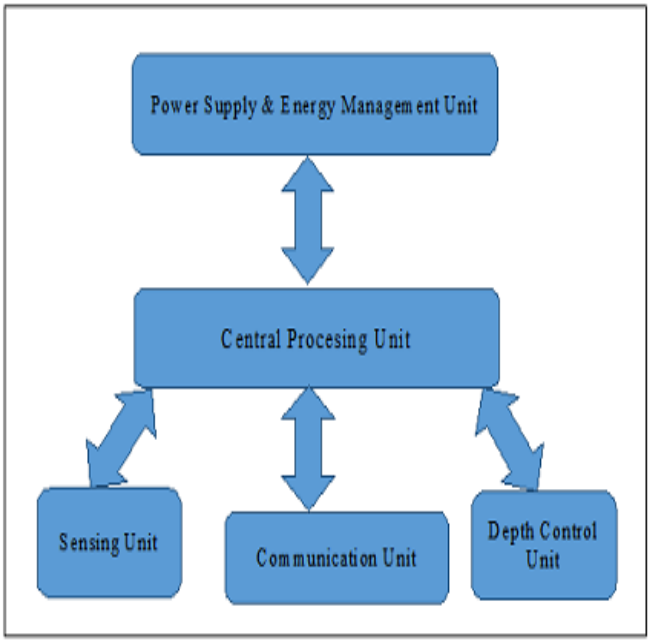


Fig. 2: Sensor Node Architecture

protocols are designed specially for underwater communication as keeping in view the limitations of the network, their low battery and low bandwidth but still there are some deficiencies that need to be addressed. Mainly routing protocols in UWSN is divided into two parts, i.e. localization based routing protocols and localization free routing protocols[8]. Localization based routing protocol comparatively requires more energy as it need prior network information[25]. Every node in the network must have detailed information of all other nodes in the network. During network initialization phase every node request other node about their current status as well as sink also broadcast ping message to know about the energy level and location of nodes. This network information is updated simultaneously after a fixed interval of time.In localization free routing protocol, it does not need any information of other nodes. This schemes consumes less energy than localization based routing protocols.

A. Depth Based Routing

Unlike localization based routing, Depth Based Routing Protocol [10] does not need any prior network information. DBR needs depth information of each node. When a node with the highest depth sense some movement, it starts sending data to higher nodes, such that it compares its depth with neighbor nodes. If send packets to only those nodes whose depth is lower than sender node. The same process continuous until packet is received by sink. This protocol is mainly concerned about depth of node. Sink are provided with continuous power. Figure 3 defines next node selection in depth based routing protocol. Where three nodes n1, n2 and n3 are in communication range of sender S. In first step depth of receiver nodes is checked. N1 and N2 are found eligible for data forwarding as their depth is less than sender node S. Now the sender S will send the data packet to two eligible nodes N1 and N2.

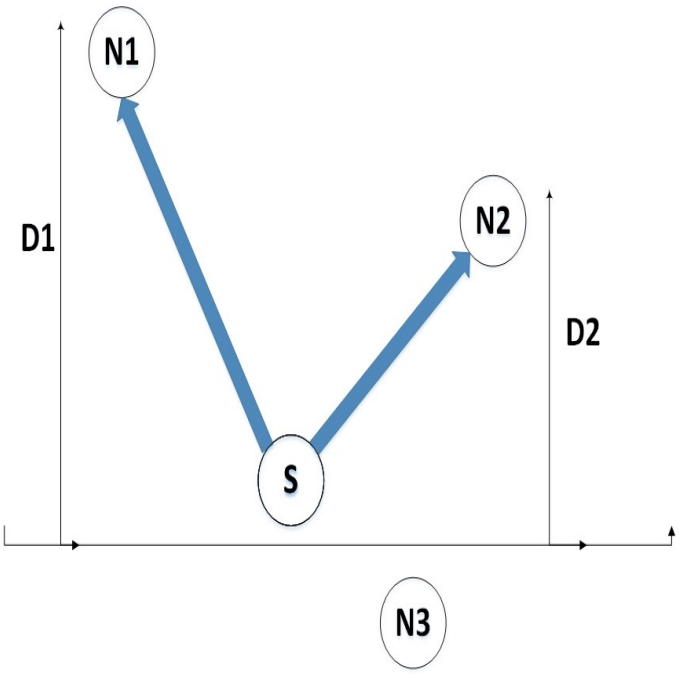


Fig. 3: Depth Based Routing Protocol

The forwarding of data in broadcast manner always result in waste of energy whether a node is sending to receiving data and always leads towards low network life. DBR does not take into account any other parameter then depth, which leads towards a few drawbacks. Network life of network where DBR is used, will be less as it will always sends the data to the same higher node. Which will decrease the number of alive nodes. There is no proper mechanism for path selection in DBR as neither proper strategy is used for efficient path nor shortest path is selected.

B. Hop by Hop Dynamic Addressing Based Routing

In H2-DAB [20], dynamic addresses are assigned to nodes and destination ID is set to 0 for all nodes. No pre-network information is required in this protocol. In first step of network setup, a hop id is assigned to each node. Every node in the network will have two type of addresses, node id and hop id. Node id is physical address of node while node id changes with change in location In H2-DAB the assignment of Hop IDs which are assigned from top to bottom. Node having lower depth are assigned lower hop id, like node which is nearest will have hop id of 1. Similarly nodes having higher depth are assigned higher hop IDs. H2- DAB supports multi sink architecture, where multiple sink are installed on shore. Those sinks are connected with each other through radio communication. Data packet received at any sink is considered received. However this approach might create problems where a node cannot find in range, any node which has lower hop id from sender node. In case of failure at finding suitable node in first attempt, sender will retransmit data packet and then wait again for specified amount of time. If results were still the same then sender node will forward data to a node having nearly or equal hop id as sender node. This process results in energy wastage.

C. Energy Efficient Depth Based Routing Protocol

In EE-DBR [14], protocol when a node forwards its data, it takes into account the depth of the receiver node and its residual energy. When a node forwards data it first compares the depth of the receiver node with itself, if the depth of receiver node is smaller than sender then it checks the residual energy of receiver node. Node with higher residual energy and less depth among the neighbors is selected as next hop for communication. Every node has information on depth and residual energy about their neighbors, so the node with most suitable parameter is selected for communication. EE-DBR has not defined any mechanism for multi-path communication. A node may forward data to node which is far away from sender and will results in higher energy consumption. Similarly no parameter has been taken into account to define a shortest and efficient path towards sink.

VI. Evaluation and Results

This paper will analyze the performance of location free routing protocols through various evaluation techniques. In this simulation three protocols i.e. DBR, EEDBR and H2-DAB are compared through simulation on the basis of network delivery ratio, path loss, network life time, number of alive nodes left and total energy consumption. Below is the parameter metrics that is taken into consideration while performing simulation.

A. Simulation Parameters

This simulation is carried out on area of 100m x 100m with 225 node and simulation time is 9000 rounds. We have simulated DBR, EEBR and H2-DAB on the basis of total end to end delay, path loss, path loss of network packet delivery ratio and total consumed energy. Simulation results are discussed below.

TABLE I: Simulation Parameters

Parameters	Value
Network Size	100m X 100m
Total Nodes	225
Initial Energy	25J
Packet Size	1024 bits
Number of Sink	4
Transmission Range	100 meter
Rounds	9000

B. Simulation Results

The above explained terminologies are taken into consideration during simulation while comparing DBR, EEDBR and H2-DAB . Results gained are discussed below.

1) End To End Delay: In figure 4, three protocols are compared with respect to end to end delay. It shows that protocol which need prior network information like H2-DAB have more end to end delay while protocols like DBR and EE-DBR have low end to end delay. DBR and EEDBR uses multi-hop mechanism, when a sender node forward a packet to all available nodes in its range. Data is forwarded on multiple path simultaneously and there is no data loss even if one path fails. H2-DAB has higher end to end delay due to unavailability of appropriate node.

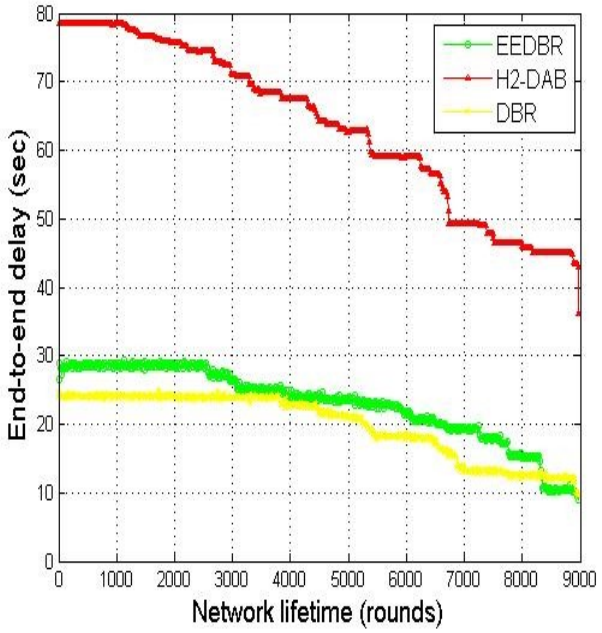


Fig. 4: End to End Delay

2) Path Loss: Figure 5, has compared path loss of network in DBR, EEDBR and H2-DAB. This figure clearly shows that path loss in H2-DAB is less while greater in DBR and EEDBR comparatively.

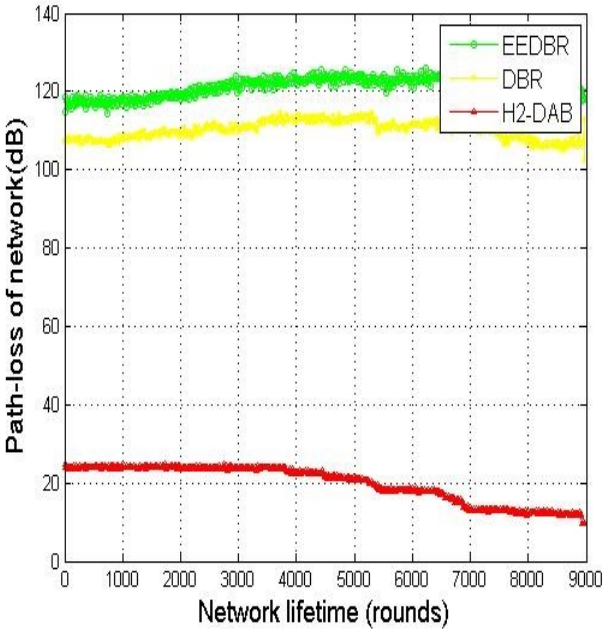


Fig. 5: Path Loss of Network

3) Packet Delivery Ratio: Figure 6 has compared packet delivery ratio in the above defined network using DBR, EEDBR and H2-DAB. The graph show delivery ratio in H2-DAB is higher as compare to other two while in DBR and EE-DBR it is almost the same.

Figure 7 shows that total amount of consumed energy where DBR has consumed more energy than other two protocol because of its flooding nature. H2-DAB has consumed less energy of all and remained consitent throughout the process. DBR, EE-DBR and H2-DAB are compared through simulation with respect to total energy consumption, end to end delay, path loss, packet delivery ratio and number of alive nodes. The graphs show that DBR showed good results in end to end delay and packet delivery ratio while number of alive nodes are less using DBR as compared with EE-DBR and H2-DAB when compared at a certain stage. Packet delivery ratio in H2-DAB is low when compared to other two. Using EE-DBR less energy is consumed when compared to other two protocols through simulation.

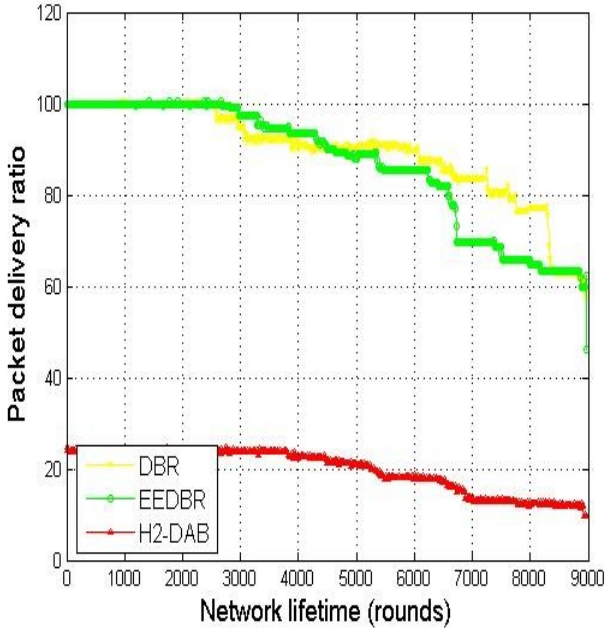


Fig. 6: Packet Delivery Ratio

4) Total Energy Consumption: Figure 7 shows that total amount of consumed energy where DBR has consumed more energy than other two protocol because of its flooding nature. H2-DAB has consumed less energy of all and remained consitent throughout the process. DBR, EE-DBR and H2-DAB are compared through simulation with respect to total energy consumption, end to end delay, path loss and packet delivery ratio. The graphs show that DBR showed good results in end to end delay and packet delivery ratio. Packet delivery ratio in H2-DAB is low when compared to other two routing protocols. Using EE-DBR, less energy is consumed when compared to other two protocols through simulation.

VII. Conclusion

In this paper we have compared the state of the art routing protocols in UWSN. Routing in UWSN is challenging and requires energy efficient techniques. While designing any

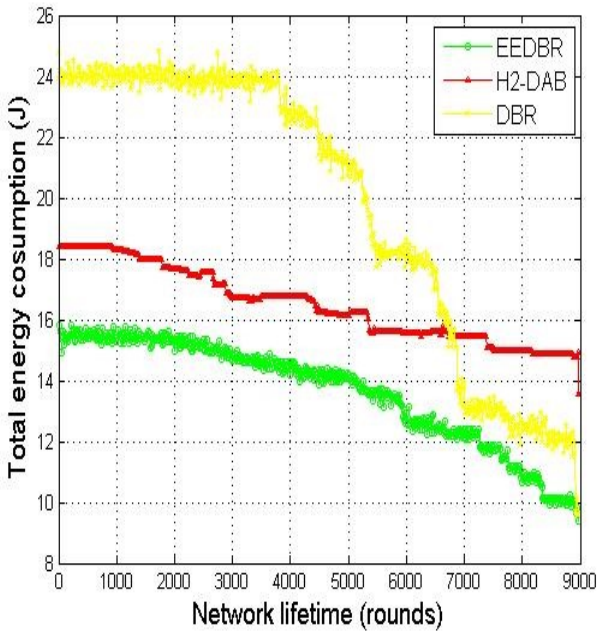


Fig. 7: Total Energy Consumption

routing protocol one should keep in view the requirement for specific application. The performance of routing protocols mostly rely on reliability, availability, energy efficiency, multi and efficient path selection, number of alive nodes and end to end delay. These all challenges attracts researchers to work in this area. Which has made UWSNs very fast growing area. Results show that H2-DAB has higher end to end delay but on the other hand it has lower network path loss than DBR and EEDBR. Packet delivery ration and network path loss of DBR and EEDBR are almost same with a little difference. There is no silver bullet for routing in UWSN and thus a lot of work could be done in the future.

REFERENCES

[1] C. Giantsis and A. A. Economides, "Comparison of routing protocols for underwater sensor networks: a survey," *International Journal of Communication Networks and Distributed Systems*, vol. 7, no. 3-4, pp. 192–228, 2011.

[2] N. Ilyas, M. Akbar, R. Ullah, M. Khalid, A. Arif, A. Hafeez, U. Qasim, Z. A. Khan, and N. Javaid, "Sedg: Scalable and efficient data gathering routing protocol for underwater wsns," *Procedia Computer Science*, vol. 52, pp. 584–591, 2015.

[3] J. L. Tangorra, S. N. Davidson, I. W. Hunter, P. G. Madden, G. V. Lauder, H. Dong, M. Bozkurttas, and R. Mittal, "The development of a biologically inspired propulsor for unmanned underwater vehicles," *IEEE Journal of Oceanic Engineering*, vol. 32, no. 3, pp. 533–550, 2007.

[4] I. F. Akyildiz, D. Pompili, and T. Melodia, "Challenges for efficient communication in underwater acoustic sensor networks," *ACM Sigbed Review*, vol. 1, no. 2, pp. 3–8, 2004.

[5] I. F. Akyildiz, D. Pompili, and Melodia, "Underwater acoustic sensor networks: research challenges," *Ad hoc networks*, vol. 3, no. 3, pp. 257–279, 2005.

[6] J. 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. IEEE, 2006, pp. 228–235.

[7] N. Chirdchoo, W.-S. Soh, and K. C. Chua, "Sector-based routing with destination location prediction for underwater mobile networks," in *Advanced Information Networking and Applications Workshops, 2009. WAINA'09. International Conference on*. IEEE, 2009, pp. 1148–1153.

[8] E. Felemban, F. K. Shaikh, U. M. Qureshi, A. A. Sheikh, and S. B. Qaisar, "Underwater sensor network applications: A comprehensive survey," *International Journal of Distributed Sensor Networks*, vol. 11, no. 11, p. 896832, 2015.

[9] R. Manjula and S. S. Manvi, "Issues in underwater acoustic sensor networks," *International Journal of Computer and Electrical Engineering*, vol. 3, no. 1, p. 101, 2011.

[10] H. Yan, Z. J. Shi, and J.-H. Cui, "Dbr: depth-based routing for underwater sensor networks," in *International conference on research in networking*. Springer, 2008, pp. 72–86.

[11] A. Wahid and D. Kim, "An energy efficient localization-free routing protocol for underwater wireless sensor networks," *International journal of distributed sensor networks*, vol. 8, no. 4, p. 307246, 2012.

[12] M. Ayaz and A. Abdullah, "Hop-by-hop dynamic addressing based (h2-dab) routing protocol for underwater wireless sensor networks," in *Information and Multimedia Technology, 2009. ICIMT'09. International Conference on*. IEEE, 2009, pp. 436–441.

[13] M. Ayaz, A. Abdullah, I. Faye, and Y. Batira, "An efficient dynamic addressing based routing protocol for underwater wireless sensor networks," *Computer Communications*, vol. 35, no. 4, pp. 475–486, 2012.

[14] M. Ayaz and A. Abdullah, "Underwater wireless sensor networks: routing issues and future challenges," in *Proceedings of the 7th International Conference on Advances in Mobile Computing and Multimedia*. ACM, 2009, pp. 370–375.

[15] E. A. Carlson, P.-P. Beaujean, and E. An, "Location-aware routing protocol for underwater acoustic networks," in *OCEANS 2006*. IEEE, 2006, pp. 1–6.

[16] M. Patil and R. C. Biradar, "A survey on routing protocols in wireless sensor networks," in *Networks (ICON), 2012 18th IEEE International Conference on*. IEEE, 2012, pp. 86–91.

[17] P. Xie, J.-H. Cui, and L. Lao, "Vbf: vector-based forwarding protocol for underwater sensor networks," in *International Conference on Research in Networking*. Springer, 2006, pp. 1216–1221.

[18] N. Nicolaou, A. See, P. Xie, J.-H. Cui, and D. Maggiorini, "Improving the robustness of location-based routing for underwater sensor networks," in *OCEANS 2007-Europe*. IEEE, 2007, pp. 1–6.

[19] D. Hwang and D. Kim, "Dfr: Directional flooding-based routing protocol for underwater sensor networks," in *OCEANS 2008*. IEEE, 2008, pp. 1–7.

[20] M. Erol and S. Oktug, "A localization and routing framework for mobile underwater sensor networks," in *INFOCOM Workshops 2008, IEEE*. IEEE, 2008, pp. 1–3.

[21] K. Anupama, A. Sasidharan, and S. Vadlamani, "A location-based clustering algorithm for data gathering in 3d underwater wireless sensor networks," in *Telecommunications, 2008. IST 2008. International Symposium on*. IEEE, 2008, pp. 343–348.

[22] N. Aitsaadi, N. Achir, K. Boussetta, and G. Pujolle, "Differentiated underwater sensor network deployment," in *Oceans 2007-Europe*. IEEE, 2007, pp. 1–6.

[23] J. M. Jornet, M. Stojanovic, and M. Zorzi, "Focused beam routing protocol for underwater acoustic networks," in *Proceedings of the third ACM international workshop on Underwater Networks*. ACM, 2008, pp. 75–82.

[24] K. Ali and H. Hassanein, "Underwater wireless hybrid sensor networks," in *Computers and Communications, 2008. ISCC 2008. IEEE Symposium on*. IEEE, 2008, pp. 1166–1171.

[25] E. Magistretti, J. Kong, U. Lee, M. Gerla, P. Bellavista, and A. Corradi, "A mobile delay-tolerant approach to long-term energy-efficient underwater sensor networking," in *Wireless Communications and Networking Conference, 2007. WCNC 2007. IEEE*. IEEE, 2007, pp. 2866–2871.

[26] R. Manjula and S. S. Manvi, "Issues in underwater acoustic sensor networks," *International Journal of Computer and Electrical Engineering*, vol. 3, no. 1, p. 101, 2011.

[27] E. Felemban, F. K. Shaikh, U. M. Qureshi, A. A. Sheikh, and S. B. Qaisar, "Underwater sensor network applications: A comprehensive

survey," *International Journal of Distributed Sensor Networks*, vol. 11, no. 11, p. 896832, 2015.

Dynamic Gesture Classification for Vietnamese Sign Language Recognition

Duc-Hoang Vo, Huu-Hung Huynh
University of Science and Technology
The University of Danang, Vietnam

Phuoc-Mien Doan
Tra Vinh University
Tra Vinh, Vietnam

Jean Meunier
DIRO, University of Montreal
Montreal, Canada

***Abstract*—This paper presents an approach of feature extraction and classification for recognizing continuous dynamic gestures corresponding to Vietnamese Sign Language (VSL). Input data are captured by the depth sensor of a Microsoft Kinect, which is almost not affected by the light of environment. In detail, each gesture is represented by a volume corresponding to a sequence of depth images. The feature extraction stage is performed by dividing such volume into a 3D grid of same-size blocks in which each one is then converted into a scalar value. This step is followed by the process of classification. The well-known method Support Vector Machine (SVM) is employed in this work, and the Hidden Markov Model (HMM) technique is also applied in order to provide a comparison on recognition accuracy. Besides, a dataset of 3000 samples corresponding to 30 dynamic gestures in VSL was created by 5 volunteers. The experiments on this dataset to validate the approach and that shows the promising results with average accuracy up to 95%.**

***Keywords*—Dynamic gesture; feature extraction; depth information; Vietnamese Sign Language**

I. Introduction

In recent decades, computer vision algorithms have been employed in many systems such as surveillance, human-computer interaction, robotic, smart home, and communication [1]. Among various vision-related problems, hand gesture recognition is the one which is widely being studied, in which a suitable approach can give supports for hard-of-hearing people in communication, as well as help to perform interacting between human and computer without touching. According to gesture types, researchers separated such problems into two sub-problems. Methods working on static gestures usually describes local and/or global features of hand shape and posture, while the hand motion is mostly estimated to represent dynamic gestures. Some researchers proposed approaches for dynamic gestures based on static ones.

This work focuses on the problem of recognizing dynamic hand gestures, which is considered to be more difficult than the similar objective on static ones. The input of this system is a sequence of depth images captured by a depth camera (Microsoft Kinect in our experiments) via an infrared (IR) sensor.

The remaining content of this article is organized as follows: some related studies as well as existing limitations are described in Section II; the details of this approach is then presented step by step in Section III; Section IV shows experiments and obtained results; and the conclusion is finally given in Section V.

II. Related Work

In recent studies, the stage of data acquisition was usually performed with the support of sensors mounted on gloves or vision-based systems. Therefore they could be separated into two categories with different pros and cons.

A. Sensor-based approach

In recent studies, many approaches focusing on the problem of recognizing hand gestures have been proposed. For example, the work [2] introduced a method for classifying 6 hand gestures of Korean signal language. The study [3] also built a game controller and performed hand symbol recognition based on the collection of a 3D acceleration sensor and electromechanical biological sensors. In [4], researchers developed a system of classifying symbols in Greek sign language using the energy obtained from bodies with biosensors (EMG) and the assembled data from an acceleration sensor mounted on the arm. For VSL, the researchers in [5] used gloves which are combined with sensors to identify 23 character gestures in the Vietnamese alphabet. These methods focused on the medical field as well as controlling, they thus still have a limited capability to identify the actual sign language gestures. Besides, it is inconvenient for users to carry the data acquisition devices on the body. Several other studies use gloves to capture the change of shapes and movements of the hand. In [6], [7], a glove which was equipped with sensors on all fingers and palm was used to detect movement and bending of the fingers. Besides, such glove also helps to retrieve the location, speed and direction of the hand under a predefined reference system. Some other studies used colored gloves combined with a computer vision system instead of using sensors [8]–[11], in which the fingers are marked by different colors. The use of gloves can support such features simplify the preprocessing step, but brings inconvenience to users when they have to wear gloves during performing sign language.

B. Vision-based approach

The work [12] published a database which consists of hand images performing 26 different gestures, in which each one includes 86 images captured from different directions in the 3D space. In total, a collection of 107328 sample images was obtained. In [13], the authors built 249 samples of 49 word symbols in American Sign Language (ASL). In order to detect and distinguish hand movements, they used 2 different colored-gloves to perform gestures. The testing stage was done in the laboratory environment with a dark background

to enhance the ability of segmenting and distinguishing the two hands. The study [14] also built another database of ASL consisting of 2576 videos corresponding to 14 gestures. The data were recorded by a RGB camera, but the performers must wear long-sleeved shirts in which the color is similar to the background. In [15] presented a database of 19 gestures with 4121022 colored image samples. Although the achieved identification accuracy was about 94%, this work strongly depended on the preprocessing since a simple skin color filter was employed to perform the segmentation. Another approach which also focused on skin color pixels was proposed by [16], in which static gestures of VSL are classified by a neural network. Although such mentioned solutions provided promising experimental results, the preprocessing for hand segmentation was significantly affected by the brightness of environment as well as the background texture.

In order to overcome mentioned limitations, recent researchers performed data acquisition using a depth camera. The study [17] built a system supporting hand gesture recognition, in which the data was collected by a Kinect sensor. An obvious drawback of this study is that characteristics, which describe the hand posture, might be significantly affected be the finger detection result. Therefore the system in [17] has to be improved much for applying on alphabetic gestures. This disadvantage also occurred in [18] where the hand gesture was represented based on detected fingers. Another gesture recognition method working on depth images was presented in [19]. The researchers proposed approaches for both static and dynamic gestures. In experiment, the reported error rate was 5%. However, their main disadvantage is that the features, which were based on hand shape, were limited on the orientation, the template matching thus could perform the classification with low accuracy.

This paper proposes a technique for extracting gesture features and classifying them using SVM. An experiment on HMM was also performed in order to provide a discussion. Since input of this approach is depth image, the mentioned drawbacks of previous works, which are related to sensor and color image, could be overcome.

III. Proposed Method

A. Microsoft Kinect

Camera Kinect is a product manufactured by Microsoft. There are two versions of Kinect which used different techniques for estimating depth information. The proposed system uses a Kinect version 2 which contains many components inside, including a color image acquisition device with a high-resolution up to 1920×1080 pixels, the depth sensor consisting of an infrared emitter and an infrared receiver which provide depth images at 512×424 pixels image at real-time rate. The depth map, which is calculated based on the infrared signals, plays an important role in extracting and recognizing objects since the changes in brightness almost have no impact on the received depth information of Kinect. Nowadays, many researches who are working on the field of sign language recognition [18], [20]–[23] use Kinect to extract objects of interest based on color information, depth as well as joint coordinates provided in the SDK. In the proposed approach, the depth image is captured frame by frame, and each dynamic gesture is represented by a sequence of such images.

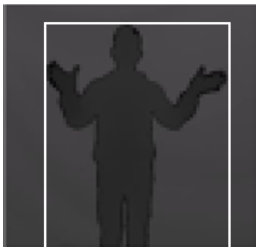


Fig. 1. Bounding box of the object in a depth frame.

B. Preprocessing

As mentioned, a dynamic hand gesture is performed continuously over time, thus the location and movement of the hands and head should be focused. Besides, the beginning and ending time of each gesture is also an important factor. In this study, the position of the hands and the head of performers are focused. The time for executing every gesture is different so that a long-time gesture corresponds to a large number of consecutive images, vice versa. In the preprocessing step, the object of interest, i.e. the performer, in each input image is determined by applying a thresholding technique. In detail, all pixels in the depth image are classified into two groups specifying the object and background (see Fig. 1). In fact, the performer can stand near or far from the camera at an arbitrary distance, as long as in the active area of Kinect. If a predefined threshold is employed to binarize a depth map, the obtained result is not really good since the intensity of object pixels depends on the distance between the object and the Kinect. Therefore, the well-known Otsu thresholding technique [24] is employed to separate the object from background.

After obtaining binary masks corresponding to all depth frame in a sequence, the smallest bounding box that covers all appeared objects in the sequence is estimated. Besides, background pixels in the frame is changed into the intensity of 255, in order to reduce the effect of background on the spatial-temporal volume representing the sequence of depth images. A 5×5 median filter is also used before thresholding for smoothing the depth image as well as noise removal. Although this filter may slightly change the information on each pixel, the thresholding could be performed more effective.

C. Feature extraction

Feature extraction is an important step, which has a great influence on the effectiveness of the process of automatic model training. The featured values are extracted from the sequences of depth image to effectively distinguish between a gesture and another one. An overview of the feature extraction is shown in Fig. 2.

Let d is the number of frames in the sequence, $h \times w$ is size of the smallest bounding box that was mentioned in the previous section, these frames are combined together along the time axis in order to form a three-dimensional array $\mathbf{A}$ with the size of $h \times w \times d$. The goal of this step is to normalize the size of such arrays to a same resolution of $n \times n \times n$ where n is a predefined whole number.

First, the array $\mathbf{A}$ is considered along the depth, i.e. time, direction. Each two-dimensional array $h \times w$ is resized to $n \times n$.

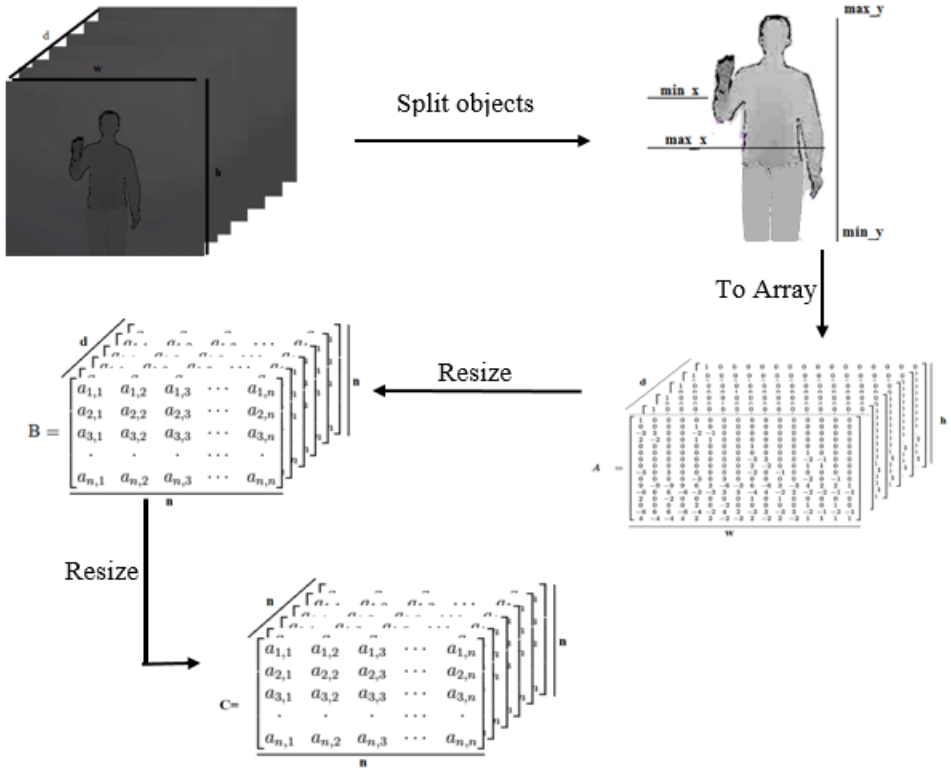


Fig. 2. The diagram of extracting feature for a sequence of depth images corresponding to a gesture.

The results after performing on d arrays are a 3D array **B** with the size of $n \times n \times d$.

$$\mathbf{A}(h, w, d) \rightarrow \mathbf{B}(n, n, d) \tag{1}$$

Next, array **B** is then processed from column 1st to *n*th, in which every 2D array of $n \times d$ is resized to $n \times n$ to obtain a 3D array with the size of $n \times n \times n$.

$$\mathbf{B}(n, n, d) \rightarrow \mathbf{C}(n, n, n) \tag{2}$$

In this work, the process of resizing array is executed by using bicubic interpolation [25] because this method could give smooth results and is used by most of image processing software, digital cameras and printers. In this technique, a new pixel value is calculated based on the mean value of 16 nearest original pixel, i.e. a neighborhood with size of 4×4.

The obtained array of size $n \times n \times n$ is not directly used to represent the feature vector of the corresponding gesture because of the large number of data dimensions and possible noise pixels inside. Therefore, this 3D array is divided into blocks in order to reduce the dimensionality of data as well as the impact of noise since noise level in each block, i.e. the ratio of noise pixels over all elements, is expected to be low. Each block is then converted into a scalar value.

First of all, the elements in the whole array **C** are aligned based on the mean value m. The new value of each element in the array is then recalculated by performing a subtraction on m. The sum of obtained elements in the new array is thus normalized to be zero in order to reduce the impact of the deviation of distance between the performer and Kinect in different sequences of depth images. Next, the new array **C** is divided into a 3D grid with the size of $z \times z \times z$ in which each cell is a block of normalized elements. Finally, each block is represented by a single value corresponding with the average of elements. The result is an array **Z** with the size of z^3, which is much smaller than the original array **C**. Figure 3 illustrates a 1D and 2D representations of array **Z**, in which z is assigned to 4. Each array shape corresponds to an input of the used machine learning models, which support classifying gestures.

D. Support Vector Machine

Support vector machine (SVM) is a supervised learning method which is popularly used for classification and regression analysis. Given a set of training data which was divided into two classes, the SVM algorithm tries to build a binary classification model to separate the input patterns into two defined classes corresponding to the positive class and negative class. Visually, a SVM model builds a super plane to separate input data points in the training set so that the distance from it to nearest points of the two classes is maximized.

In order to create a multi-class support vector machine, one of these two strategies including one-against-all and one-against-one was usually used. In this study, the latter one is selected because of its efficiency and stability. Specifically, a collection of $k(k-1)/2$ binary SVM classifiers, where k is the number of gestures, is built. Each classifier is trained based on data extracted from two classes. In summary, multi-class classification problem is solved by an ensemble of binary SVM classifiers in this work.

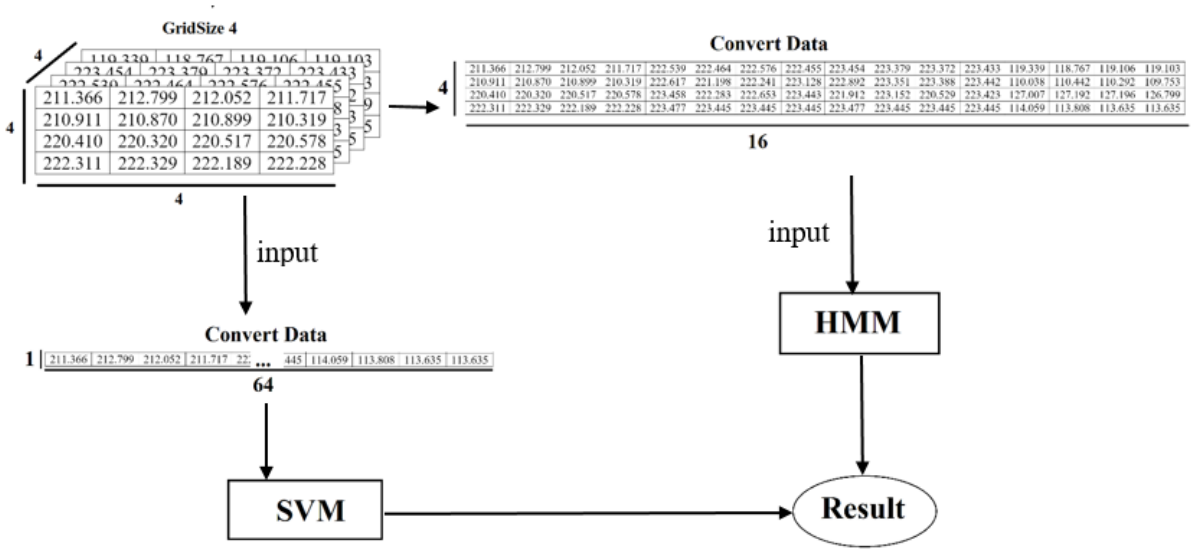


Fig. 3. Low-dimensional array **Z** and different representations for various training models.

E. Hidden Markov Model

Hidden Markov Model (HMM) is a statistical model in which the system being modeled is assumed to be a Markov process with unobserved (hidden) states. This technique was proposed and developed in [26]. With a system involving N states which are numbered in order from 1 to N, HMM is characterized by following elements:

- N is the number of states
- $S = \{s_1, s_2, ..., s_N\}$ is the set of states
- M is the number of distinct observations
- $V = \{v_1, v_2, ..., v_M\}$ is the set of observations
- $A = \{a_{ij}\}$ is the set of transition probabilities
- $a_{ij} = \mathrm{P}(q_{t+1} = s_j | q_t = s_i)$, in which $1 \leq i, j \leq N$, q_t is the actual state at time t
- B is the set of output probability distribution
- π is the initial state distribution, i.e. $\pi_i = \mathrm{P}(q_1 = s_i)$
- $\lambda = (A, B, \pi)$ is the compact notation of a HMM

The objective of HMM related problems includes determining the probability which a sequence of observations is generated from a HMM model in case of testing, and approximating a HMM based on a training set of series of observations in case of training. This technique is selected to perform a comparison because it was employed to solve many problems related to temporal information (e.g. gait assessment [27], [28]).

IV. Experiment

A. Dataset

The dataset that was used in the experiments was built from 5 volunteers with the average distance between each one and the camera is about 2.5m. Each volunteer performed 30 predefined gestures with 20 times for each one, corresponding to 600 sequences of depth images. Each depth image is created at 30 fps with the resolution of 512×424 pixels. Figure 4 illustrates some gestures in the recorded dataset.

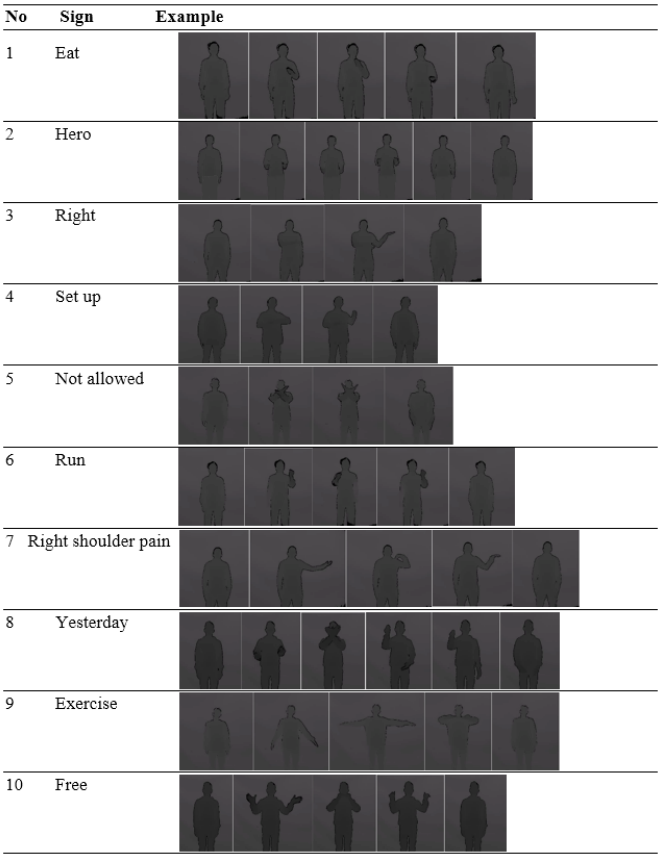


Fig. 4. Examples of some words in our dataset containing depth sequences.

In total 3000 patterns recorded by 5 volunteers, the training

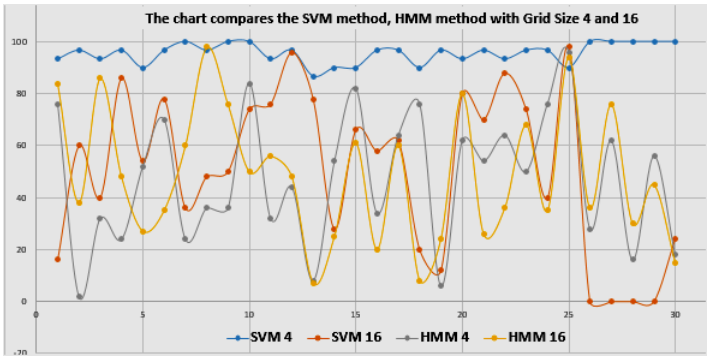


Fig. 5. The result of SVM, HMM when the grid size z is assigned to 4 and 16, respectively.

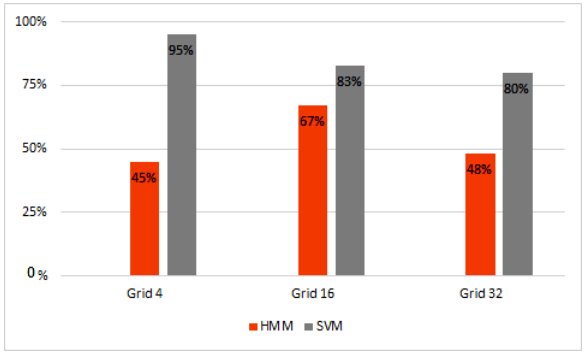


Fig. 6. Accuracy of different machine learning techniques on 30 gestures with different values of z.

set is formed by 1800 samples corresponding to 3 subjects, and the remaining is used for testing stage. This experiment was performed with different grid sizes z as mentioned in section III-C.

B. Experimental results

Figure 5 and 6 show the accuracy with different experiments, using HMM and SVM approach, respectively. It is obvious to see that the SVM method with the 3D grid size $4\times4\times4$ gives the highest result, which is about 95%, compared with other ones.

In order to perform a comparison with a state-of-the-art approach, the method proposed in [29] was applied on the dataset in section IV-A. Differently to the preprocessing in [29] where the stage of hand segmentation is performed on color images, the built-in hand detector of the Kinect was employed to determine the hand silhouette in this experiment. The resulting average accuracy was 93%, which is lower than the best one (95%) of the proposed method. This is because the employed dataset focuses on motion trajectory of the hand while the hand geometry is more difficult to be described compared with the dataset in [29]. The classification accuracies of the proposed approach with different sizes of grid and the method [29] are shown in Table I.

TABLE I. CLASSIFICATION ACCURACIES OF PROPOSED APPROACH AND THE METHOD [29]

Method	**HMM**			**SVM**			**[29]**
Grid size	**4**	**16**	**32**	**4**	**16**	**32**	
Accuracy	0.45	0.67	0.48	**0.95**	0.83	0.80	0.93

V. CONCLUSION

This paper propose an approach for performing feature extraction and recognizing of hand gesture in VSL with the data collected from a Kinect camera. The experimental result on 3000 gestures has confirmed the classification ability of this approach on VSL since the highest accuracy is up to 95%.

With such promising results, we intend to expend the experiment with more words as well as complicated gestures, e.g. combining the hands with motion of other body parts such as head and shoulder, and also to build a system supporting communication between hard-of-hearing people, which focus on the deaf in the context of VSL.

ACKNOWLEDGMENT

The authors would like to thank Trong-Nguyen Nguyen, DIRO, University of Montreal, for his helpful comments. This work was supported by the Polytechnic Computer Vision Group (PCVG), University of Science and Technology, The University of Danang.

REFERENCES

[1] Y. Zhu, Z. Yang, and B. Yuan, "Vision based hand gesture recognition," in *Service Sciences (ICSS), 2013 International Conference on*. IEEE, 2013, pp. 260–265.

[2] K. K. Jung, J. W. Kim, H. K. Lee, S. B. Chung, and K. H. Eom, "Emg pattern classification using spectral estimation and neural network," in *SICE Annual Conference 2007*, Sept 2007, pp. 1108–1111.

[3] X. Zhang, X. Chen, W.-h. Wang, J.-h. Yang, V. Lantz, and K.-q. Wang, "Hand gesture recognition and virtual game control based on 3d accelerometer and emg sensors," in *Proceedings of the 14th International Conference on Intelligent User Interfaces*, ser. IUI '09. New York, NY, USA: ACM, 2009, pp. 401–406. [Online]. Available: http://doi.acm.org/10.1145/1502650.1502708

[4] V. E. Kosmidou and L. J. Hadjileontiadis*, "Sign language recognition using intrinsic-mode sample entropy on semg and accelerometer data," *IEEE Transactions on Biomedical Engineering*, vol. 56, no. 12, pp. 2879–2890, Dec 2009.

[5] T. D. Bui and L. T. Nguyen, "Recognizing postures in vietnamese sign language with mems accelerometers," *IEEE Sensors Journal*, vol. 7, no. 5, pp. 707–712, May 2007.

[6] T. Kuroda, Y. Tabata, A. Goto, H. Ikuta, M. Murakami, and T. Limited, "Consumer price data-glove for sign language recognition," in *In: Proc. of 5th Intl Conf. Disability, Virtual Reality Assoc. Tech*, 2004, pp. 253–258.

[7] S. A. Mehdi and Y. N. Khan, "Sign language recognition using sensor gloves," in *Neural Information Processing, 2002. ICONIP '02. Proceedings of the 9th International Conference on*, vol. 5, Nov 2002, pp. 2204–2206 vol.5.

[8] H. Brashear, V. Henderson, K.-H. Park, H. Hamilton, S. Lee, and T. Starner, "American sign language recognition in game development for deaf children," in *Proceedings of the 8th International ACM SIGACCESS Conference on Computers and Accessibility*, ser. Assets '06. New York, NY, USA: ACM, 2006, pp. 79–86. [Online]. Available: http://doi.acm.org/10.1145/1168987.1169002

[9] K. Assaleh and M. Al-Rousan, "Recognition of arabic sign language alphabet using polynomial classifiers," *EURASIP J. Appl. Signal Process.*, vol. 2005, pp. 2136–2145, Jan. 2005. [Online]. Available: http://dx.doi.org/10.1155/ASP.2005.2136

[10] X. Li, "Gesture recognition based on fuzzy c-means clustering algorithm," *Department Of Computer Science The University Of Tennessee Knoxville*, 2003.

[11] S. Mitra and T. Acharya, "Gesture recognition: A survey," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 37, no. 3, pp. 311–324, May 2007.

[12] V. Athitsos and S. Sclaroff, *Database Indexing Methods for 3D Hand Pose Estimation*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 288–299.

[13] R. Bowden, D. Windridge, T. Kadir, A. Zisserman, and M. Brady, *A Linguistic Feature Vector for the Visual Interpretation of Sign Language*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 390–401.

[14] A. C. Kak, "Purdue rvl-slll asl database for automatic recognition of american sign language," in *Proceedings of the 4th IEEE International Conference on Multimodal Interfaces*, ser. ICMI '02. Washington, DC, USA: IEEE Computer Society, 2002, pp. 167–. [Online]. Available: http://dx.doi.org/10.1109/ICMI.2002.1166987

[15] D. V. Hieu and S. Nitsuwat, "Image preprocessing and trajectory feature extraction based on hidden markov models for sign language recognition," in *2008 Ninth ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing*, Aug 2008, pp. 501–506.

[16] T.-N. Nguyen, H.-H. Huynh, and J. Meunier, "Static hand gesture recognition using artificial neural network," *Journal of Image and Graphics*, vol. 1, no. 1, pp. 34–38, 2013.

[17] Z. Ren, J. Yuan, and Z. Zhang, "Robust hand gesture recognition based on finger-earth mover's distance with a commodity depth camera," in *Proceedings of the 19th ACM International Conference on Multimedia*, ser. MM '11. New York, NY, USA: ACM, 2011, pp. 1093–1096. [Online]. Available: http://doi.acm.org/10.1145/2072298.2071946

[18] Y. Li, "Hand gesture recognition using kinect," in *2012 IEEE International Conference on Computer Science and Automation Engineering*, June 2012, pp. 196–199.

[19] X. Liu and K. Fujimura, "Hand gesture recognition using depth data," in *Sixth IEEE International Conference on Automatic Face and Gesture Recognition, 2004. Proceedings.*, May 2004, pp. 529–534.

[20] D. H. Vo, T. N. Nguyen, H. H. Huynh, and J. Meunier, "Recognizing vietnamese sign language based on rank matrix and alphabetic rules," in *2015 International Conference on Advanced Technologies for Communications (ATC)*, Oct 2015, pp. 279–284.

[21] Z. Zafrulla, H. Brashear, T. Starner, H. Hamilton, and P. Presti, "American sign language recognition with the kinect," in *Proceedings of the 13th International Conference on Multimodal Interfaces*, ser. ICMI '11. New York, NY, USA: ACM, 2011, pp. 279–286. [Online]. Available: http://doi.acm.org/10.1145/2070481.2070532

[22] S. Lang, M. Block, and R. Rojas, *Sign Language Recognition Using Kinect*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 394–402.

[23] X. Chai, G. Li, Y. Lin, Z. Xu, Y. Tang, X. Chen, and M. Zhou, "Sign language recognition and translation with kinect," in *IEEE Conf. on AFGR*, 2013.

[24] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, no. 1, pp. 62–66, Jan 1979.

[25] R. Keys, "Cubic convolution interpolation for digital image processing," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 29, no. 6, pp. 1153–1160, Dec 1981.

[26] A. A. Markov, "An example of statistical investigation in the text of 'Eugene Onyegin' illustrating coupling of 'tests' in chains," in *Proceedings of the Academy of Sciences*, vol. 7 of VI, St. Petersburg, 1913, pp. 153–162.

[27] T.-N. Nguyen, H.-H. Huynh, and J. Meunier, "Skeleton-based abnormal gait detection," *Sensors*, vol. 16, no. 11, 2016. [Online]. Available: http://www.mdpi.com/1424-8220/16/11/1792

[28] L. Tao, A. Paiement, D. Damen, M. Mirmehdi, S. Hannuna, M. Camplani, T. Burghardt, and I. Craddock, "A comparative study of pose representation and dynamics modelling for online motion quality assessment," *Computer vision and image understanding*, vol. 148, pp. 136–152, 2016.

[29] D.-H. Vo, H.-H. Huynh, and T.-N. Nguyen, "Modeling dynamic hand gesture based on geometric features," in *2014 International Conference on Advanced Technologies for Communications (ATC 2014)*, Oct 2014, pp. 471–476.

High Performance of Hash-based Signature Schemes

Ana Karina D. S. de Oliveira
FACOM-UFMS
Federal University of Mato Grosso do Sul
MS, Brazil

Julio López
IC-UNICAMP
University of Campinas
SP, Brazil

Roberto Cabral
UFC-CRATEÚS
Federal University of Ceará
CE, Brazil

***Abstract*—Hash-based signature schemes, whose security is based on properties of the underlying hash functions, are promising candidates to be quantum-safe digital signatures schemes. In this work, we present a software implementation of two recent standard proposals for hash-based signature schemes, Leighton and Micali Signature (LMS) scheme and Extended Merkle Signature Scheme (XMSS), using a set of AVX2 instructions on Intel processors. The implementation uses several optimization techniques for speeding up the underlying hash functions SHA2 or SHA3, and other building block functions which lead to high performance for signature operations on both schemes. On an Intel Skylake processor, using a tree of height 60 with 12 layers, the signing operation for XMSS takes 3,841,199 cycles (1,043 signatures per second) at 128-bit security level (against quantum attacks). For an equivalent security, the LMS system computes a signature in 1,307,376 cycles (3,065 signatures per second). We also provide the first comparative performance results for signing and verification of both schemes using different parameters. The results of our implementation indicate that both schemes LMS and XMSS can achieve high performance using vector instructions on modern processors.**

***Keywords*—post-quantum cryptography; digital signature; Merkle signature; LMS; XMSS**

I. Introduction

A digital signature scheme is an important cryptographic tool for public-key cryptography. Digital signature scheme are widely used for providing authenticity, integrity, and non-repudiation of data. Nowadays, the most commonly used digital signature schemes are ECDSA [1], RSA [2] and DSA [3]. These schemes have their security based on the difficulty of factoring large integers or computing discrete logarithms. In [4], Shor introduced a polynomial-time quantum algorithm for factoring and computing discrete logarithms. Thus, digital signature schemes that can resist an attack by quantum computers are an active area of research.

A One-Time Signature (OTS) scheme allows using a key pair to sign exactly one message [5]. These schemes are inadequate for the most practical situations since each key pair is used only for a single signature. In [6], Merkle proposed N-Time Signature (NTS), that are built out of one-time signature schemes. The Merkle Signature Scheme (MSS) makes one-time signatures practical by combining $N = 2^h$ of them in a single structure, which is a complete binary tree of height h. These systems have regained interest since 2006 because of their resistance against quantum-computer-aided attacks. Since the security of these schemes is based on the underlying cryptographic hash function, they are called hash-based signature schemes.

Independently of the actual realization of quantum computing, governmental and standardization organizations are encouraging the transition to post-quantum cryptography, i.e. cryptographic schemes not known to be vulnerable to quantum computer attacks [7]. Standardization efforts are under way, for example, the National Institute of Standards and Technology (NIST) is now accepting submissions for quantum-resistant public-key cryptographic algorithms [8].

Some MSS variants were proposed: An improved Merkle signature scheme (CMSS) [9] builds two chained trees allowing the signature of 2^{40} messages and also reduce the runtime of key pair and signature generation. The Merkle signatures with virtually unlimited signature capacity (GMSS) [10] allow to sign a significant number of messages (2^{80}) with one key pair. XMSS [11] introduced a signature scheme with minimal security requirements. A hierarchical based-hash signature XMSS^{MT} [12] allows signing a large but a fixed number of messages. SPHINCS [13] is a practical stateless hash-based signature scheme and introduces a new method to randomize tree-based stateless signatures. SPHINCS has significantly larger signatures, which could make it impractical in some scenarios. In 2016, the work [7] analyzes the state management in hash-based schemes N-times and proposes a hybrid stateless/stateful scheme to protect against unintentional copies of the state of the private key and has smaller and faster signatures.

There are two proposals for standards of hash-based signature schemes: the first one [14] describes the LMS, an adaptation of the one-time signature scheme Lamport-Diffie-Winternitz-Merkle [15]. The second one [16] describes XMSS. Therefore, the design and efficient implementation of secure and practical digital signature schemes are crucial for applications that require data integrity assurance and data origin authentication.

Our Contribution. In this work, we present a software implementation of two recent standard proposals for hash-based signatures schemes, LMS and XMSS, using the Intel AVX2 vector instruction set. We use parallel optimization techniques for improving the performance of the underlying hash functions SHA2 and SHA3. We also show how to speed up the main building blocks of LMS and XMSS by taking advantage of the fastest implementation of SHA2 and SHA3. We provide a comparative performance analysis of both schemes.

Organization. The rest of this paper is organized as follows. We describe: the Winternitz One-Time Signature (WOTS) and (WOTS^{+}) in Section II, the MSS and XMSS in Section III, Hierarchical Signatures Scheme (HSS) in Section

IV and LMS and XMSS drafts in Section V. We present the target microarchitectures in Section VI. We discuss the software optimizations in Sections VII and VIII. In Section IX we show the performance results. In Section X we present the conclusions.

II. Winternitz One-Time Signature (WOTS) and (WOTS$^+$)

The OTS are used to validate the authenticity of a message by associating a secret private key with a shared public key [14]. In these one-time signatures, each private key must be used only one time to sign any given message. As a part of the signing process, a digest of the original message is computed using a cryptographic hash function H, and the resulting digest is signed. The WOTS [15] is a modification of the Lamport One-Time Signature (LOTS) [5]. WOTS uses a parameter w which is the number of bits to be signed simultaneously. This scheme produces smaller signatures than Lamport, but increases the number of one-way function evaluations from 1 to $2^w - 1$, in each element of the signing key. Hülsing [17] proposed WOTS$^+$, a modification of WOTS that uses a chaining function f^e starting from random inputs. This modification allows eliminating the requirement to use a collision-resistant hash function.

WOTS uses a one-way function $f : \{0,1\}^n \rightarrow \{0,1\}^n$ and a cryptographic hash function $g : \{0,1\}^* \rightarrow \{0,1\}^n$, where n is positive integer. The WOTS chaining function f^e computes e iterations of f on input $x \in \{0,1\}^n$ where $e \in \mathbb{N}$ $(e < w)$.

The WOTS chaining function is defined as:

$$f^e(x) = \begin{cases} x & \text{if } e = 0; \\ f(f^{e-1}(x)) & \text{if } e > 0. \end{cases}$$

Similarly, the WOTS$^+$ chaining function f^e computes e iterations of f_K on inputs key $K \in \{0,1\}^n$ chosen randomly, $x \in \{0,1\}^n$ and bitmask $bm = (bm_1, ..., bm_{w-1})$ chosen randomly, with $e \in \mathbb{N}$. Then, the chaining function f^e is defined as:

$$f^e(x, bm) = \begin{cases} x & \text{if } e = 0; \\ f_K(f^{e-1}(x, bm) \oplus bm_e) & \text{if } e > 0. \end{cases}$$

These schemes are parameterized by a security parameter n and the Winternitz parameter $w \in \mathbb{N}$, for $w > 1$. The values n and w are used to compute len (number of elements of the signature), where $len = len_1 + len_2$.

In WOTS: $len_1 = \lceil n/w \rceil$, $len_2 = \lceil (\lfloor log_2 len_1 \rfloor + 1 + w)/w \rceil$.

In WOTS$^+$: $len_1 = \lceil n/(log_2(w)) \rceil$, and $len_2 = \lfloor (log_2(len_1(w-1)))/(log_2(w)) \rfloor + 1$.

A. WOTS/WOTS$^+$ key pair generation

The private keys $sk = (sk_0, \ldots, sk_{len-1})$ can be generated uniformly at random, or via a pseudorandom process. The public verification key is $pk = (pk_0, ..., pk_{len-1}) \in \{0,1\}^{(n,len)}$.

In WOTS: $pk_i = f^{2^w - 1}(sk_i)$.

In WOTS$^+$: $pk_i = f^{w-1}(sk_i, bm)$.

B. WOTS/WOTS$^+$ signature generation

To generate the signature of a message M, first compute the message digest $d = g(M)$. Then, d is split into len_1 binary blocks, resulting in $d = (m_0||...||m_{len_1-1})$, where $||$ denotes concatenation. The checksum c is computed and added to d, where c can be divided into len_2 blocks $c = (c_0|| \ldots ||c_{len_2-1})$.

In WOTS: $c = \sum_{i=0}^{len_1-1}(2^w - m_i)$.

In WOTS$^+$: $c = \sum_{i=0}^{len_1-1}(w - 1 - m_i)$.

Let $b = d||c$ be the concatenation of the extended string d with the extended string c. Thus $b = (b_0||b_1|| \ldots ||b_{len-1}) = (m_0|| \ldots ||m_{len_1-1}||c_0|| \ldots ||c_{len_2-1})$. The signature of the message M is $sig\_ots = (sig_0, ..., sig_{len-1})$, where:

In WOTS: $sig_i = f^{b_i}(sk_i)$.

In WOTS$^+$: $sig_i = f^{b_i}(sk_i, bm)$.

C. WOTS/WOTS$^+$ verification

To verify the signature sig_ots of the message M, we compute $(b_0, ..., b_{len-1})$ in the same way as it was calculated during signature generation. Then, we compute: $temp\_sig = (sig'_0, \ldots, sig'_{len-1})$.

In WOTS: $sig'_i = f^{2^w - 1 - b_i}(sig_i)$.

In WOTS$^+$: $sig'_i = f^{w-1-b_i}(sig_i, bm)$.

If $temp\_sig = pk_i$ for $i = \{0, 1, ..., len - 1\}$, then the signature is valid, otherwise is invalid.

III. Merkle Signature Scheme (MSS)

MSS [15] is a digital signature scheme that consists of three algorithms: key generation, signing and verification. This scheme constructs a binary tree where the leaves are the verification keys, and the public key is the root of the tree. This key pair can sign/verify messages. A tree of height h and 2^h leaves will have 2^h one-time key pairs. The digest of the one-time verification public key $(g(pk_0||...||pk_{t-1}))$ will be a leaf of the Merkle tree.

A. MSS key pair generation

First, the signer must select the tree height $h \in N$, $h \geq 2$. Merkle uses a cryptographic hash function $g : \{0,1\}^* \rightarrow \{0,1\}^n$, where n is a positive integer. The treehash algorithm [15] is used to generate the public key that is the root of the tree. The authentication path (Aut) is formed by the sibling right nodes, connecting the leaf up to the tree root, which is used to validate the public key. Aut is saved during the execution of the treehash algorithm.

B. MSS signature generation

The signature generation consists of two steps: first, the signature of the message digest $g(M)$ is generated using the WOTS signature algorithm and the corresponding secret key sk_s of the leaf s. Then, the signature $SIG = (s, sig_s, Aut)$ contains the index of the leaf s, the WOTS signature sig_s, and the authentication path Aut. In the second step, the next authentication path Aut is generated. This step can be

done efficiently with the algorithm proposed by [18] which is a modification of the classic authentication path algorithm proposed by Merkle [6].

C. MSS verification

The signature verification consists of two steps: first, the signature sig_s is used to recover a leaf of the tree. Second, the public key of the Merkle tree is validated in the following way. The receiver can reconstruct the path $(p_0, ..., p_h)$ from leaf s to root. The index s is used to decide the order in which the authentication path is reconstructed. Initially, $p_0 = Y_s$. For each $i = 1, 2, \ldots, h$, p_i is computed using the condition (if $\lfloor s/(2^{i-1}) \rfloor \equiv 1 \bmod 2$) and the recursive formula:

$$p_i = \begin{cases} g(Aut_{i-1} \| p_{i-1}) & \text{if } \lfloor s/(2^{i-1}) \rfloor \equiv 1 \bmod 2; \\ g(p_{i-1} \| Aut_{i-1}) & \text{otherwise.} \end{cases}$$

Finally, if the value p_h is equal to the public key pub, the signature is valid.

D. Extended Merkle Signature Scheme (XMSS)

XMSS [11] is a modification of MSS. This scheme uses a slightly modified version of Winternitz WOTS$^+$ described in Section II. XMSS is provably forward-secure and efficient when instantiated with two secure and efficient function families: one second-preimage resistant hash function family G_n and the other a pseudorandom function family F_n, where $G_n = \{g_K : \{0,1\}^{2n} \rightarrow \{0,1\}^n | K \in \{0,1\}^n\}$ and $F_n = \{f_K : \{0,1\}^n \rightarrow \{0,1\}^n | K \in \{0,1\}^n\}$.

The parameters of XMSS are: $n \in \mathbb{N}$, the security parameter; $w \in \mathbb{N}(w > 1)$, the Winternitz parameter; $m \in \mathbb{N}$, the message digest length; and $h \in \mathbb{N}$, the height of the binary tree.

An XMSS binary tree is constructed to generate the public key pub. The XMSS tree is a modification of the Merkle tree. A tree of height h has $h + 1$ levels. The nodes on level j are $node_{i,j}$, for $0 < j \leq h$ and $0 \leq i < 2^{h-j}$. XMSS uses the hash function g_K and bitmask (bitmaskTree) $bm \in \{0,1\}^{2n}$, chosen uniformly at random, where bm_{2i+2j} is the left bitmask and $bm_{2i+2j+1}$ is the right bitmask. The bitmasks are the main difference among the others Merkle tree constructions since they allow to replace the collision-resistant hash function family by a second-preimage resistant hash function family [11]. The nodes are computed as: $node_{i,j} = g_K((node_{2i,j-1} \oplus bm_{2i+2j}) || (node_{2i+1,j-1} \oplus bm_{2i+2j+1}))$.

To generate a leaf in the XMSS tree, a Ltree is used. The Ltree [11] uses bitmasks in the same form as in the XMSS tree. The WOTS$^+$ public verification keys $(pk_0, \ldots, pk_{len-1})$ are the first len leaves of a Ltree. If len is not a power of 2, then there are not sufficiently leaves to build a binary tree. Therefore, a node that has a no right sibling is lifted to a higher level of the Ltree until it becomes the right sibling of another node.

IV. Hierarchical Signatures Scheme (HSS)

A hierarchical signature scheme is an N-time signature scheme that uses other hash-based signatures in its construction [7]. Some schemes use this constructions as in CMSS [9], GMSS [10], XMSS^{MT} [12], LMS [14] and SPHINCS [13]. The basic construction of HSS consists of a tree with d layers of subtrees, for $i = 0, \ldots, d-1$, where the lower layer is $i = d-1$. The trees on top and intermediate layers are used to sign the root nodes of the trees on the respective layer below. Trees on the lowest layer are used to sign the actual messages. All trees can have equal height.

An HSS private key consists of the private keys of each level. The public key is the root of the top level. A signature HSS consists of the public keys of levels 1 to $(d-1)$, along with the signatures in each level, and the signature of the message M with the private key of the lower level $(d-1)$. Hierarchical signatures allow for shorter signing time of a message M while offering a larger number of signed messages.

V. LMS and XMSS Drafts

Among the variants of the Merkle scheme, we chose the two standard proposals for hash-based signatures to implement: LMS [14] and XMSS[16]. LMS system is an adaptation of the original Lamport-Diffie-Winternitz-Merkle one-time signature system [15] and uses the WOTS and the HSS. XMSS specifies the one-time signature scheme (WOTS$^+$), a single-tree (XMSS) and a multi-tree variant (XMSS^{MT}) of XMSS.

A. LMS

Leighton and Micali [14], introduce a "security string" that is distinct for each invocation of H to improve security against attacks that amortize their effort against multiple invocations of the hash function H. The following fields can appear in a security string: (I, D_ITER, D_PBLC, D_MESG, D_LEAF, D_INTR, C, r, q, i, j) as described in [14]. The values I, $D\_$ and C must be chosen uniformly at random, or via a pseudorandom process; r is the node number associated with a particular node of a hash tree; q is set to be the leaf number of the hash tree; i is the index of the private key element ($pk[i]$); and j is the iteration number used when the private key element is being iteratively hashed. To generate a leaf ($leaf[q]$) in the LMS tree, the hash functions are used: $tmp = H\_leaf(X)$=$Hash(X)$, where $X = (S||pk[0]|| \ldots ||pk[p-1]||D\_PBLC)$ and $leaf[q] = H\_node(Y)$=$Hash(Y)$, where $Y = (I||tmp||u32str(r)||D\_LEAF)$.

B. XMSS

XMSS [16] randomize each hash function call; this means that aside of the initial message digest, for each hash function call a different key and different bitmask is used. These values are pseudorandomly generated using a pseudorandom function that takes a key $SEED$ and a 32-byte address $ADRS$ and outputs a n-byte value, where n is the security parameter. There are three different types of addresses; one type for the hashes used in one-time signature schemes, one for hashes used within the main Merkle-tree construction, and one for hashes used in the Ltrees.

C. Functions used in LMS and XMSS

This section describes the differences between the main functions of both schemes LMS and XMSS. Let F be the chain function used to generate the private keys, sign and verify messages. Let G be the function used to generate the inner nodes of the tree. Let I,r,i,q,j be the security strings defined in Section V-A; S=I+q; $Left$ and $Right$ be nodes left and right; KEY be a key; BM, BM_0 and BM_1 be bitmasks; sk_i be the WOTS secret key. The function $u\mathbf{Y}str(X)$ takes a nonnegative integer X as input and return $\mathbf{Y}/8$ byte strings.

In the LMS, the main functions have the following input sizes:

- $F(X)$ = Hash(X), where X is composed of $(S||sk_i||u16str(i)||u8str(j)||D\_ITER)$ and $|X|$ is $(3n+8)$ bytes.
- $G(Y)$ = Hash(Y), where Y is composed of $(I||node[2.r]||node[2.r+1]||u32str(r)||\ D\_INTR)$ and $|Y|$ is $(4n+5)$ bytes.

In the XMSS, the main functions have the following input sizes:

- $F(X)$ = Hash(X), where X is composed of $((toByte(0,n)||KEY||sk_i\ XOR\ BM))$ and $|X|$ is $(3n)$ bytes.
- $G(Y)$ = Hash(Y), where Y is composed of $((toByte(1,n)||KEY||(Left\ \ XOR\ \ BM_0)||(RightXORBM_1))$ and $|Y|$ is $(4n)$ bytes.

D. Keys LMS and XMSS

The sizes of the private key(SK), the verification key (PK) and the signature (Sig) are described below.

In LMS:

- $SK = (q, SEED\_sk, SEED\_I)$ has $2n+4$ bytes, given that q requires 4 bytes, the seed to generate the secret key and the seed to generate the identifier I have n bytes.
- $PK = (I, T[1])$ has $3n$ bytes, given that the identifier I has $2n$ bytes and the root of the tree $(T[1])$ has n bytes.
- $Sig = (q, sig\_ots, auth[0], \ldots, auth[h-1])$ has $(p+1+h)n+4$ bytes, given that the index q has 4 bytes, the WOTS signature has a random value C with n bytes and sig with pn bytes; the authentication path has hn bytes.

In XMSS:

- $SK = (idx, wots\_sk, SK\_PRF, root, SEED)$ has $4n+4$ bytes, given that the index leaf idx requires 4 bytes, and the secret key $wots\_sk$, the key SK_PRF, the root $root$ and the seed $SEED$ require n bytes.
- $PK = (root, SEED)$ has $2n$ bytes, given the $root$ and $SEED$ require n bytes.
- $Sig = (idx\_sig, r, sig\_ots, auth[0], \ldots, auth[h-1])$ has $(len+h+1)n+4$ bytes, given that the idx_sig has 4 bytes, the random value r has n bytes, the WOTS+ signature require $lenn$ bytes, and an authentication path require hn bytes.

E. Security considerations

LMS is provably secure in the random oracle model, as shown by Katz [19]. From Theorem 8 of that reference: *for any adversary attacking arbitrarily many instances of the one-time signature scheme, and making at most q hash queries, the probability with which the adversary can forge a signature with respect to any of the instances is at most* $q2^{(1-8n)}$ [14]. The format of the inputs to the hash function have the property that each invocation of that function has an input that is distinct from all others, with high probability. This property is important for a proof of security in the random oracle model. Let n be the number of bytes in the output of the hash function. Therefore, we use n = 32 to have a security level of 128 bits, even assuming that there are quantum computers that can compute the input to an arbitrary function with a computational cost equivalent to the square root of the size of the domain of that function.

XMSS provides strong security guarantees and is even secure when the collision resistance of the underlying hash function is broken. Parameters are accompanied by a bit security value. The meaning of bit security is that a parameter set grants b bits of security if the best attack takes at least $2^{(b-1)}$ bit operations to achieve a success probability of $1/2$. Hence, to mount a successful attack, an attacker needs to perform 2^b bit operations on average [20]. According to the security proof in [16], it is not sufficient to break the collision resistance of the hash functions to generate a forgery. More specifically, the requirements on the used functions are that F and G are post-quantum multi-function multi-target second-preimage resistant keyed functions, F fulfills an additional statistical requirement that roughly says that most images have at least two preimages, PRF is a post-quantum pseudorandom function, H_msg is a post-quantum multi-target extended target collision resistant keyed hash function.

VI. Target Microarchitectures

In this section, we describe the microarchitecture details of the Intel processors (Haswell and Skylake) used in this work. The Haswell microarchitecture, launched in 2013, supports the AVX2 vector instruction set, which expanded the integer arithmetic instructions of 128-bit to 256-bit registers. A single AVX2 instruction can operate eight 32-bit values or four 64-bit values at the same time. These instructions allowed four hashes could be processed concurrently for the SHA2-512/SHAKE128 and eight hashes for SHA2-256.

The Skylake microarchitecture, released in 2015, is based on the Haswell and Broadwell microarchitecture [21]. Skylake improved the latency of some instructions. Some instructions in Skylake (such as `vmov, vpand` and `vpsllq`) have better throughput and can be used to better schedule the instructions.

In the following, we describe general aspects of these micro-architectures, and the most relevant vector instructions used in this work.

A. Vector operations

In the late 1990s, processor manufacturers focused their efforts on exploiting data parallelism rather than instruction parallelism. Thus, they incorporated functional units that could execute a single instruction over a set of data. This processing fits into the paradigm of parallel computing known as Single Instruction Multiple Data (SIMD) [22].

In 1997, Intel launched its first set of instructions to implement the SIMD paradigm; called Multimedia eXtensions (MMX). MMX added 64-bit registers and vector instructions that enabled the processing of two 32-bit operations; at that time, the architectures had native 32-bit registers [21].

In 1999, Intel released the Streaming SIMD Extensions (SSE) that included eight 128-bit registers (XMMs); the number of registers was doubled in the next year when the size of native registers increased to 64. In the following years, the SSE has evolved with the launch of the new instructions sets SSE2, SSE3, e SSE4 [21].

In 2011, it was launched the Advanced Vector eXtensions (AVX) instruction set, which introduced significant contributions to the architecture; were included 256-bit registers, called YMMs, that are overlapped on XMMs registers. Also, AVX introduced a new encoding format that allows the use of three-operand assembly code, making the assignment of registers more flexible.

The code that is compiled for an instruction set can be executed only if both the CPU and the operating system support such set. Some compilers, like GCC, Clang and ICC can perform vector operations automatically (without programmer interference); however, it is not easy to determine whether the code can be vectorized. It is possible to vectorize code explicitly, by writing the code in assembly or using intrinsic functions. The intrinsic functions are primitive operations in the sense that each intrinsic function is translated into one or more machine instructions.

B. Haswell

The Haswell microarchitecture, Intel's 4th generation Core processor family, was launched in early 2013 and presenting a series of improvements on performance and also new instructions. There are instructions in the Bit Manipulation Instruction (BMI), feature group that aid in SHA2 (RORX) and RSA (MULX) performance increases. Also besides, the new instruction set AVX2 that promote vector operations from 128 bits to 256 bits, increasing performance of integer operations [23]. AVX2 has permutation and combination instructions that allow moving the words contained in vector registers [24].

C. Skylake

The Skylake microarchitecture was launched in 2015. Skylake offers the following enhancements: larger internal buffers, higher cache bandwidth, higher throughput, better branching predictor, low power consumption, throughput balancing, and reduced floating point. A significant portion of the SSE, AVX, AVX2 and general purpose instructions also had latency improvements [21].

D. Relevant instructions

According to Agner Fog [24], the latency of an instruction is the delay that the instruction generates in a dependency chain, the unit of measure is clock cycles. Another factor that influences performance is throughput, which is the maximum number of instructions of the same type that can be executed per clock cycle when the operands of each instruction are independent of the previous instructions.

In Table I are highlighted some instructions of the AVX2 set that are relevant to the context of the efficient implementation of XMSS and LMS. In this table are shown the latency, the throughput and the execution ports in Haswell and Skylake [24].

TABLE I. LATENCY, THROUGHPUT AND EXECUTION PORT [24], [25]

Haswell			Execution port			
Vector instruction	Latency	Throughput	0	1	2 . . . 4	5
`vmov`	1	3	x	x		x
`vpadd`	1	2		x		x
`vpand/vpor/vpxor`	1	3	x	x		x
`vpsllq`	1	1	x			
Skylake			Execution port			
Vector instruction	Latency	Throughput	0	1	2 . . . 4	5
`vmov`	1	4	x	x		x
`vpadd`	1	3	x	x		x
`vpand/vpor/vpxor`	1	3	x	x		x
`vpsllq`	1	2	x	x		

The ports 0, 1 and 5 in Table I are the most used ports, and therefore, the most critical in determining the efficiency of the implementation. Note that some instructions in Skylake have better throughput; this can be used to schedule instructions to take advantage of this fact.

VII. SOFTWARE OPTIMIZATIONS

In this section, we will discuss the software optimization aspects applied in this work for Intel micro-architectures: Haswell and Skylake. Software optimization is committed to making software faster and smaller and goes beyond of writing a program with few lines of code. One must consider the costs of software development, the programming language used, the security of the code, and the computing power of processors. We will show the most critical parts of our program and how we apply optimizations using AVX2 instructions.

One of the general objectives of this work was to provide techniques that enable the efficient use of vector instruction sets in the implementation of the XMSS and LMS. Because both schemes are based on hash functions, this work shows the results of an efficient implementation that uses 256-bit registers to compute four hash values using SHA2-512/SHAKE128 or eight hash values using SHA2-256 concurrently.

The first optimization for improving both signature schemes uses the computation of multiple hashes at the same time. We call this approach multi-buffer optimization. As data buffers are independent of each other and have messages of the same size, it is possible to take advantage of the data-level parallelization of the hash algorithms. In addition, once the data is loaded into the registers, the data is processed several times by the hash function, performing several iterations of the hash algorithm on the same data, avoiding memory accesses. The result of hash is also returned in the same order as it

was sent, making it easy to implement. This optimization was applied in both hash functions SHA2 and SHA3.

A. SHA2 optimizations

The 256-bit vector instructions can process four 64-bit words or eight 32-bit words using only one instruction. The SHA2-512/SHAKE128 algorithm works internally with 64-bit words and SHA2-256 with 32-bit words. Thus, taking advantage of the 256-bit registers, four hashes could be processed concurrently for the SHA2-512/SHAKE128 and eight hashes for SHA2-256.

For example, in the calculation of $W[t]$:

$$W[t] = \sigma_1^{256}(W[t-2]) + W[t-7] + \sigma_0^{256}(W[t-15]) + W[t-16].$$

The operations with the values of $W[t]$ for the eight messages can be performed in parallel using 256-bit registers. Each $W[t]$ receives and processes 32-bit values. Thus, it is possible to compute the operations for the calculation of eight hashes at the same time with SHA2-256.

1) Optimizations based on processor execution ports: Another optimization was to reformulate the code of the functions that compute the hash of the messages, scheduling the execution of the instructions to improve the throughput. The core strategy of our implementation was to analyze the main functions that process the messages to generate the hash. We look for the instructions for these functions, the ports available to execute them, the latency, and throughput, and the dependencies in the code. We have eliminated several dependencies in the SHA2 algorithm code. This approach allows us to parallelize calculations when there is no dependence between instructions.

The most critical functions are:

- $T_1(h, e, f, g) = h + Sig1(e) + Ch(e, f, g) + K^{256} + W.$
- $T_2(a, b, c) = Sig0(a) + Maj(a, b, c).$

Analyzing the dependency chain in the function T_1:

- $Sig1(x) = Rot^6(x) \oplus Rot^{11}(x) \oplus Rot^{25}(x).$
- $Rot^n(x) = (SL) \vee (SR) = (x \ll (32-n)) \vee (x \gg n).$
- $Ch(e, f, g) = (((f \oplus g) \wedge e) \oplus g).$

As an example, the function $Sig1(x)$ makes three calls to the rotation function $Rot^n(x)$. $Sig1(x)$ performs three shifts to the left (SL), three shifts to the right (SR) and three OR operations. If each call to the $Rot^n(x)$ is performed separately, then the instructions of this function will also be executed separately, underutilizing the available ports on the processor.

Both microarchitectures offer three ports to execute the OR instruction; then we unroll the $Rot^n(x)$ function to perform all (SL) first, followed by (SR). Then, three SL values and the three SR values will be available to execute three OR operations in parallel. In particular, Skylake has one additional port to perform the shift operation; then it is possible to execute two shifts at the same time.

This analysis of the logical functions of the SHA2 algorithm has resulted in an implementation that takes advantage of the available ports by the processor used and improves the throughput. As an example, we illustrate in Figure 1, the execution sequence of the T_1 function instructions in the Haswell processor; the graph represents dependency in the bottom-up design; where the nodes represent the operations of the instructions and the numbers below each node represent the time in clock cycles.

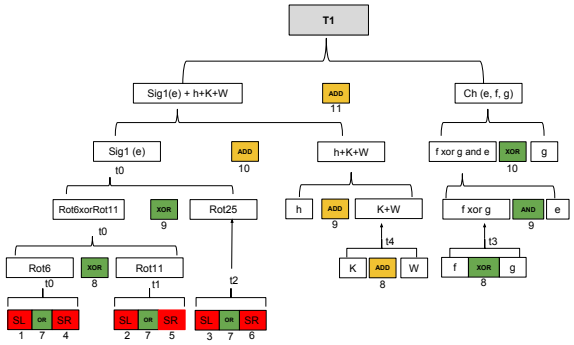


Fig. 1. Instruction scheduling of the function T_1

.

The shift (SH) to the left (SL) and right (SR) of Figure 1 can be calculated in time 1 to 6. According to Table I, this instruction has one cycle of latency and throughput one on Haswell. The three operations OR are executed at time seven because the throughput of this logical operation is three. Since the latency of the OR instruction is one cycle, at time eight the next instructions already can be executed.

Overall, the latency will be one cycle if we look isolated instructions, but if we look at a long chain of instructions of the T_1 function, the total latency will be eleven cycles, where the most critical parts are bit rotations. We can calculate, on average, the total latency in Haswell of the T_1 function, with the following formula:

$$\text{Lat\_Haswell} = 6(SH) + 4/2(ADD) + 8/3(LOGIC).$$

$$\text{Lat\_Haswell} \approx 11\ cycles.$$

Analyzing the latency and the throughput in the Skylake processor for the T_1 function, we observe that some operations have the same latency, but a better throughput. The SH operation has a throughput of two and the operation ADD a throughput of three. Thus, the latency calculation for the Skylake processor can be expressed as:

$$\text{Lat\_Skylake} = 6/2(SH) + 4/3(ADD) + 8/3(LOGIC).$$

$$\text{Lat\_Skylake} \approx 8\ cycles.$$

Section IX shows how these optimizations improved the performance of SHA2 in Haswell and Skylake.

B. SHA3 optimization

The Secure Hash Algorithm-3 (SHA-3) is a family of functions that was standardized by NIST in 2015 [26]. This family consists of four cryptographic hash functions and two extendable-output functions (XOFs), called SHAKE128 and SHAKE256. The permutation function used in the SHA-3 family is KECCAK-p [1600,24] and is the one responsible for algorithm efficiency.

The permutation function KECCAK-p [1600,24] is composed of five steps that are processed 24 times. The steps are: the θ step, where is computed an XOR of each word of the state with the parity of the left column and the right column rotated one bit; the ρ step, where each word of the state is rotated a fixed amount of bits; the π step where the words of the state are permuted; the χ step, where is processed a non-linear function between the elements of the same row; the ι step, where is computed an XOR between the first element of the state with a constant value. The KECCAK-p [1600,24] function uses a state of 25 words of 64 bits. The use of AVX2 instructions allows to gather four words in the same register and to process four states at the same time. To map four states are required 25 variables of 256 bits; after the mapping, each one of the 25 variables will be composed by one word of each state.

To implement the θ step, we need only XORs and rotations; the AVX instructions `vpsllq` and `vpsrlq` can be used to emulate rotation instructions. The other four steps can be implemented in blocks of five words at the same time; it is important to process these words together to avoid a large number of memory accesses because the Intel architecture has 16 256-bit registers and this implementation uses 25 variables.

In the ρ step is required to rotate a different amount of bits in each word of the state. It is possible to process this step in parallel using the AVX2 instructions `vpsllvq` and `vpsrlvq` to emulate a variable rotate. The π step permutes the words of the state; as each word of each state was mapped in the same variable, the permutation just change the name of the variables, that in fact, no instruction is required.

The χ step is processed in parallel by using one XOR and the `vpandn` instruction and the ι step is just one XOR of the first word of the state with a constant. The complete code can be found in [27].

VIII. Optimizations in LMS and XMSS

The following software optimizations were applied to the standard proposal LMS and XMSS. We will show how these optimizations improved the algorithms of key generation, signature, and verification of these schemes. Each of these operations is based on hash functions. Thus, by optimizing the underlying hash functions, we speedup the execution of the signature operations of both schemes.

The optimized functions, based on hash algorithms, were:

- the keyed hash functions of LMS and XMSS;
- the function F of LMS and XMSS;
- the functions PRF and PRG of LMS and XMSS;
- the function H of XMSS.

A. Optimization of keyed hash functions

The keyed hashed functions of both schemes LMS and XMSS always work with message blocks of the same size. Then, in order to accelerate the computation of the keyed hash function, we made a specialized implementation based on the size of the message input to be processed and set the $block$ values and the pad values. Since the pad values are fixed, there is no need to calculate the pad each time the function is called. Figure 2 shows the optimization of the function F of the XMSS with fixed pad for SHA2-256.

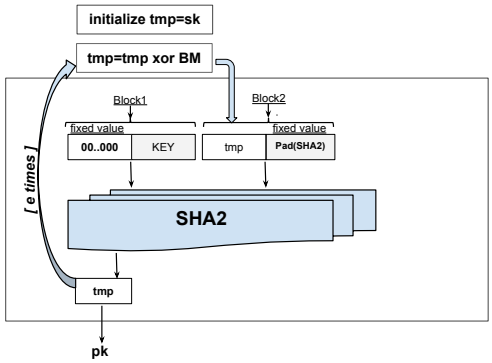


Fig. 2. Function F of the WOTS+.

In the specialized implementation of these functions, we have created an interface to receive and processes 32-bit message blocks on the SHA2-256 and 64 bits on SHA2-512/SHAKE128. The creation of an implementation of these functions with input, processing, and output with values of 32/64-bits, has significantly reduced processing time over generic functions that receive 8-bit characters because the conversion from 32/64 bit to 8 bits is time-consuming for the processor.

The hash function SHA2 processes blocks of 512 bits while the SHA3/SHAKE128 can handle up to 1344 bits of the message at the same time. An implementation of the function F of XMSS with SHAKE128 needs to process just a single block while in SHA2 must process two blocks to generate the hash value.

B. Optimization of the function F

The function F is used in the chaining function algorithm to generate the verification keys OTS. In the signature, the chaining function algorithm is also used to update the leaves in the authentication path. Thus, reducing the execution time of the function F reflected a significant improvement in the performance of both schemes LMS and XMSS.

Figure 3 shows our implementation of the function F with SHA2-512 which computes four instances in parallel, generating four public keys pk at the same time. We load four secret keys sk into the 256-bit vector registers, perform e iterations of the function F and then store four private keys pk on memory. We can compute the private keys pk in parallel because its generation is independent. Additionally, we store four instances of the pad value in a 256 register because these values will be used multiple times.

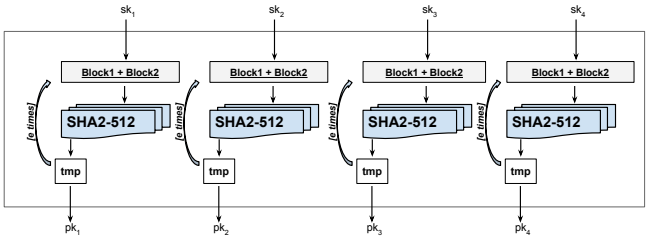


Fig. 3. Implementation of the chaining function F with SHA2-512.

.

The use of SIMD instructions helped to reduce the runtime of the function F, which are the computationally most expensive parts for key and signature generation.

1) The Function F in the signature and verification: For the generation of OTS keys, the optimization of the function F was simple, because in the generation of keys pk, the function F is performed the same number of times on all elements of the secret key sk. However, for OTS signature and verification generation, the application of the function F in each signature element depends on the message digest M. So we made a small change in the one-time signature and verification algorithms to apply the function F in parallel in elements of the signature.

We have added a sort algorithm before the function F. The vector msg, which contains how many times the function F will be performed, is sorted according to the number of applications of the function F. After sorting, we select the msg elements that have the same value to run the F in parallel. The function F is executed, and at the end, the signature elements are scaled according to the original order.

C. Optimizing the functions PRG and PRF

The PRG pseudo random generator generates the secret key elements sk using the PRF function. The secret keys are calculated as $sk[i] = PRF(S, toByte(i, 32))$, to $0 \leq i \leq len$. The string S is a secret value generated randomly and is used as a seed to generate all keys sk. The value i is concatenated with the value S to generate the values of sk. Since there is no dependence on the generation of the elements of sk, we could generate eight values of sk at the same time with SHA2-256 and four values of sk with SHAKE128, reducing the execution time of these functions.

The PRF function is used to generate the pseudorandom values. This generator was implemented in key generation and the signature of the LMS and the XMSS. The $PRF : Hash(toByte(3, n)||KEY||M)$ function receives the values of KEY and M as input. We created the function PRF_SIMD, which receives eight values of KEY and eight values of M in SHA2-256 and four values in SHAKE128. Then, processes these values in parallel and returns eight or four pseudorandom values.

D. Optimizing the implementation of the Ltree

The Ltree from XMSS [11] is used to generate the leaves of XMSS. In this section, we show an optimization in the generation of the Ltree for improving the performance of generating each leaf of the XMSS tree. This optimization was suggested in [13]. The function G is applied to each concatenation of children nodes to generate the parent node. Then, we modified the Ltree algorithm to perform eight evaluations of the function G at the same time. We generate eight internal nodes at the same time, from 16 children nodes, which are concatenated two by two. If the number of remaining nodes is not multiple of 16, we generate the next internal nodes one by one as the traditional way. If len is not a power of two, then there are not sufficient leaves to build a binary tree. Therefore, a node that has not a right sibling is lifted to a higher level of the Ltree until it becomes the right sibling of another node. Figure 4 shows the optimization performed in the generation of the Ltree for $w = 16$ and $l = 67$.

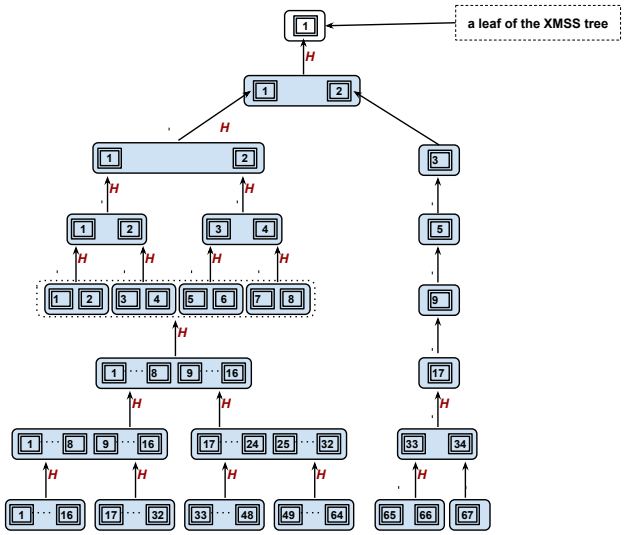


Fig. 4. Construction of the Ltree.

IX. Performance Results

This section shows the experimental results for LMS and XMSS of our implementation using AVX2 instructions. These results were obtained by running benchmarks on a Haswell processor Core i7-4770 at 3.4 GHz and a Skylake processor Core i7-6700K at 4.0 GHz. The Intel Turbo Boost and the Intel Hyper-Threading technologies were disabled to ensure the reproducibility of the results. Our implementation was written in C language and compiled using the GNU C Compiler v6.2.0. In our work, the runtimes for signing and verifying for $H > 20$, are calculated using the arithmetic average of the first one million signatures.

A. Scheme parameters

We have selected a set of parameters provided by the drafts LMS [14] and XMSS [16]. The parameters selected were: $w \in \mathbb{N}$, the Winternitz parameter; $h \in \mathbb{N}$, the total height of the tree; d, the number of layers; n, the output of the hash function.

The output of the chosen hash function influences system security. Considering classic computers, the parameter $n = 32$ provides a 256-bit security level and $n = 64$ provides 512-bit security level. Considering quantum computers, for 128-bit

security, we use the SHA2-256 and SHAKE256 functions in the LMS and the SHA2-256 and SHAKE-128 functions in the XMSS. For 256-bit security, we use the SHA2-512 and SHAKE256 functions in XMSS.

The value w influences the execution time and the size of the signature. Larger values for w imply larger execution times, but smaller signature sizes. The H and d values affect the signature size. The output n of the chosen hash function also influences the size of the public, private, and signature keys. We used $w = 2$ or $w = 4$ for LMS and $w = 16$ for XMSS. The maximum height of XMSS^{MT} was $H = 60$ and the maximum number of layers was $d = 12$ according to [16].

B. SHA2/SHA3 implementation results

In Table II, we show the performance of the implementation of the SHA2-256 and the SHAKE128 single-buffer (64-bit) and multi-buffer(256-bit) on Haswell and Skylake. The input sizes of these functions have been selected according to the size of the functions F and G of LMS and XMSS.

TABLE II. PERFORMANCE FIGURES OF SHA2-256 AND SHAKE128.

runtime(cycles/bytes)					
		Single Buffer		Multi Buffer	
hash	function	Haswell	Skylake	Haswell	Skylake
SHA2-256	F(104 bytes)	15.28	14.64	3.31	2.98
SHAKE128	F(104 bytes)	12.75	12.37	5.24	4.73
SHA2-256	G(133 bytes)	17.17	16.52	4.20	3.50
SHAKE128	G(133 bytes)	9.96	9.67	4.09	3.48

The function F processes message of 104 bytes in LMS and 96 bytes in XMSS. The G function processes message of 133 bytes in LMS and 128 bytes in XMSS. Then, the function F on SHA2-256 processes two data blocks (512 bits) and the functions G processes three data blocks (512 bits). In SHAKE128, the functions F and G processes a single block (1344 bits). Therefore, the implementation of F and G single-buffer functions with SHAKE128 has better results.

The computation of the functions F and G with SHA2-256 were faster than the versions using SHAKE128, for multi-buffer implementations. The speedup of the function F with SHAKE single-buffer is $1.2\times$ compared to SHA2-256 single-buffer implementation. SHA2-256 processes 8 independent hash values simultaneously and SHAKE128 processes only four independent hash values in parallel. Also, the speedup with SHA2 multi-buffer is approximate $4.6\times$, and with SHAKE multi-buffer is approximate $2.4\times$ compared to the single-buffer.

We also note in Table II that the function F with SHA2-256 presents a speedup of $4.6\times$ per hashing on Haswell and $4.9\times$ per hashing on Skylake. Performance on Skylake is better because of the computer architecture features presented in Section VI.

In the following sections, we will show that the performance obtained in the multi-buffer implementation of the hash functions impacts on the performance of the key generation, signature, and verification algorithms.

C. XMSS/LMS implementation results

In this section, we present the results of our XMSS/LMS single-buffer (64-bit) and multi-buffer (256-bit) implementation with the hash functions SHA2 and SHAKE.

In Table III, we compare our XMSS single-buffer (64-bit) implementation with the single-buffer implementation presented on the author's website [28]. The results were obtained on a Haswell processor. A speedup of $1.4\times$ is observed for key generation, signature, and verification of our implementation over the implementation [28]. This improvement was due to the specialized implementation of each function of XMSS.

TABLE III. PERFORMANCE FIGURES OF XMSS FROM [28] AND OUR XMSS IMPLEMENTATION ON HASWELL

runtime(ms) XMSS SHA2-256						
	Single Buffer [28]			Single Buffer(our)		
h	KeyGen	Sig	Ver	KeyGen	Sig	Ver
10	2241	9.78	1.2	1613	7.04	0.87
16	143329	16.31	1.22	103268	11.76	0.88
20	2286505	20.63	1.22	1652284	14.91	0.89

Table IV presents the results of the single-buffer (64-bit) and multi-buffer (256-bit) implementation of XMSS with SHA2 and SHAKE for different security levels. We compare our results using single-buffer and using a multi-buffer for $h = 20$.

We show that the speed up due to the multi-buffer optimization, for key generation, signing, and verification respectively, is: with SHA2-256 ranges from $4.4\times$, $4.2\times$ and $2.4\times$; with SHAKE128 ranges from $2.6\times$, $2.5\times$ and $2.0\times$; with SHA2-512 ranges from $2.4\times$, $2.4\times$ and $2.0\times$; and with SHAKE256 ranges from $3.3\times$, $3.3\times$ and $2.8\times$.

TABLE IV. PERFORMANCE FIGURES OF XMSS FOR PARAMETERS $h = 20$ AND $w = 16$ FOR DIFFERENT SECURITY LEVELS ON SKYLAKE

runtime(ms) XMSS							
	Function	Single Buffer			Multi Buffer		
security	F/G	KeyGen	Sig	Ver	KeyGen	Sig	Ver
128	SHA2-256	1369770	12.36	0.73	312702	2.92	0.30
128	SHAKE128	1056084	9.49	0.60	410818	3.74	0.30
256	SHA2-512	3537106	32.49	1.85	1452600	13.57	0.90
256	SHAKE256	5339130	49.12	2.77	1586750	14.70	0.99

For key generation, the performance is greater because the function F_SIMD executes the same amount of times in all elements of the secret key. However, in the WOTS signing and verification process, it was necessary to sort the elements of the signature before of the function F_SIMD, because the number of applications of the function depends on the bits of the message.

The runtimes with SHA2-512/SHAKE256 are larger than using SHA2-256/SHAKE128, but we get a higher level of security (256-bit security level). For 128-bit security level, the performance of the XMSS single-buffer with SHAKE128 is better than with SHA2-256. However, the multi-buffer version of SHA2-256 has better runtimes than the multi-buffer version of SHAKE128, due to the performance of these functions presented in Table II.

Table V represents the timing results of our software for the multi-buffer version of LMS with SHA2-256 and SHAKE256 at 128-bit security level. We observed that the acceleration obtained with SHA2-256 multi-buffer and $w = 4$ is $4.2\times$, $4.1\times$ and $2.0\times$ for key generation, signature, and verification respectively. The implementation with SHAKE256 multi-buffer and $w = 4$ ranges from $2.7\times$, $2.6\times$ and $1.8\times$ for key generation, signing, and verification.

A larger value of w results in shorter signatures but slower overall signing operations; it has little effect on security. For

TABLE V. PERFORMANCE FIGURES OF LMS 128-BIT SECURITY LEVEL FOR DIFFERENT VALUES OF w ON SKYLAKE

runtimes(ms) LMS								
			Single Buffer			Multi Buffer		
HASH	h	w	KeyGen	Sig	Ver	KeyGen	Sig	Ver
SHA2-256	20	2	225069	2.06	0.12	61987	0.56	0.06
SHA2-256	20	4	436627	3.96	0.23	103053	0.96	0.11
SHAKE256	20	2	186187	1.73	0.10	76870	0.71	0.06
SHAKE256	20	4	353786	3.20	0.18	130802	1.20	0.10

keys and signature generation, performance is higher for the same reasons as for XMSS implementation. As in XMSS, the LMS single-buffer with SHAKE256 is faster than SHA2-256, and the LMS multi-buffered with SHA2-256 is faster than with SHAKE256.

D. Hierarchical signatures scheme implementation results

In this section, we show the performance results of our software for both schemes HSS and XMSS^{MT} on the Skylake processor. We use the parameter $w = 4$ for LMS and $w = 16$ for XMSS, then the length of the signature OTS is $len = 67$ for both schemes. In Table VI, are given the runtimes of HSS multi-buffer using the hash functions SHA2-256 and SHAKE256 at 128-bit security level.

TABLE VI. PERFORMANCE FIGURES OF HSS MULTI-BUFFER FOR $w = 4$ AND DIFFERENT VALUES OF h AND d ON SKYLAKE

runtimes(ms) HSS multi-buffer					
HASH	h	d	KeyGen	Sig	Ver
SHA2-256	40	4	402	0.57	0.39
SHA2-256	40	8	27	0.32	0.73
SHA2-256	60	6	602	0.57	0.60
SHA2-256	60	12	40	0.32	1.11
SHAKE256	40	4	552	0.76	0.42
SHAKE256	40	8	36	0.41	0.79
SHAKE256	60	6	827	0.76	0.62
SHAKE256	60	12	55	0.41	1.19

Table VII presents the results of our implementation of XMSS^{MT} with SHA2-256 and SHAKE128 for the 128-bit security level.

TABLE VII. PERFORMANCE FIGURES OF XMSS^{MT} MULTI-BUFFER FOR $w = 16$ AND DIFFERENT VALUES OF h AND d ON SKYLAKE

runtimes(ms) XMSS^{MT} multi-buffer					
HASH	h	d	KeyGen	Sig	Ver
SHA2-256	40	4	1236	1.74	1.15
SHA2-256	40	8	83	0.96	2.18
SHA2-256	60	6	1883	1.75	1.76
SHA2-256	60	12	124	0.96	3.32
SHAKE128	40	4	1667	2.28	1.14
SHAKE128	40	8	111	1.23	2.23
SHAKE128	60	6	2498	2.28	1.67
SHAKE128	60	12	167	1.23	3.33

Notice that by increasing the number of layers the runtime is reduced for key generation and signing; however, the runtime for verification increases because the signatures of all layers must be checked. For subtrees that have the same height, the signature time remains constant. Then, increasing the tree height allows producing more signatures without impacting the performance of signing and verifying. Additionally, by increasing the number of layers the size of the secret key and signature is larger because they store information of each layer.

E. Analysis of results

In this section, we examine the results with AVX2 and compare the schemes LMS and XMSS.

Figure 5 shows the performance of XMSS/LMS multi-buffer $\times$ single-buffer with AVX2. The multi-buffer implementations with SHA2-256 have better performance because it allowed executing eight hashes at the same time whereas the SHAKE allowed to perform only four hashes in parallel.

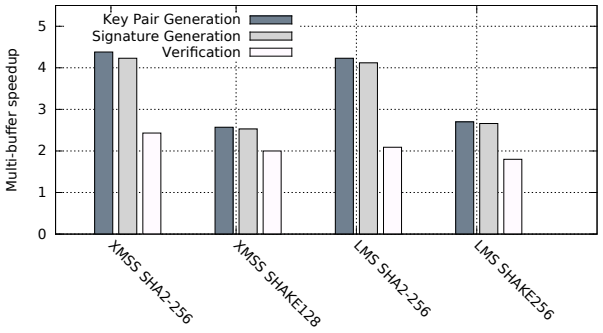


Fig. 5. Performance multi-buffer x single-buffer implementation.

Table VIII shows a summary of the code size, in C language, for the main functions of the schemes LMS and XMSS. We execute the GNU command nm in the Linux compiler, and it returned the size of the objects in a file. Note that the LMS has code size approximately $1.23\times$ greater than the XMSS code for the single-buffer implementation because the LMS uses two hash functions H_leaf and H_node for the generation of the leaves of the tree and XMSS uses only Ltree.

TABLE VIII. SIZE OF LMS AND XMSS CODES

	size (bytes)			
scheme	F	G	H_leaf/Ltree	H_no
LMS	9413	14168	15577	9552
XMSS	9664	13996	15819	

Table IX shows the keys length and runtimes of both schemes LMS and XMSS for a tree with height $h = 20$ and 128-bit security level ($n = 32$ bytes). If one uses $w = 4$ for LMS and $w = 16$ for XMSS, then the length of the signature OTS is $len = 67$ for both schemes.

TABLE IX. SIZES AND RUNTIMES OF THE LMS AND XMSS WITH SHA2-256 FOR 2^{20} SIGNATURES

		sizes (bytes)			runtimes(ms)		
draft	w	SK	PK	Sig	KeyGen	Sig	Ver
LMS	4	68	96	2820	103053	0.96	0.11
XMSS	16	132	64	2820	410818	3.74	0.30

Note that for the selected parameters, LMS secret key size is shorter than XMSS secret key. On the other hand, LMS public key size is larger than XMSS public key, and the signature key of the both schemes has the same size. Since LMS has fewer calls to the underlying hash function than XMSS, the implementation of LMS with SHA2-256 is approximate $2.7\times$ faster than the implementation of XMSS with SHA2-256. In addition, LMS does not use a Ltree tree and performs fewer operations on generating the internal nodes of the binary tree.

According to the results presented, the use of AVX2 contributes significantly to the implementation of the proposals standard for the Merkle scheme and its variants. For 128-bit security level, if the computer does not have instructions $AVX2$, then the implementation of the schemes LMS/XMSS with the hash function SHAKE128 single-buffer is a good option for presenting better runtimes. However, if these instructions are available on the computer, the implementation of the LMS/XMSS multi-buffer with SHA2-256 would be the best option.

Also, if the choice of the signature scheme is based on the runtimes, the LMS could be used because of better execution times. However, if there is a greater preoccupation with information security, XMSS would be a better option because the XMSS scheme provides strong security guarantees, XMSS is existentially unforgeable under adaptively chosen message attacks (EUCMA), it is forward security, and it is considered safe even when the collision resistance of the underlying hash function is broken.

X. Conclusion

The emerging transition to post-quantum cryptography requires digital signature schemes that are immune to quantum computers. Hash-based signatures schemes are promising candidates for replacing the current signatures schemes because they do not depend on arithmetic operations such as the problem of factorization of integers. These schemes are the object of current standardization efforts. Many improvements have already been made to the MSS making it feasible for many nowaday applications. However, some additional issues also appear as some signatures, storage resources, state management and slow generation of the key pair, leading to an important question: How can we apply the Merkle scheme in current applications?

New variants have emerged to improve the storage resource problem, such as the use of pseudo-random generators, reducing key size. The use of multi-trees, allowed to increase the number of signatures and reduce the time of generation of signature and verification keys.

In this work, we present an efficient software implementation of the Merkle scheme proposals (LMS and XMSS) using the set of vector instructions AVX2 on Intel processors. We show that our implementation presents significant improvements in the execution times of the key generation algorithms, signature, and verification of these standards. We have used several optimization techniques for increasing the performance in the software of both schemes. Our results show the feasibility of using these post-quantum schemes in practical applications.

Acknowledgment

The authors would like to thank the anonymous referees for their valuable comments and suggestions to improve the quality of this paper. The second author was partially supported by a research productivity scholarship from CNPq Brazil.

References

[1] D. Johnson, A. Menezes, and S. Vanstone, "The elliptic curve digital signature algorithm (ecdsa)," *International Journal of Information Security*, vol. 1, no. 1, pp. 36–63, 2001. [Online]. Available: http://dx.doi.org/10.1007/s102070100002

[2] 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, Feb. 1978. [Online]. Available: http://doi.acm.org/10.1145/359340.359342

[3] N. FIPS, "186 digital signature standard," 1994.

[4] P. W. Shor, "Algorithms for quantum computation: Discrete logarithms and factoring," in *Foundations of Computer Science, 1994 Proceedings., 35th Annual Symposium on*. IEEE, 1994, pp. 124–134.

[5] L. Lamport, "Constructing digital signatures from a one-way function," Technical Report CSL-98, SRI International Palo Alto, Tech. Rep., 1979.

[6] R. C. Merkle, "A certified digital signature," in *Conference on the Theory and Application of Cryptology*. Springer, 1989, pp. 218–238.

[7] D. McGrew, P. Kampanakis, S. Fluhrer, S.-L. Gazdag, D. Butin, and J. Buchmann, "State management for hash-based signatures," in *Security Standardisation Research*. Springer, 2016, pp. 244–260.

[8] NIST. (2016) Post-quantum cryptography standardization. [Online]. Available: http://csrc.nist.gov/groups/ST/post-quantum-crypto/index.html

[9] J. Buchmann, L. C. C. García, E. Dahmen, M. Döring, and E. Klintsevich, "Cmss–an improved merkle signature scheme," in *International Conference on Cryptology in India*. Springer, 2006, pp. 349–363.

[10] J. Buchmann, E. Dahmen, E. Klintsevich, K. Okeya, and C. Vuillaume, "Merkle signatures with virtually unlimited signature capacity," in *Applied Cryptography and Network Security*. Springer, 2007, pp. 31–45.

[11] J. Buchmann, E. Dahmen, and A. Hülsing, "Xmss-a practical forward secure signature scheme based on minimal security assumptions," in *International Workshop on Post-Quantum Cryptography*. Springer, 2011, pp. 117–129.

[12] A. Hülsing, L. Rausch, and J. Buchmann, "Optimal parameters for xmss-mt," in *International Conference on Availability, Reliability, and Security*. Springer, 2013, pp. 194–208.

[13] D. J. Bernstein, D. Hopwood, A. Hülsing, T. Lange, R. Niederhagen, L. Papachristodoulou, M. Schneider, P. Schwabe, and Z. Wilcox-O'Hearn, "Sphincs: practical stateless hash-based signatures," in *Annual International Conference on the Theory and Applications of Cryptographic Techniques*. Springer, 2015, pp. 368–397.

[14] D. McGrew, M. Curcio, and S. Fluhrer, "Hash-based signatures," 2016, work in progress, draft-mcgrew-hash-sigs-05.

[15] R. C. Merkle, "Secrecy, authentication, and public key systems," 1979, ph.D. thesis, Electrical Engineering, Stanford.

[16] A. Hülsing, D. Butin, S. . Gazdag, and A. Mohaisen, "Xmss: Extended hash-based signatures," 2016, work in progress, Crypto Forum Research Group, Internet Draft, draft-xmss-06.

[17] A. Hülsing, "W-ots+-shorter signatures for hash-based signature schemes." *Africacrypt*, vol. 7918, pp. 173–188, 2013.

[18] J. Buchmann, E. Dahmen, and M. Schneider, "Merkle tree traversal revisited," in *International Workshop on Post-Quantum Cryptography*. Springer, 2008, pp. 63–78.

[19] J. Katz, "Analysis of a proposed hash-based signature standard," 2015. [Online]. Available: http://www.cs.umd.edu/ jkatz/papers/HashBasedSigs.pdf

[20] A. Hülsing, J. Rijneveld, and F. Song, "Mitigating multi-target attacks in hash-based signatures," in *Public-Key Cryptography–PKC 2016*. Springer, 2016, pp. 387–416.

[21] INTEL. (2016) Intel 64 and ia-32 architectures optimization reference manual. [Online]. Available: www.intel.com/content/dam/www/public/us/en/documents/manuals/64-ia-32-architectures-optimization-manual.pdf.

[22] M. J. Flynn, "Some computer organizations and their effectiveness," *IEEE Transactions on Computers*, vol. C-21, no. 9, pp. 948–960, 1972.

[23] S. Gulley and V. Gopal, "Haswell cryptographic performance," 2013.

[24] A. Fog, "Instruction tables: Lists of instruction latencies, throughputs and micro-operation breakdowns for intel, amd and via cpus," *Copenhagen University College of Engineering*, 2016.

[25] A. Faz-Hernández, R. Cabral, D. F. Aranha, and J. López, "Implementação Eficiente e Segura de Algoritmos Criptográficos," in *Simpósio Brasileiro em Segurança da Informação e de Sistemas Computacionais - Minicursos*, vol. XV. Sociedade Brasileira de Computação, 2015, pp. 93–140.

[26] National Institute of Standards and Technology, *FIPS PUB 202 SHA-3 Standard: Permutation-Based Hash and Extendable-Output Functions*. Gaithersburg, MD, USA: National Institute for Standards and Technology, Aug. 2015. [Online]. Available: http://nvlpubs.nist.gov/nistpubs/FIPS/NIST.FIPS.202.pdf

[27] R. Cabral. (2017) Implementation of the sha-3 family using avx/avx2 instructions. [Online]. Available: https://github.com/rbCabral/SHA-3

[28] A. Hülsing and J. Rijneveld. (2016) Implementation of xmss and xmssmt as specified in draft-huelsing-cfrg-hash-sig-xmss-06. [Online]. Available: https://huelsing.wordpress.com/code

JSEA: A Program Comprehension Tool Adopting LDA-based Topic Modeling

Tianxia Wang
School of Software Engineering
Tongji University
China

Yan Liu
School of Software Engineering
Tongji University
China

***Abstract*—Understanding a large number of source code is a big challenge for software development teams in software maintenance process. Using topic models is a promising way to automatically discover feature and structure from textual software assets, and thus support developers comprehending programs on software maintenance. To explore the application of applying topic modeling to software engineering practice, we proposed JSEA (Java Software Engineers Assistant), an interactive program comprehension tool adopting LDA-based topic modeling, to support developers during performing software maintenance tasks. JSEA utilizes essential information automatically generated from Java source code to establish a project overview and to bring search capability for software engineers. The results of our preliminary experimentation suggest the practicality of JSEA.**

Keywords*—Java program comprehension; Topic models; Interactive tool

I. Introduction

Before performing software maintenance and extension tasks, developers must understand what a program does and how it does [1]. Program comprehension is the activity of understanding how a software system or a part of it works [2]. According to Corbi [3], program comprehension accounts for more than half of the software maintenance time. Researchers and practitioners developed tools that can assist developers in program comprehension, such as JRipples [4], Codecrawler [5], and SonarQube [6]. However, tools applying topic models to program comprehension has not been given as much attention.

With the development of Information Retrieval, topic models appeared as an effective way of extracting semantic information. Topic models are probabilistic models for uncovering the underlying semantic structure of a document collection based on a hierarchical Bayesian analysis of the original texts [7]. Researchers apply topic models to mine source code and get linguistic topics automatically. These topics extract semantic information from programs and tend to correspond to features implemented by the software [8]. Using these topics appropriately, developers should be able to understand programs more effectively.

Researchers have used topic modeling to support varied software engineering tasks, including traceability link recovery [9], concept/feature location [10], source code metrics [11], and many other tasks [12], [13], however, the application of using topic modeling to support interactive program comprehension has not got as much attention. Also, there are gaps between knowing and doing when applying topic modeling to software engineering practice. For instance, feature location is the act of identifying the set of source code fragments in a software system that implement a particular concept, but the boundary lines and definitions of features in programs are vague [14], so it is hard to directly locate features via automatic topic modeling. We need to explore topic modeling and its application on interactive program comprehension further.

In this paper, we explored the source code preprocessing procedure of topic modeling based on the characteristic of source code and introduced JSEA (Java Software Engineers Assistant), an interactive program comprehension tool adopting LDA, which provides a project overview page and a search model. JSEA aims at helping developers learn unfamiliar source code in a faster manner to carry on software maintenance and extension tasks. Considering different programming languages have different characters, in this paper, we focus on Java language, but the concept can be reused for other languages.

This paper makes the following contributions: (1) Based on the characteristic of source code, explore the procedure of source code preprocessing to support topic modeling. (2) Design and develop an interactive program comprehension tool, which extracts semantic information from source code in a useful manner, to support developers during software maintenance. (3) Verify the practicality of JSEA through a preliminary evaluation.

This paper is organized as follows. In the next section, we introduce background. Then, section III detail the preprocess procedure used for JSEA. In section IV and section V, we introduce the interactive program comprehension tool based on LDA and its strategies for the number of topics. In section VI, we provide our evaluation results. Finally, the section **??** concludes and introduces future works.

II. Background

A. Program Comprehension

Some influential theories about program comprehension were proposed in the past, including top-down [15], [16], bottom-up [17], and a combination of the two [18], [19]. Brooks [15] describes top-down theories of program comprehension as hypothesis-driven. When a program is comprehended, the knowledge are organized into distinct domains which bridge between the original problem and the final program. The program comprehension process is reconstructing

knowledge about these domains and the relationship among them. In button-up model, Shneiderman et al. [17] theorizes that programmers first read source code line-by-line and the program comprehension is accomplished by a hierarchical chunking process that organizes several statements into a functional unit. Then, these units can be organized into still higher level units which convey the overall operation of the program. The integrated metamodel of program comprehension have also been proposed [18], [19], in which programmers switch flexibly from top-down to bottom-up comprehension strategies depending on the situation.

Furthermore, Corritore and Wiedenbeck [1] reported that the object-oriented programmers tend to use a strongly top-down approach to program understanding during the early phase of familiarization with the program but use an increasingly bottom-up approach during the subsequent maintenance tasks. Koenemann and Robertson [20] argued that comprehension activities are mostly top-down in a larger program. The tool, JSEA, is aimed at assist developers in comprehending unfamiliar programs, especially large scale programs, so the top-down model for JSEA is suitable. JSEA is also an interactive tool, which facilitates the information exchange between users and the system and then foster program comprehension during software maintenance.

B. Topic Modeling

Topic models were originally developed as a means of automatically indexing, searching, clustering, and structuring large corpora of unstructured and unlabeled documents. Using topic models, topics are extracted from documents and are used to represent the corpora. A topic is a collection of terms that co-occur frequently in the documents of the corpus, so the documents can be clustered by topics and the entire corpus can be indexed and organized in terms of this discovered semantic structure [7], [13]. Latent Dirichlet Allocation (LDA) is a popular probabilistic topic model [21].

C. LDA

Latent Dirichlet Allocation (LDA) is a popular probabilistic topic model. It models each document as a multi-membership mixture of K corpus-wide topics, and each topic as a multi-membership mixture of the terms in the corpus vocabulary. This means that there is a set of topics that describe the entire corpus, each document can contain more than one of these topics, and each term in the entire repository can be contained in more than one of these topics. Therefore, LDA is able to discover a set of ideas or themes that well describe the entire corpus. Blei reported LDA in detail [21]. In this paper, LDA-based topic modeling is used to mine source code and then support developers.

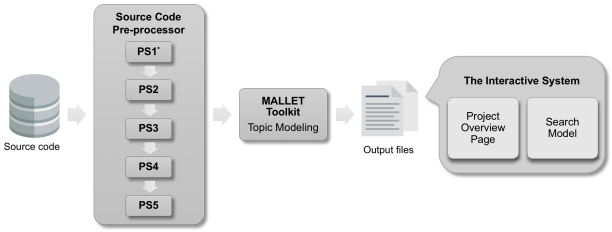
III. Preprocessing Procedure

Before topic modeling, several preprocess steps are generally taken to reduce noise and improve the modeling results [13]. Compared to natural language text, Hindle et al. [22] reported that text extracted from source code is much more repetitive and predictable. Based on the characteristic of source code and prior researches [23], [24], the preprocess procedure was customized through experimentation and brought the following five preprocess steps:

- **Remove programming language information:** Characters related to the syntax of the programming language (e.g., &&, $\rightarrow$) are removed; programming language keywords (e.g., *if*, *while*) are removed. In this paper, a Java language keywords list was customized referring to Oracle official documentation [25]. At first, the customized Java language keywords list was same as the list that Oracle provided, then this list was used to topic modeling and got a result. Secondly, useless Java language keywords were selected from the result and were added into the list. Then, this process was repeated until the result did not have useless Java language keywords. See Table I for a full list. In future, the Java language keywords list can be generated by machine learning instead of manual processing.
- **Split words:** Identifier names are split into multiple parts based on common naming conventions, such as camel case (oneTwo), underscores (one_two), and dot separators (one.two). According to the research of Grant et al. [24], in this paper, we tried to put original identifier names into the source of topic modeling too. For instance, an identifier name called "addMenu", was first split into "add" and "menu", and then "addmenu", "add" and "menu" were put into the source of topic modeling. However, the results were worse. We suppose the reason is the original words like "addmenu" are not in natural language, so adding original words into assets influences the results for topic modeling. Therefore, in this step, identifier names were just split and the original identifier names were abandoned.
- **Remove stop words:** Common English-language stopwords (e.g., *the*, *it*, *on*) are removed. In this paper, the common English-language stopwords list of MALLET LDA toolkit [26] is used. In future, the common English-language stopwords list can be configurable.
- **Remove copyright information:** For open source code, copyright information almost exists in each document. Topic modeling with such information will generate a topic containing copyright information like *author*, *copyright*, *license*. This kind of information is easy to obtain without the help of topic modeling, and it is useless for developers to get a general understanding of programs, so copyright information need to be removed.
- **Remove nondescript information:** In source code, libraries can be imported and packages can be declared, thus topics extracted from source code can contain some common used library names and package names. However, some library names and package names is useless for developers to acquire a general understanding of the program and even might confuse developers. For example, {*set drawing button* ***org*** *init layout panel components pane variables*} is a topic extracted from JHotDraw [27], and "org" is useless in

TABLE I. THE CUSTOMIZED JAVA LANGUAGE KEYWORDS

abstract array arg assert boolean break byte catch case char class code continue default ddouble do don double else enum error exception exist exists extends false file final finally float for id if implementation implemented implements import instanceof int integer interface interfaces invoke invokes java lead long main method methodname methods native new null object objects overrides package packages param parameters precison println private protected public return returned returns short static string strictfp super switch synchronized system this throw throws transient true try version void volatile while



*PS: Preprocess step (see Sect. III)

Fig. 1. JSEA Overview

this topic. We define this kind of library names and package names as nondescript information and remove them in preprocessing.

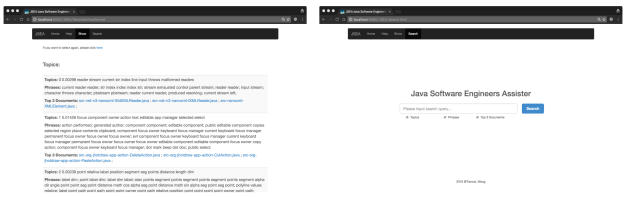
IV. JSEA

JSEA (Java Software Engineers Assistant) is a web application implemented in Java language, setting up in Tomcat server. It infuses topic modeling into program comprehension in an interactive manner, aiming at supporting developers during software maintenance. All the data, results and source code of JSEA is available in Github[1].

A. Design

Fig. 1 provides an overview of JSEA. At the first stage, the Source Code Pre-processor tackles the source code with comments of a Java program, using four preprocess steps mentioned in Sect. III. Then, the data is sent to MALLET LDA Toolkit [26]. MALLET is a Java-based package for statistical natural language processing, document classification, clustering, topic modeling, information extraction, and other machine learning applications to text. Topic modeling based on LDA is completed in this stage. As for the parameters of the model, it affects the practicability of JSEA in a large extent, so the values of parameters were tuned, especially the number of topics (See Section V), through experiments. By default, the parameter settings are as follows: the maximum iteration (–num-iterations) is 1000, the number of most probable words (–num-top-words) is 10, the number of iterations between reestimating dirichlet hyperparameters (–optimize-interval) is 10 and the initial topic model parameters are the default values in the MALLET LDA toolkit [26]. We also allow users to set the value of parameters as their demands. Next, JSEA processes topic modeling results to acquire semantic and structure information about the program as follows:

- **Extracting Topics:** Topic modeling directly generates topics, so JSEA just need to store these topics with index for the following procedure.



(a) The Project Overview Page (b) JSEA-Search

Fig. 2. The Screenshots of JSEA

TABLE II. AN EXAMPLE OF A TOPIC IN THE PROJECT OVERVIEW PAGE

Top Words: 78* 0.00985† *action menu add set bar put item tool actions open*
Phrases: action action; menu item; menu bar; menu add; tool bar; action put action; put action; menu open; action add; menu menu labels
Top 3 Documents:
src-org-jhotdraw-app-DefaultOSXApplication.java-src.txt;
src-org-jhotdraw-app-DefaultMDIApplication.java-src.txt;
src-org-jhotdraw-app-DefaultSDIApplication.java-src.txt

*The index of topics
†The value of Dirichlet parameters

- **Applying Descriptive Phrases to Topics:** When using MALLET LDA toolkit to topic modeling, the LDA toolkit automatically summarizes related phrases for each topic. For example, {*stroke color width fill grow basic font line shape factor*} is a topic extracted from JHotDraw [27], and the LDA toolkit adds phrases like "fill color", "stroke width" and "basic stroke" to this topic. JSEA links these descriptive phrase with their related topics, in order to increase the readability of topics.
- **Applying Documents to Topics:** The probabilistic topic distributions of documents are generated by topic modeling. Using the distributions, JSEA assigns top related documents to each topic.

Finally, JSEA utilizes these information to establish a system allowing developers to explore the programs in an interactive manner, including:

- **A Project Overview Page:** The Project Overview Page is designed for helping developers get an overview of Java projects in a short time. It shows all topics and some related information about the program. Which kind of information will be shown depends on which style is selected by users. JSEA provides four styles: (1) Topics; (2) Topics and Phrases; (3) Topics, Phrases and Top 3 Documents (Recommended); (4) Topics, Phrases and More Top Documents. Note that "More Top Documents" means top 100 related documents for each topic. For example, if users select the third style, the Project Overview Page will be like Fig. 2(a) and Table II gives an example of a topic in Project Overview Page. Users can select a suitable style according to their demands and habits.
- **JSEA-Search:** JSEA-Search (Fig. 2(b)) is a search model, which is designed for satisfying developer's immediate information need. Through JSEA-Search, developers can obtain information semantically and

[1]https://github.com/jseaTool/JSEA

structurally related to the search query. These information can include topics, descriptive phrases, top related documents for each topic and an access to related source code, all of which can help developers quickly determine where to start during performing software maintenance or extension tasks. Users determine searching which kind of information through the checkboxes below the search bar, and all checkboxes are selected as default.

Developers can select suitable sub systems based on their needs. JSEA-Search is more practical than the Project Overview Page, because the latter is helpful just for developers who want to have a general overview of projects in a short time, but MAT-Search can effectively assist developers when they face maintenance or extension tasks.

B. Configuration

JSEA is configured in the following steps: (1) Deploy Tomcat Server; (2) Configure the *web.xml* of JSEA; (3) Configure MALLET LDA toolkit.

V. The Number of Topics

The number of topics greatly affects topic modeling results and thus the results of the study [13]. Too many topics assigns related words to different topics but brings more meaningful information, while too few topics rarely brings related words into different topics but leads to results containing less meaningful information. Therefore, the number of topics need to be estimated.

A few articles proposed approaches determining the number of topics [28], [29], but they were task-specific. Our paper explore the application of topic modeling in a generic perspective other than a task-driven style, so we need to be groping for other estimating approach. Chen et al. [13] reported that many articles choose an optimal value of K (the number of topics) by testing a range of K values and evaluating each in some way. We decided to use the same strategy.

As for how to evaluate topic modeling results, a Naive Criterion was proposed: (1) Label each topics. The categories include *functionality*, *Java library topic*, *design*, *repetitive*, *multiple meaning* and *useless*. The first three are positive, while the last three are negative; (2) Calculate the percentage of each category for each K value; (3) Compare and analysis all results generated by varied K. The basic idea is that the result is better when the percentage of positive label is higher.

We recommend the users of JSEA estimating the number of topics by testing a range of K values and evaluating each using Naive Criterion. For instance, we used JHotDraw [27], a Java GUI framework, as our learning object. Referring to Grant and Cordy [29], where they think 100 to 200 is the best area for the number of topics of JHotDraw, we tested the number of topics ranging from 50 to 250 in 10 increments and evaluated each result using our Naive Criterion. We found that 80 is the most optimum value for the number of topics of JHotDraw.

TABLE III. Task Assignment

Systems	FLT_1 *	FLT_2	RT_1 †	RT_2
JHotDraw	IDE and JSEA	only IDE	IDE and JSEA	only IDE
MALLET	IDE and JSEA	only IDE	IDE and JSEA	only IDE

*FLT: feature location task
†RT: reuse task

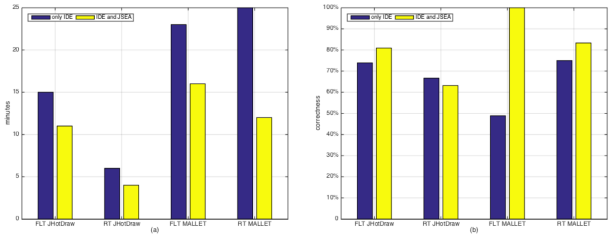


Fig. 3. (a) The time spent on each task; (b) The correctness of each task. FTL means feature location task, and RT means reuse task.

VI. Evaluation

We conducted several experiments on two open-source systems, in order to evaluate the utility of JSEA on software maintenance. The overall metric of our experiments was: Does JSEA save effort for developers when they perform tasks?

One of the open-source systems is JHotDraw version 7.0.6, a Java GUI framework for technical and structured graphics. Another is MALLET version 2.0.8, a Java-based package for statistical natural language processing, document classification, clustering, topic modeling, information extraction, and other machine learning applications to text. JHotDraw version 7.0.6 has 54KLOC (thousands of lines of code), while MALLET version 2.0.8 has 114KLOC, which is 2.11 times larger than JHotDraw version 7.0.6. Note that, 170 is the most optimum value for the number of topics of MALLET, using the same strategy of determining the number of topics with JHotDraw (See Sect. V).

To evaluate the practicality of JSEA, we recruited an experienced Java developer as our volunteer to perform two kinds of software maintenance tasks, and one of authors, who is greatly familiar with both systems and JSEA, was asked to evaluate the correctness of results. As for the two kinds of software maintenance tasks, one is feature location task (FLT), and another is reuse task (RT). For each kind of task, we conducted two tasks for each systems. One is with the help of commonly used IDE (IntelliJ IDEA), another is with the help of IDE and JSEA. Table III summarizes the task assignment. The volunteer performed tasks in Table III from left to right and from JHotDraw to MALLET.

During experimentation, we recorded the time spent on each task and analyzed the correctness of the results. Fig. 3(a) shows the time spent on each task. Fig. 3(b) shows the correctness of the results. The correctness is calculated in the following way: We define the files that considered relevant to each task by the author as *total files*, and define the intersection of *total files* and the files that considered relevant to each task by the volunteer as *correct files*. The correctness is the percentage that dividing the number of *correct files* by the number of *total files*.

In Fig. 3, the blue one means only using IDE, while the

yellow one means using IDE and JSEA. We can see from the left bar chart that the yellow bar is lower than the blue bar for each pair of bars, especially the third pair and the fourth pair. It means JSEA can save time for developers when they perform feature location tasks and reuse tasks. Besides, JSEA is more suitable to support developers on larger scale programs. In the right bar chart, for each pair of bars, the yellow bar is higher than the blue bar or is similar with the blue bar. It means using JSEA do not affect the correctness of the results. In conclusion, JSEA can help developers comprehend programs and then perform tasks faster.

VII. Conclusion and Future Work

In this paper, we explored the source code preprocessing procedure of topic modeling based on the characteristic of source code and developed JSEA (Java Software Engineers Assistant), an interactive tool adopting LDA-based topic modeling, to support developers during software maintenance. JSEA utilizes essential information automatically generated from Java source code to establish a project overview and to bring search capability for developers. The preliminary experimentation indicates that JSEA can effectively help developers comprehend programs, and assist them in maintaining software projects or developing new features with less time.

In future, we plan to integrate topic modeling of MALLET toolkit into JSEA, and calculate the correlation between topics and search queries. Other interesting direction for future work would be to combine the code search system into JSEA, assisting software engineers with more powerful functionality. Besides, we plan to recruit developers with different professional levels as volunteers and repeat the experimentation, in order to verify the practicality of JSEA more effectively.

References

[1] C. L. Corritore and S. Wiedenbeck, "An exploratory study of program comprehension strategies of procedural and object-oriented programmers," *International Journal of Human-Computer Studies*, vol. 54, no. 1, pp. 1–23, 2001.

[2] W. Maalej, R. Tiarks, T. Roehm, and R. Koschke, "On the comprehension of program comprehension," *ACM Transactions on Software Engineering and Methodology (TOSEM)*, vol. 23, no. 4, p. 31, 2014.

[3] T. A. Corbi, "Program understanding: Challenge for the 1990s," *IBM Systems Journal*, vol. 28, no. 2, pp. 294–306, 1989.

[4] J. Buckner, J. Buchta, M. Petrenko, and V. Rajlich, "Jripples: A tool for program comprehension during incremental change," in *Program Comprehension, 2005. IWPC 2005. Proceedings. 13th International Workshop on*. IEEE, 2005, pp. 149–152.

[5] M. Lanza, S. Ducasse, H. Gall, and M. Pinzger, "Codecrawler-an information visualization tool for program comprehension," in *Software Engineering, 2005. ICSE 2005. Proceedings. 27th International Conference on*. IEEE, 2005, pp. 672–673.

[6] S. S.A, "SonarQube," https://www.sonarqube.org/, 2017, accessed 27 Feb 2017.

[7] D. M. Blei and J. D. Lafferty, "Topic models," *Text mining: classification, clustering, and applications*, vol. 10, no. 71, p. 34, 2009.

[8] P. F. Baldi, C. V. Lopes, E. J. Linstead, and S. K. Bajracharya, "A theory of aspects as latent topics," in *ACM Sigplan Notices*, vol. 43, no. 10. ACM, 2008, pp. 543–562.

[9] S. K. Lukins, N. A. Kraft, and L. H. Etzkorn, "Bug localization using latent dirichlet allocation," *Information and Software Technology*, vol. 52, no. 9, pp. 972–990, 2010.

[10] E. Linstead, P. Rigor, S. Bajracharya, C. Lopes, and P. Baldi, "Mining concepts from code with probabilistic topic models," in *Proceedings of the twenty-second IEEE/ACM international conference on Automated software engineering*. ACM, 2007, pp. 461–464.

[11] Y. Liu, D. Poshyvanyk, R. Ferenc, T. Gyimóthy, and N. Chrisochoides, "Modeling class cohesion as mixtures of latent topics," in *Software Maintenance, 2009. ICSM 2009. IEEE International Conference on*. IEEE, 2009, pp. 233–242.

[12] D. Andrzejewski, A. Mulhern, B. Liblit, and X. Zhu, "Statistical debugging using latent topic models," in *European conference on machine learning*. Springer, 2007, pp. 6–17.

[13] T.-H. Chen, S. W. Thomas, and A. E. Hassan, "A survey on the use of topic models when mining software repositories," *Empirical Software Engineering*, pp. 1–77, 2015.

[14] S. Grant, J. R. Cordy, and D. B. Skillicorn, "Reverse engineering co-maintenance relationships using conceptual analysis of source code," in *2011 18th Working Conference on Reverse Engineering*. IEEE, 2011, pp. 87–91.

[15] R. Brooks, "Towards a theory of the comprehension of computer programs," *International Journal of Man-Machine Studies*, vol. 18, no. 6, pp. 543–554, 1983.

[16] V. Rajlich and N. Wilde, "The role of concepts in program comprehension," in *Program Comprehension, 2002. Proceedings. 10th International Workshop on*. IEEE, 2002, pp. 271–278.

[17] B. Shneiderman and R. Mayer, "Syntactic/semantic interactions in programmer behavior: A model and experimental results," *International Journal of Parallel Programming*, vol. 8, no. 3, pp. 219–238, 1979.

[18] S. Letovsky, "Cognitive processes in program comprehension," *Journal of Systems and Software*, vol. 7, no. 4, pp. 325–339, 1987.

[19] M.-A. Storey, "Theories, methods and tools in program comprehension: Past, present and future," in *Program Comprehension, 2005. IWPC 2005. Proceedings. 13th International Workshop on*. IEEE, 2005, pp. 181–191.

[20] J. Koenemann and S. P. Robertson, "Expert problem solving strategies for program comprehension," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1991, pp. 125–130.

[21] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[22] A. Hindle, E. T. Barr, Z. Su, M. Gabel, and P. Devanbu, "On the naturalness of software," in *Software Engineering (ICSE), 2012 34th International Conference on*. IEEE, 2012, pp. 837–847.

[23] S. Thomas, "Mining unstructured software repositories using ir models," 2012.

[24] S. Grant, J. R. Cordy, and D. B. Skillicorn, "Using heuristics to estimate an appropriate number of latent topics in source code analysis," *Science of Computer Programming*, vol. 78, no. 9, pp. 1663–1678, 2013.

[25] Oracle, "Java Language Keywords," http://docs.oracle.com/javase/tutorial/java/nutsandbolts/_keywords.html, 2015, accessed 2 Aug 2016.

[26] A. K. McCallum, "MALLET: A Machine Learning for Language Toolkit - Topic Modeling," http://mallet.cs.umass.edu/topics.php, 2002, accessed 21 May 2016.

[27] W. Randelshofer, "JHotDraw as Open-Source Project," http://www.jhotdraw.org/, 2007, accessed 27 Feb 2017.

[28] B. Dit, A. Panichella, E. Moritz, R. Oliveto, M. Di Penta, D. Poshyvanyk, and A. De Lucia, "Configuring topic models for software engineering tasks in tracelab," in *2013 7th International Workshop on Traceability in Emerging Forms of Software Engineering (TEFSE)*. IEEE, 2013, pp. 105–109.

[29] S. Grant and J. R. Cordy, "Estimating the optimal number of latent concepts in source code analysis," in *Source Code Analysis and Manipulation (SCAM), 2010 10th IEEE Working Conference on*. IEEE, 2010, pp. 65–74.

Missing Data Imputation using Genetic Algorithm for Supervised Learning

Waseem Shahzad
National University of Computer and Emerging Sciences, Islamabad, Pakistan

Qamar Rehman
National University of Computer and Emerging Sciences, Islamabad, Pakistan

Ejaz Ahmed
National University of Computer and Emerging Sciences, Islamabad, Pakistan

***Abstract*—Data is an important asset for any organization to successfully run its business. When we collect data, it contains data with low qualities such as noise, incomplete, missing values etc. If the quality of data is low then mining results of any data mining algorithm will also below. In this paper, we propose a technique to deal with missing values. Genetic algorithm (GA) is used for the estimation of missing values in datasets. GA is introduced to generate optimal sets of missing values and information gain (IG) is used as the fitness function to measure the performance of an individual solution. Our goal is to impute missing values in a dataset for better classification results. This technique works even better when there is a higher rate of missing values or incomplete information along with a greater number of distinct values in attributes/features having missing values. We compare our proposed technique with single imputation techniques and multiple imputations (MI) statistically based approaches on various benchmark classification techniques on different performance measures. We show that our proposed methods outperform when compare with another state of the art missing data imputation techniques.**

***Keywords*—genetic algorithm; information gain; missing data; supervised learning**

I. Introduction

Data is available in every sphere of life which is collected and used for various purposes. Processing and analysis of the collected data after being processed usually provides useful insights and knowledge about the system which has produced such data. The field of data mining basically deals with mining useful information from raw data instead of using all the data that also has some unimportant information. Data mining is a collection of techniques used for extracting or mining of previously unknown, useful and understandable patterns from large databases. Data mining integrates techniques from multiple disciplines such as database technology, machine learning, statistics, pattern recognition, neural networks, and image processing and data visualization. There is always a requirement for efficient and scalable data mining algorithms and it is a subject of ongoing research [1].

The process of data mining is to extract information from data. The first step is to extract data from the database and then perform preprocessing steps on it. Data mining techniques are used to extract data patterns. Evaluation and presentation mean to represent the knowledge which is understandable to users. The result is the empowerment of users with knowledge.

There are different data mining techniques including supervised classification, association rules mining or market basket analysis, unsupervised clustering, web data mining, and regression. One important technique of data mining is the classification of data. The objective of classification is to build one or more models based on the training data, which can correctly predict the class of test objects. There are several problems with a large scale of domains which can be cast as classification problems [1]. The classification has several important applications in our lives [2-5]. Examples include customer behavior prediction, portfolio risk management, identifying suspects, medical applications, sports and fraud detection etc. This research deals mainly with the data preprocessing evaluated on the basis of classification technique of data mining.

One of the challenging problems is to transform huge amount of data into an accessible and actionable knowledge. This knowledge is utilized by domain experts for decision making. Therefore, the core focus is on the knowledge discovery process in the databases (KDD). KDD is defined as a non-trivial process of identification and extraction of implicitly, previously unknown, and potentially useful information from the data [1].

The collected data may contain several states of the art deficiencies such as missing values, non-discredited data, inconsistent, incomplete and noise etc. If data is not of high quality it may hinder the discovery of useful patterns later in the process. The main purpose of the preprocessing step is to enhance the quality of data used in the experiment. All the data mining techniques are applicable once the data has been preprocessed and the objective of preprocessing is simple. Data collected from the real world is dirty and needs to be cleaned. The word dirty in data perspective means state of the art deficiencies described earlier. There can be various reasons due to which these issues arise, overcoming these problems is done by using KDD process, and there are different techniques that are proposed by various researchers which we will describe later in this paper.

In this paper, we address an important area of data preprocessing which is missing values imputation. Missing values in a dataset mislead the learning model. We have proposed a new approach based on GA and IG to impute the missing values. The proposed technique has been evaluated on different classification methods. The proposed technique has a higher accuracy rate and is well suited for large dimensional search spaces with a higher rate of missing values.

The rest of the paper is organized as follows. Section 2

describes the background of missing values, section 3 presents different classification algorithms, section 4 provides a detail description of proposed technique, section 5 presents experimentation results and finally section 6 concludes proposed technique and gives some future directions.

II. Background of Missing Data Imputation

A. Importance of Complete Datae

Basically, in data mining, the focus is on extracting useful information from a large amount of data that is collected from various sources and to take decisions using such data. Decisions are made on the basis of science, business and economic approaches on data available. As an example, sales and other information allow business class and investors to evaluate and make critical decisions regarding their investments with their future outcomes, whereas advances in research are based on the discovery of knowledge from various experiments and measured parameters.

During fault detection and identification, it is observed that most data is corrupt or incomplete. Predictive models that take observed data as an input are used for many decision-making processes, such models do not tolerate any incompleteness in data provided for prediction and as a result, such models are normally broken down. In many applications, simply ignoring the incomplete record is not an option. Most decision-making tools such as the commonly used neural networks, support vector machines, and many other computational intelligence techniques cannot be used for decision making if data is not complete. This is mainly due to the fact that ignorance can lead to biased results in statistical modeling or even damages in machine control [6]. For this reason, it is often essential to making the decision-based approach on available data [7].

The challenge missing data pose to the decision-making process is more evident in on-line applications where data have to be used almost instantly after being obtained. The biggest challenge is that the standard computational intelligence techniques are not able to process input data with missing values and hence, cannot perform classication or regression. Some of the reasons for missing data are sensor failures, omitted entries in databases and on- response to questions in questionnaires. There have been many techniques reported in the literature to estimate the missing data for some applications [7]. There are several reasons why data might be missing, and missing data may follow an observable pattern. Exploring the pattern is important and may lead to the possibility of identifying cases and variables that eect the missing data [7, 8]. A proper estimation method can be derived by identifying the variables that predict the pattern.

B. Missing Data Mechanisms

Missing data randomness is divided into three classes [9] such as missing completely at random missing at random, not missing at random [5] and missing data handling techniques (Ignoring data).

To discard data with missing values two core methods are used. One is called complete case analysis. It is available in every one of statistical packages and is the default method in many programs. The other method is discarding instances or attributes called listwise deletion. In this method, the level of missingness is determined on each instance and attribute and deletes the instances or attributes with high extents of missing data. Prior to deleting any attribute, it is vital to evaluate its connotation to the investigation. The methods, complete case analysis and discarding are executed only if missing data is missing completely at random. The missing data that are not missing completely at random contain non-random elements that may prejudice the results [9]. The deletion can bring in significant bias into the experimentation. In addition, the reduced sample size can significantly hamper the analysis. The thumb rule for deletion instances is, if attributes have more than 5

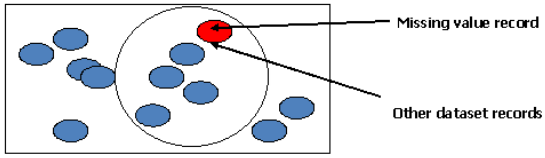


Fig. 1. KNN search space

1) Mean-fill approach: Most common technique in missing data imputation is finding the estimates of the values and then these estimates are replaced with the missing entries, the focus of our work is related to the estimation of values and its comparison to proposed technique. These estimates include statistical calculation i.e., means, zero filling, min replacement and max replacement.

These estimation techniques are used in datasets with missing values as observed values results are observed in the form of classifiers accuracy and other output measures like precision, recall, f-measure and Area under ROC. The main reason of calculating other results is just because if the classifier does not satisfy the accuracy reported. Then these measures can also be observed in the case of finding a better result.

Mean-fill approach is one of the most common statistical estimation approaches that is actively used as filling up missing values attributes of data with missing values, which is provided by various open source data mining toolboxes or packages. Also in latest researchers are using comparison technique and their majority cases research provides promising results. But it is observed that for data with a large amount of missing information this approach do not work very well.

Mean of the attribute values (in case, of numeric values, for discrete values MODE is taken) set is taken and all the missing values are replaced by the mean value in that particular attribute, similar is the case for all attributes for any dataset. Min fill approach, Max fill approach, Zero fill approach and K-Nearest Neighbor approach [10] are most common approaches being used.

K-Nearest Neighbors are determined on the bases of some kind of distance between points. It has the biggest disadvantage since it looks for the most similar instances, the whole dataset should be searched. On the other hand, how to select the value k and the measure of similar will impact the result greatly.

2) Multiple imputations (MI): It is one of the most attractive methods for general purpose handling of missing data in

the multivariate analysis. Rubin [9] described MI as a three step process, imputation, analysis, and pooling.

The most challenging step is imputation, that is, the construction of the m-completed datasets. This step accounts for the process that causes the creation of the missing data. First, sets of plausible values for missing values are created using an appropriate model chosen, reflects the uncertainty due to the missing data. Each of these sets of plausible values is used to fill-in the missing values and creates a completed dataset. Typical problems are:

- Missingness could be related to the value of information (e.g., people with higher incomes tend to skip income questions more often).
- Missing entries can appear anywhere in the data.
- The method used in the imputation step must foresee the intended complete-data analysis.

The repeated ANALYSIS step on the imputed data is actually somewhat simpler than the same analysis without imputation because there is no need to bother with the missing data. Each of these datasets can be analyzed using complete data methods.

The POOLING step consists of computing the mean over the m repeated analysis, its variance, and its confidence interval or P value. Results are combined finally. In general, these computations are relatively simple.

There are various ways to generate imputations. The implementation program for MI of continuous multivariate data (NORM) is available in [12]. However, it is not necessarily true that any particular method will perform better for any particular empirical study. It is well known that methods for handling nonignorable data require the analyst to make assumptions about the model of missingness [11]. Recent overviews of NMAR modeling are given in [13, 14, 15]. Selection and Pattern mixture models are used for NMAR data Models need more statistical formulas to impute the data. If the chosen model is incorrect then MNAR model may perform even less well than standard MAR methods [9]. Different types of weighting methods are also used for non-ignorable missing data. Even though many methods are available, they could not be used by researchers due to lack of familiarity and computational challenges and researchers often opt for ad-hoc approaches that may do more harm [7].

3) Auto-associative Neural Networks: An auto-associative referred to as autoencoder neural network is a specific neural network, trained to recall its inputs [19]. Given a set of inputs, the network predicts these inputs as outputs and thus has the same number of output nodes as there are inputs. However, the hidden layer is characterized by a bottleneck, with fewer hidden nodes than output nodes.

The smaller hidden layer projects the inputs onto a smaller space, extracting linear and non-linear interrelationships such as covariance and correlation, from the input space and also removes redundant information [19]. This means that they can be used in applications to recall the inputs and missing data estimation applications.

III. Classification

Data mining learning models are categorized into two, the one in which class to which training sample is known while there is a learning stage; it is called labeled training data. The predictive models are built on the basis of supervised learning data, whereas unlabeled data is used to test the model. One example is the classification method in which class labels are known. Other is unsupervised learning method where the class label for the training data is unknown. Here the training data is grouped according to their similarities, clustering is the example of unsupervised learning where data is unlabeled.

A fundamental aim of this research work in the field of classification is to perform preprocessing on data available and to make clean data available to the classifiers highly accurate models from the available data that can be learned. Other objective includes verification of correctness of proposed technique on the basis of classification results. Decision Tree (C4.5), PART, NB-Tree and RIPPER are the most common classifiers used in the field of machine learning and these are also used in this research [23, 24, 25, 26].

IV. Proposed Technique

We have used GA with IG for imputation of missing values. Following subsections will describe the proposed technique.

A. Genetic Algorithm

GAs are basically evolutionary ideas of natural selection and genetics [16, 17]. GAs are adaptive heuristic search algorithm. Inspired by Darwins theory of evolution survival of the fittest, it is common in nature that in a competition where individuals are looking for resources fittest individuals dominate over weaker ones. Evolutionary computing today holds GAs as one of the important parts. Among random search methods employed to solve optimization problems, GAs represent an intelligent structure which is easy to implement.

For any particular problem GAs works for solving it is by mimicking processes nature use, like selection, crossover, mutation and acceptance, to evolve a good solution for that problem.

1) Operators of GA: GAs use genetic operators to maintain genetic diversity. It is important that genetic diversity or variation is maintained for the process of evolution. Inspired by natural genetic structure, genetic operators are the same. Following are operators used in genetic algorithms.

1) *Reproduction/ Selection:* Usually, the first operator applied on population is a reproduction, from the population the chromosomes are selected to be parents for the crossover step and producing offsprings.
   According to Darwins theory survival of fittest, the best ones should survive and create new offsprings. Reproduction operator is also called selection operator because it is basically extraction of genes subset from existing population based on some quality criteria or definition. The fitness function is the quality measurement that can be performed to select best genes subset, as every gene contains some meaning.
2) *Crossover/ Recombination* This genetic operator is called crossover because it mates (combines) two

parents (chromosomes) to produce a new offspring (chromosome).
Most commonly used methods for the selection of parents to crossover are:

- Roulette wheel selection.
- Rank selection.
- Boltzmann selection.
- Steady state selection.
- Tournament selection.

The idea behind crossover is that after mating any chromosomes (parents) that are selected based on some function, offsprings (chromosomes) will be fitter as they are derived as a result of best characteristics of their parents. According to user-defined crossover probability, it takes place during evolution stage.

3) *Mutation:* During the evolution stage mutation occurs where the user defines mutation probability, this probability is usually set to a fairly low value, like 0.01 is a good first choice. Mutation is the genetic operator used to maintain genetic diversity from one generation of a population of chromosomes to the next generation.

B. Proposed Technique

This section provides detail of the proposed technique along with fitness function used

1) General Description: GA is used for missing data imputation, the importance of missing data imputation varies from problem to problem, and we use this technique to clean the dirty data for classification problem. The missing values are imputed in the datasets using GA and GA is run for each attribute which is treated as a chromosome. We divided these chromosomes into frames for further accurate measures; these frames are explained by the example in the following section. Frames are dependent upon the no of classes in the dataset. i.e., there is n number of class labels in a dataset.

The flow chart describes the working of proposed technique as shown in figure 2. Using attribute instances first we create an initial solution of population size defined in parameter section. Evaluate the fitness of each solution. Check termination criteria for a maximum number of generations. For generation number 1 initial size of the new population is 0. Select individuals randomly from the population according to tournament size for selection using tournament selection. Select genetic operator to be applied to the selected individuals probabilistically. Perform crossover or mutation on the selected individual's bases on the probability of selection for crossover or mutation. The resultant of the genetic operator is inserted in the new population. Check for population size on every iteration, if population size is equal to maximum population size then start a new generation and check for termination criteria else continue to select new individuals from the current population. When the population size is reached maximum new generation become started, if the current population has the fitness of individual then how previous populations best fitted then we keep that individual from the previous population in current population (Elitism= keep best).

To illustrate how GA works in improving data quality by imputing missing values based on estimation, the following is

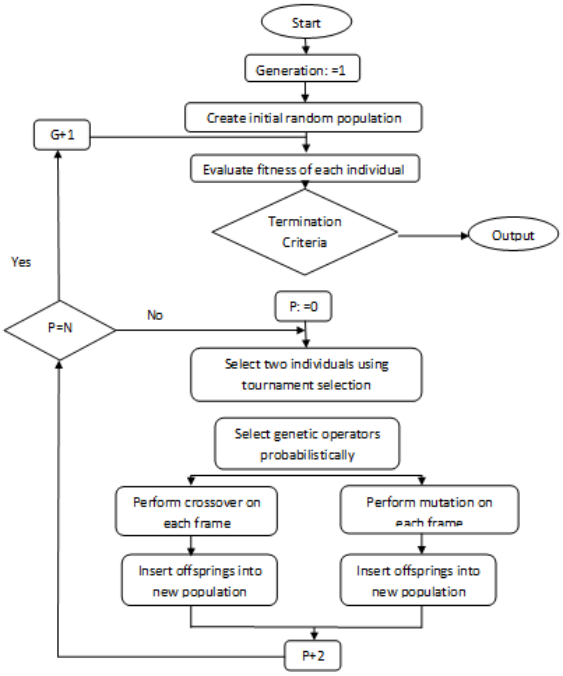


Fig. 2. Flow Chart of proposed method

Attr#1	Attr#2	Attr#3	Attr#4	Attr#5	Class
5	-1	3	5	9	1
4	43	1	1	-1	1
5	58	4	5	3	1
-1	28	1	1	3	0
5	74	1	5	-1	1
-1	59	-1	4	5	1
4	-1	2	1	-1	0
5	51	4	-1	-1	1
4	39	1	1	-1	0
11	44	4	4	3	0
4	53	-1	3	4	0
2	-1	1	3	0	1
9	22	3	0	6	0
2	54	4	-1	-1	1
2	45	1	6	-1	0
1	65	-1	4	-1	1

Fig. 3. Sample Datasets

a simple example that describes the technique. This sample dataset is given in figure 3 is a part of dataset named Memographicmasses. Remember -1 represents missing values.

A chromosome split into n number of frames (sub-chromosome). N is the number of classes in the dataset. Each frame initialized independently to another frame, within restricted range that it must contain values obtained by the attribute to a specific class, in the first generation. Merging all frames into one makes the valid structure of a chromosome in population.

Each frame is treated as full fledged independent chromosome at the time of applying genetic operators on it. One point crossover used on each frame so n number of cross points are used for every chromosome. Mutation operator mutates randomly n number of genes depending on the probability of mutation criteria. Each gene belongs to a specific frame so, during mutation of a gene, gene value is replaced by a specific set of domain values of class from the dataset.

1) *Structure of Chromosome:* The data illustrated above belongs to two class problem so N = 2. The number of the frame will be two in each chromosome as shown

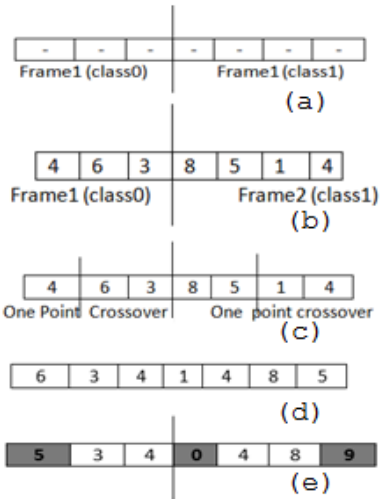


Fig. 4. Chromosome Structure

in figure 4 (a-Chromosome Structure). The size of the frame depends on missing values related to specific class as explained in the previous section.
According to figure 3 let us take the case of attribute # 5, the range of selecting values for frame1 will be min gene max. According to dataset values of the gene are assigned between 3 gene 6. Similarly for frame 2 the values between 0 gene 9. As shown in figure 4 (b- Values of genes assigned).

2) Operate on genetic operators:

1) *Crossover:* One point crossover is performed. Figure 4 (c- Crossover Performed) shows the crossover points on the chromosome. As a result of crossover, offspring is created that is shown in figure 7. The chromosome in figure 4 (d- Result of crossover) shows the result of crossover operation performed on it in the previous step.
2) *Mutation:* In mutation gene values mutate according to the domain of each frame defined according to the range of distinct values that are available in the particular data attribute, here figure 8 illustrates the outcome of the mutation operation performed on chromosome shown in figure 4 (e- Mutation performed). After mutation is performed the fitness of the resultant chromosome is calculated if it is greater than the previous data fitness the values are saved and next iteration takes its place, until the termination criteria are met that can be the end of condition or some value which achieved terminates GA.

C. Fitness function

In the proposed GA for missing data imputation, we are using IG as Fitness function that is based on the entropy of each attribute regarding its class label in the given dataset. Remember that when we calculate the fitness of an attribute then the whole attribute is used for the calculation of fitness after imputing missing values.

Following is the brief description of Fitness function used in the proposed work.

$$\mathbf{H(X)} = -\sum_i \boldsymbol{P(x_i) log_2}\ (\boldsymbol{(P(x_i))} \qquad \textbf{(1)}$$

$$\mathbf{H\ (X|Y)} = -\sum_j \boldsymbol{P(y_j)} \sum_i \boldsymbol{P(x_i|y_j) log_2 (P(x_i|y_j))} \qquad \textbf{(2)}$$

$$\mathbf{IG\ (X|Y)} = \mathbf{H(X)} - \mathbf{H\ (X|Y)} \qquad \textbf{(3)}$$

IG is a correlation-based measure. It is based on an information theoretical concept of entropy i.e. a measure of the uncertainty of a random variable. Following is the equation of entropy of X eq(1).

The entropy of X when the value of another random variable Y is known, following is the conditional entropy eq(2).

In the above equation (2), P(x_i) is the prior probability for all the values of X. P(x_i—y_j) is the posterior probability of X after the values of Y are known. The amount of which the entropy of X decreases, it depicts the decrease in uncertainty level. This is achieved through the additional information regarding X provided by Y. This measure is called IG. Following is the formula for information gain eq(3).

V. Experimentation and Analysis

A. Experimentation Framework

The population size is defined as 500, for 100 generations and tournament size is kept 6. These parameters setting have been chosen after performing several experiments. All the other parameters used are defined in the following table 1. Experimentation has been performed with different combinations of these parameters and best values are kept same for all the experimentation as shown in table 1.

In the experimentation, worth of a missing data imputation through GA is evaluated on 5 key measures, i.e. predictive accuracy, along with precision, recall, f-measure, and ROC.

Following table 2 elaborates about the datasets used in the experimentation. All the datasets used are publicly available and taken from UCI repository [31]. We have used standard implementation of MI which is available as NORM [12], and classifiers like NB tree, PART, JRIP, j48, NAVE bases and K-Nearest Implementation of these algorithms is provided by data mining software Weka [30]. All the algorithms are used with their default values and no tweaking is done over the methods. Since these algorithms are implemented by their authors, therefore, it is assumed that parameter setting is already incorporated.

Following table 2 describes datasets used for experimentation along with total number of features, the number of instances and percentage of missing values in these datasets.

The techniques that are used for comparison with the proposed method are Multiple Imputation, Mean filling, Min filling, Max fills and Zero fill. Table 3 shows a comparison of classification accuracies and their standard deviations after being imputed by various techniques including proposed technique of GA fill.

TABLE I. PARAMETERS USED IN GA

Parameters	Values
Population Size	500
Generation	100
Chromosome size	Missing values in attributes
Fitness Function	IG
Selection	Tournament
Tour Size	6
Crossover	One point crossover
Probability of crossover	0.8
Probability of mutation	0.2
Probability of gene mutate	0.5
Elitism	Keep-best
Runs	20

TABLE II. DATASETS USED FOR EXPERIMENTATION

Dataset	Total Attributes	No of Instances	%age missing
Labor	17	57	33.82
Echocardiogram	12	132	5.16
Cylinderbands	34	540	5.26
Memographicmasses	6	961	4
Colic-Horse	22	368	23.90

B. Comparison with other techniques

For comparison, four single imputation techniques have been adopted; filling missing values using mean, min, max and zero by replacing all missing data by 0, Multiple Imputation (MI) of missing data is also used for comparison. The results are compared for NB-Tree, JRIP, PART, NAVE Bayes, IBK (Lazy) and j48 (C4.5) classifiers. For performing experimentation with these classifiers we used Weka machine learning tool [30]. We used supervised discretization filter of Weka-3.4 machine learning tool [30] to discretize continuous attributes as a preprocessing step. The GA has seven user-defined parameters. The values of these parameters are given in table 1. The predictive accuracies of the compared algorithms are shown in tables 3 and 4. Ten-fold cross validation is used to obtain the results. The Bold value represents the highest accuracy achieved.

From tables 3 and 4, it is observed that missing data imputation using GA clearly out marks by 70% of the datasets than other estimation and predictive model techniques. These predictive accuracies show worth of the proposed approach. A genetic algorithm is an evolutionary algorithm and has much diversity when imputing missing values.

Our method performs better in three datasets for NAVE Bayes algorithm, similarly, for K-Nearest neighbors classifier, it achieves better accuracies on three datasets and better in four datasets for the j-48 classifier.

Table 5 shows results of precision results of proposed algorithm and single, and MI technique is presented. Different classifiers are used for classification with 10-fold cross validation. The Bold value represents the highest accuracy achieved.

In the above-mentioned table, our method performs comparable and/or better in 70% of the datasets. It can be observed that our proposed method achieves better/ comparable classification precisions as compared to single and MI techniques n most of the cases. It is observed that missing data imputation using GA clearly out marks/ comparable with other estimation and predictive model techniques.

In table 6, recall measures of proposed algorithm and single and MI technique is shown.

TABLE III. COMPARISON OF GA WITH DIFFERENT TECHNIQUES BASED ON CLASSIFIERS ACCURACIES ALONG WITH STANDARD DEVIATIONS

Datasets	NB-Tree						
	With missing value	GA filled	MI filled	Mean filled	Min filled	Max filled	Zero filled
Labor	80.80±15.66	**95.43±8.80**	87.53±13.38	85.47±13.99	81.07±16.70	80.73±14.86	90.07±13.20
Echocardiogram	87.29±7.70	**90.90±7.89**	89.17±7.01	86.10±8.00	88.62±7.85	86.58±8.33	88.10±7.63
Cylinderbands	67.41±6.90	68.78±6.14	66.24±6.26	68.63±6.23	**69.57±6.14**	68.39±6.10	68.37±5.76
Memographicmasses	81.85±3.04	**83.08±3.19**	82.74±3.39	81.98±3.30	82.24±3.28	82.84±3.34	82.35±3.29
Horsecolic	83.53±6.51	83.85±5.64	83.02±6.97	83.74±6.45	**83.98±5.56**	83.96±6.17	83.36±6.28
				PART			
Labor	76.77±16.28	**93.30±9.58**	87.70±15.29	85.57±15.37	81.53±15.01	79.13±12.09	83.43±12.72
Echocardiogram	88.00±7.54	**89.59±8.38**	88.09±7.85	86.92±8.71	85.55±7.86	85.82±7.97	86.35±8.47
Cylinderbands	69.56±5.67	69.93±6.57	68.81±6.11	72.39±6.76	72.52±5.66	70.89±6.34	**73.09±5.25**
Memographicmasses	81.53±3.36	**82.23±3.31**	82.43±3.53	80.56±3.83	81.46±3.66	81.15±3.32	81.46±3.34
Horsecolic	84.77±6.40	**85.34±5.70**	79.76±6.58	78.83±6.16	79.36±7.24	78.33±6.93	78.56±6.43
				JRIP			
Labor	**88.40±15.86**	86.97±11.1	85.59±8.90	81.60±15.95	80.33±16.73	84.00±14.21	79.03±14.86
Echocardiogram	83.86±8.67	**86.26±8.46**	83.58±3.50	84.23±8.29	84.47±9.21	83.67±8.88	84.03±8.47
Cylinderbands	67.02±6.29	68.37±5.99	67.59±6.51	69.67±5.90	69.91±6.34	69.98±6.11	**70.63±6.00**
Memographicmasses	82.38±3.22	**83.81±3.37**	82.99±5.64	81.89±3.41	82.60±3.36	82.55±3.43	82.79±3.31
Horsecolic	83.28±6.55	**85.15±5.28**	84.73±13.98	82.01±6.47	84.18±5.98	81.79±5.99	83.06±6.18

TABLE IV. COMPARISON OF GA WITH DIFFERENT TECHNIQUES BASED ON CLASSIFIERS ACCURACIES ALONG WITH STANDARD DEVIATIONS

Datasets	NAIVE Bayes						
	With missing value	GA filled	MI filled	Mean filled	Min filled	Max filled	Zero filled
Labor	89.67±13.38	**98.20±5.45**	91.93±11.83	92.07±11.72	90.53±10.43	87.97±13.56	65.60±18.24
Echocardiogram	89.59±7.03	**92.18±6.64**	89.05±7.24	85.19±8.62	88.92±7.85	86.56±8.01	74.64±9.77
Cylinderbands	65.67±6.35	66.24±5.91	65.44±5.00	63.72±6.93	67.09±5.75	**68.74±4.54**	67.63±4.50
Memographicmasses	**82.40±3.18**	79.09±3.49	79.16±4.17	77.26±4.19	81.12±3.24	81.29±3.32	80.84±3.39
Horsecolic	79.73±6.71	**85.83±4.67**	82.60±5.76	80.73±7.04	77.69±6.57	79.41±6.53	76.67±6.92
				K-Nearest			
Labor	82.33±16.57	90.10±10.74	**98.93±4.88**	92.57±10.30	85.87±13.38	82.83±16.68	75.63±17.15
Echocardiogra	84.03±7.83	**86.14±7.92**	86.04±7.93	85.20±8.48	85.97±7.70	82.34±9.00	84.37±8.25
Cylinderbands	68.22±5.57	70.04±5.84	68.35±5.97	70.85±5.75	75.09±5.84	75.07±5.29	**77.48±5.82**
Memographicmasses	74.69±3.78	**76.57±3.74**	74.64±3.69	75.10±3.76	75.34±4.07	73.01±3.90	75.39±4.02
Horsecolic	77.25±5.98	**81.64±5.90**	79.64±6.44	77.77±5.56	77.59±6.14	76.34±7.64	73.46±5.77
				J-48			
Labor	71.67±15.38	**89.20±11.50**	84.60±13.59	83.33±13.83	86.83±14.58	81.40±14.10	87.80±14.6
Echocardiogra	85.77±7.17	**89.31±7.09**	86.81±7.66	85.40±8.28	86.53±8.37	83.70±7.74	86.14±8.14
Cylinderbands	68.52±6.10	70.13±5.83	68.37±6.34	71.96±5.85	71.76±6.10	72.33±5.95	**73.43±6.07**
Memographicmasses	81.38±3.14	**83.00±3.03**	82.36±3.42	81.10±3.24	81.85±3.31	81.13±3.33	82.08±3.40
Horsecolic	85.15±5.78	**86.36±5.30**	84.54±6.22	83.94±6.17	82.80±6.22	83.85±5.75	82.74±5.86

In the below-mentioned table, our method performs comparable or better in most of the datasets.

In table 7, F-Measure of proposed algorithm and single, and MI technique is presented. The proposed approach also has better F-measure values in most of the datasets.

In table 8, AREA under ROC of proposed algorithm, single and MI techniques are presented. The AREA under ROC approach is high on most of the datasets.

These experimentation results of different data sets are evaluated on different benchmarks evaluation methods. This has shown the worth of proposed approach when compared with well-known missing data imputation algorithms. These results indicated that GA is a suitable method for the imputation of missing values.

TABLE V. COMPARISON OF GA WITH DIFFERENT TECHNIQUES BASED ON CLASSIFIERS PRECISION RESULTS

Datasets	NB-Tree						
	With missing value	GA filled	MI filled	Mean filled	Min filled	Max filled	Zero filled
Labor	0.88	**0.98**	0.94	0.90	0.86	0.87	0.93
Echocardiogram	0.62	**0.81**	0.73	0.63	0.69	0.61	0.65
Cylinderbands	0.60	0.66	0.60	0.68	0.68	0.68	**0.74**
Memographicmasses	0.81	0.80	0.82	0.80	0.81	**0.84**	0.81
Horsecolic	0.84	**0.89**	0.87	0.85	0.86	0.85	0.85
PART							
Labor	0.86	0.93	**0.94**	0.89	0.88	0.87	0.89
Echocardiogram	0.68	**0.74**	0.64	0.67	0.61	0.60	0.63
Cylinderbands	0.66	0.66	0.64	0.69	0.69	0.67	**0.70**
Memographicmasses	0.81	0.81	**0.83**	0.81	0.82	0.82	0.82
Horsecolic	0.86	**0.87**	0.85	0.83	0.84	0.83	0.83
JRIP							
Labor	0.91	**0.93**	0.92	0.91	0.84	0.91	0.86
Echocardiogram	0.55	**0.63**	0.64	0.58	0.58	0.58	0.55
Cylinderbands	0.62	0.64	0.63	0.67	0.67	0.67	**0.68**
Memographicmasses	0.82	0.79	**0.84**	0.81	0.83	0.83	0.83
Horsecolic	0.87	0.87	**0.89**	0.85	0.86	0.85	0.85
NAIVE Bayes							
Labor	0.93	**0.98**	0.94	0.91	0.93	0.92	0.83
Echocardiogram	0.69	**0.76**	0.87	0.60	0.68	0.62	0.14
Cylinderbands	0.63	0.66	0.69	0.59	0.67	**0.73**	0.57
Memographicmasses	**0.78**	0.73	0.75	0.72	0.78	0.77	0.78
Horsecolic	0.84	**0.89**	0.86	0.84	0.85	0.82	0.84
K-Nearest							
Labor	0.96	**0.96**	0.99	0.93	0.89	0.89	0.89
Echocardiogram	0.56	0.62	0.59	0.57	**0.64**	0.51	0.59
Cylinderbands	0.66	0.69	0.64	0.69	0.74	0.74	**0.76**
Memographicmasses	0.72	**0.73**	0.72	0.73	0.73	0.70	0.73
Horsecolic	0.78	**0.85**	0.85	0.83	0.82	0.81	0.80
J-48							
Labor	0.79	0.88	0.91	0.88	**0.89**	0.88	0.89
Echocardiogram	0.63	**0.70**	0.65	0.61	0.65	0.57	0.61
Cylinderbands	0.66	0.68	0.65	0.70	0.70	0.71	**0.72**
Memographicmasses	0.81	0.81	0.82	0.82	0.83	0.82	**0.84**
Horsecolic	0.85	**0.87**	0.89	0.87	0.86	0.87	0.86

TABLE VI. COMPARISON OF GA WITH DIFFERENT TECHNIQUES BASED ON CLASSIFIERS RECALL MEASURES

Datasets	NB-Tree						
	With missing value	GA filled	MI filled	Mean filled	Min filled	Max filled	Zero filled
Labor	0.84	**0.96**	0.87	0.90	0.88	0.85	0.94
Echocardiogram	0.79	**0.92**	0.81	0.75	0.87	0.72	.86
cylinderbands	**0.66**	0.56	0.62	0.53	0.56	0.53	0.46
Memographicmasses	0.80	**0.85**	0.82	0.82	0.81	0.79	0.82
Horsecolic	**0.91**	0.88	0.89	0.90	0.90	0.91	0.90
PART							
Labor	0.81	**0.98**	0.89	0.93	0.86	0.84	0.89
Echocardiogram	0.68	**0.73**	0.64	0.64	0.60	0.61	0.63
cylinderbands	0.59	0.60	0.61	**0.64**	**0.64**	0.63	0.64
Memographicmasses	0.80	0.81	0.79	0.76	**0.82**	0.77	0.79
Horsecolic	**0.92**	0.89	0.85	0.84	0.84	0.82	0.83
JRIP							
Labor	0.90	**0.92**	0.86	0.87	0.89	0.86	0.85
Echocardiogram	0.81	**0.90**	0.80	0.82	0.85	0.81	0.83
cylinderbands	0.58	0.57	0.59	0.57	0.59	0.59	**0.60**
Memographicmasses	0.80	**0.84**	0.80	0.80	0.83	0.79	0.80
Horsecolic	0.88	**0.90**	0.89	0.88	0.89	0.87	0.89
NAIVE Bayes							
Labor	0.92	**1.00**	0.95	0.98	0.95	0.91	0.58
Echocardiogram	0.87	**0.91**	0.87	0.78	0.90	0.79	0.15
cylinderbands	0.48	0.42	0.33	**0.51**	0.45	0.33	0.30
Memographicmasses	**0.86**	0.76	0.84	0.83	0.78	0.86	0.83
Horsecolic	0.85	**0.89**	0.86	0.86	0.79	0.87	0.79
K-Nearest							
Labor	0.75	0.90	**0.99**	0.98	0.92	0.85	0.74
Echocardiogram	0.57	**0.72**	0.59	0.59	0.62	0.64	0.57
cylinderbands	0.52	0.54	0.58	0.58	0.64	0.65	**0.70**
Memographicmasses	0.74	**0.76**	0.75	0.75	0.73	0.73	0.75
Horsecolic	**0.89**	0.87	0.85	0.82	0.83	0.83	0.78
J-48							
Labor	0.81	**1.00**	0.88	0.89	0.94	0.85	0.96
Echocardiogram	0.73	**0.83**	0.65	0.68	0.68	0.62	0.66
cylinderbands	0.55	0.57	0.55	0.59	0.59	0.58	**0.62**
Memographicmasses	0.79	**0.83**	0.79	0.77	0.83	0.77	0.77
Horsecolic	**0.93**	0.92	0.89	0.88	0.87	0.88	0.88

TABLE VII. COMPARISON OF GA WITH DIFFERENT TECHNIQUES BASED ON CLASSIFIERS F-MEASURES

Datasets	NB-Tree						
	With missing value	GA filled	MI filled	Mean filled	Min filled	Max filled	Zero filled
Labor	0.84	**0.96**	0.89	0.89	0.85	0.84	0.92
Echocardiogram	0.67	**0.79**	0.71	0.64	0.73	0.63	0.71
Cylinderbands	**0.62**	0.60	0.60	0.58	0.60	0.57	0.53
Memographicmasses	0.80	**0.82**	0.81	0.81	0.81	0.81	0.81
Horsecolic	0.87	0.87	0.87	0.87	**0.88**	0.88	0.87
PART							
Labor	0.81	**0.95**	0.90	0.89	0.85	0.83	0.87
Echocardiogram	0.65	**0.70**	0.63	0.62	0.57	0.59	0.60
Cylinderbands	0.62	0.63	0.62	**0.66**	0.66	0.64	0.66
Memographicmasses	0.80	**0.81**	0.81	0.78	0.80	0.79	0.80
Horsecolic	**0.88**	**0.88**	0.84	0.83	0.84	0.83	0.83
JRIP							
Labor	**0.90**	**0.90**	0.88	0.85	0.85	0.87	0.83
Echocardiogram	0.63	**0.71**	0.66	0.64	0.66	0.64	0.64
Cylinderbands	0.59	0.60	0.60	0.61	0.62	0.62	**0.63**
Memographicmasses	0.81	**0.82**	0.82	0.80	0.81	0.81	0.81
Horsecolic	0.87	**0.88**	0.87	0.86	0.88	0.86	0.87
NAIVE Bayes							
Labor	0.92	**0.99**	0.94	0.94	0.93	0.91	0.65
Echocardiogram	0.75	**0.81**	0.74	0.65	0.75	0.67	0.13
Cylinderbands	**0.54**	0.51	0.44	0.54	0.53	0.46	0.43
Memographicmasses	**0.82**	0.77	0.79	0.77	0.80	0.81	0.80
Horsecolic	0.84	**0.89**	0.86	0.85	0.82	0.81	0.81
K-Nearest							
Labor	0.83	0.92	**0.99**	0.95	0.89	0.85	0.78
Echocardiogram	0.53	**0.64**	0.57	0.56	0.60	0.54	0.54
Cylinderbands	0.58	0.60	0.60	0.62	0.68	0.69	**0.72**
Memographicmasses	0.73	**0.74**	0.73	0.73	0.74	0.71	0.74
Horsecolic	0.83	**0.86**	0.84	0.82	0.82	0.81	0.79
J-48							
Labor	0.78	**0.93**	0.88	0.87	0.90	0.85	0.78
Echocardiogram	0.53	**0.72**	0.60	0.61	0.63	0.56	0.59
Cylinderbands	0.59	0.61	0.59	0.64	0.64	0.64	**0.66**
Memographicmasses	0.80	**0.82**	0.81	0.79	0.80	0.79	0.80
Horsecolic	0.89	**0.90**	0.88	0.87	0.86	0.87	0.86

TABLE VIII. COMPARISON OF GA WITH DIFFERENT TECHNIQUES BASED ON AREA UNDER ROC

Datasets	NB-Tree						
	With missing value	GA filled	MI filled	Mean filled	Min filled	Max filled	Zero filled
Labor	0.88	**1.00**	0.95	0.89	0.84	0.86	0.92
Echocardiogram	0.92	**0.94**	0.92	0.88	0.93	0.89	0.92
Cylinderbands	0.70	**0.74**	0.70	0.73	0.73	**0.74**	0.73
Memographicmasses	0.89	**0.90**	0.89 |	0.89	0.89	0.88	0.85
Horsecolic	0.84	**0.89**	0.84	0.84	0.85	0.84	0.83
PART							
Labor	0.79	**0.91**	0.89	0.87	0.78	0.81	0.81
Echocardiogram	**0.93**	0.90	0.90	0.88	0.88	0.89	0.89
Cylinderbands	0.75	0.72	0.73	**0.76**	0.75	0.75	0.76
Memographicmasses	0.88	**0.89**	0.88	0.87	0.88	0.87	0.88
Horsecolic	**0.87**	0.85	0.79	0.78	0.79	0.78	0.77
JRIP							
Labor	**0.88**	0.86	0.85	0.81	0.76	0.85	0.76
Echocardiogram	0.83	**0.88**	0.83	0.83	0.85	0.83	0.84
Cylinderbands	0.66	0.69	0.67	0.70	0.71	0.71	**0.72**
Memographicmasses	0.84	**0.86**	0.85	0.84	0.84	0.84	0.84
Horsecolic	0.82	**0.85**	0.81	0.80	0.82	0.81	0.81
NAIVE Bayes							
Labor	0.89	**1.00**	0.96	0.94	0.99	0.90	0.80
Echocardiogram	0.96	**0.97**	0.96	0.91	0.96	0.93	0.71
Cylinderbands	0.70	0.70	0.71	0.69	0.71	**0.74**	0.73
Memographicmasses	**0.89**	0.86	0.86	0.85	0.89	0.89	0.89
Horsecolic	0.85	**0.93**	0.86	0.84	0.84	0.83	0.81
K-Nearest							
Labor	0.89	0.91	**0.99**	0.91	0.81	0.83	0.75
Echocardiogram	0.83	**0.97**	0.79	0.82	0.84	0.82	0.72
Cylinderbands	0.65	0.69	0.73	0.69	0.74	0.74	**0.75**
Memographicmasses	0.79	0.79	0.79	0.79	**0.80**	0.77	**0.80**
Horsecolic	0.73	**0.80**	0.78	0.77	0.77	0.74	0.71
J-48							
Labor	0.71	0.85	**0.88**	0.82	0.84	0.83	0.85
Echocardiogram	0.93	**0.93**	0.88	0.88	0.87	0.87	0.87
Cylinderbands	0.70	0.72	0.69	0.72	0.72	**0.74**	0.74
Memographicmasses	0.87	**0.87**	0.86	0.85	0.87	0.86	0.87
Horsecolic	**0.85**	0.84	0.83	0.81	0.81	0.82	0.79

VI. CONCLUSION AND FUTURE WORK

Data mining is an active area of research and in this area data is the most vital and valuable asset. Without applying automatic data mining techniques and preprocessing methods it is difficult to effectively analyze large amounts of data. Researchers are interested in finding efficient and accurate technique/method that cleans dirty and noisy data so that achieve higher accuracy rate, are comprehensible and can be learned in reasonable time, even for large databases.

In this paper, we addressed the problem of missing data imputation. First, we have elaborated on the importance of clean data (complete) in KDD. We have proposed an evolutionary technique for filling missing data on the basis of good estimation using GAs. Our main objective was to embed population-based search mechanisms to explore more search

space along with exploitation. The datasets used are standard datasets having by default missing values. We have also demonstrated that proposed technique works well for datasets with a greater percentage of missing values also for datasets where attributes are having a large range of distinct values, as GA gets into real play where there is space for more and more combination of different values. In future, we like to extend our algorithm to the domain of Noise reduction/removal.

REFERENCES

[1] J. Han, and M. Kamber, Data Mining: Concepts and Techniques, 2nd ed., Morgan Kaufmann Publishers, 2006.

[2] M. J. Berry, and G. Linoff. Data Mining Techniques for Marketing, Sales, and Customer Support. New York: John

[3] J. Pesce, Stanching hospitals, Financial hemorrhage with information technology, Health Management Technology, Vol. 24, No. 8, pp. 6-12, 2003.

[4] W. Ceusters, Medical natural language understanding as a supporting technology for data mining in healthcare Chapter 3 in: Cios K.J., eds. Medical Data Mining and Knowledge Discovery, Heidelberg: Springer-Verlag, pp. 32-60, 2000.

[5] A.C. Tessmer, "What to learn from near misses: an inductive learning approach to credit risk assessment," Decision Sciences, Vol. 28, No. 1, pp. 105-120, 1997.

[6] Roth, P. L. and Switzer III, F. S.: 1995, A Monte Carlo analysis of missing data techniques in a HRM setting, Journal of Managment 21, 10031023.

[7] Schafer, J. L. and Graham, J. W.: 2002, Missing data: Our view of the state of the art, Psychological Methods 7(2), 147177

[8] Allison, P. D.: 2002, Missing Data: Quantitative Applications in the Social Sciences, Thousand Oaks, CA: Sage.

[9] Little, R. J. A. and Rubin, D. B.: 1987, Statistical Analysis with Missing Data, Wiley, New York

[10] Batista, G. and Monard, M.C. (2003). An Analysis of Four Missing Data Treatment Methods for Supervised Learning, Applied Artificial Intelligence, 17, pp. 519-533.

[11] Graham JW, Cumsille PE, Elek-Fisk E. 2003. Methods for handling missing data. In Research Methods in Psychology, ed. JA Schinka, WF Velicer, pp. 87114.Volume 2 of Handbook of Psychology, ed. IB Weiner.New York:Wiley

[12] Multiple Imputation [Online], www.multiple-imputation.com

[13] Beunckens, C., Molenberghs, G., Verbeke, G, and Mallinckrodt, (2008). A latent- class mixture model for incomplete longitudinal Guassian data. Biometrics, 64, 96- 105.

[14] Little, R.J.(2009). Selection and patter mixture models. In Fitzmaurice, G., Davidian, M, Verbeke,G. & Molenberghs, G42 (eds.), Longitudinal Data Analysis , pp. 409-431. Boca Raton: Chapman & Hall/CRC Press

[15] Albert, P.S. & Follman, D.A. (2009). Shared parameter models.

[16] Papagelis A. and Kalles D. 2000. GAtree: Genetically Evolved Decision Trees, Proceedings 12th International Conference on Tools with Artificial Intelligence 13-15 November 2000 pages 203-206..

[17] Goldberg D.1999. Genetic Algorithms in Search, Optimization, and Machine Learning, Addison-Wesley.

[18] Nelwamondo F V, Mohamed S, Marwala T. Missing Data: A Comparison of Neural Network and Expectation Maximisation Techniques, eprint arXiv:0704.3474, April 2007

[19] Betechuoh B L, Marwala T, TetteyT, Autoencoder networks for HIV classification, Current Science, Vol 91, No 11, December 2006, pp 1467-1473.

[20] I.H. Witten, and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques, 2nd ed. Morgan Kaufmann, 2005.

[21] J.R. Quinlan, Generating production rules from decision trees, in Proceedings of International Joint Conference of Artificial Intelligence, pp. 304-307, San Francisco, USA, 1987.

[22] P. Eklund, and A. Hoang, A performance survey of public domain machine learning algorithms, Technical Report, School of Information Technology, Griffith University, 2002.

[23] R. Rastogi, and K. Shim, A decision tree classifier that integrates building and pruning, Data Mining and Knowledge Discovery, Vol. 4, pp. 315344, 2000.

[24] Eibe Frank, Ian H. Witten: Generating Accurate Rule Sets Without Global Optimization. In: Fifteenth International Conference on Machine Learning, 144-151, 1998.

[25] W. Cohen, Fast effective rule induction, in Machine Learning: Proceedings of the Twelfth International Conference (ML95), pp. 852-857, 1995.

[26] Ron Kohavi: Scaling Up the Accuracy of Naive-Bayes Classifiers: A Decision-Tree Hybrid. In: Second International Conference on Knoledge Discovery and Data Mining, 202-207, 1996.

[27] M.L. Zhang and Z.H. Zhou, A k-nearest neighbor based algorithm for multi-label classification, in 1st IEEE International Conference on Granular Computing, Vol. 2, pp 718721, 2005.

[28] N. Friedman, D. Geiger, and M. Goldszmidt, Bayesian network classifiers, Journal of Machine Learning Research, Vol. 29, pp. 131163, Dec. 1997.

[29] A. Hedar, J. Wang, and M. Fukushima, "Tabu search for attribute reduction in rough set theory", presented at Soft Comput, 2008, pp.909-918.

[30] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten, The WEKA Data Mining Software: An Update, SIGKDD Explorations, Volume 11, Issue 1, 2009.

[31] S. Hettich, and S.D. Bay, The UCI KDD Archive. Irvine, CA: Dept. Inf. Comput. Sci., Univ. California, 1996 [Online]. Available: http://kdd.ics.uci.edu.

Multitaper MFCC Features for Acoustic Stress Recognition from Speech

Salsabil Besbes
University of Tunis El Manar
National School of Engineers of Tunis
Signal, Image and Information Technology laboratory
BP. 37 Le Belvdre, 1002, Tunis, Tunisia

Zied Lachiri
University of Tunis El Manar
National School of Engineers of Tunis
BP. 37 Le Belvdre, 1002, Tunis, Tunisia

***Abstract*—Ameliorating the performances of speech recognition system is a challenging problem interesting recent researchers. In this paper, we compare two extraction methods of Mel Frequency Cepstral Coefficients used to represent stressed speech utterances in order to obtain best performances. The first method known as traditional is based on single window (taper) generally the Hamming window and the second one is a novel technique developed with multitapers instead of a single taper. The extracted features are then classified using the multiclass Support Vector Machines. Experimental results on the SUSAS database have shown that the multitaper MFCC features outperform the conventional MFCCs.**

***Keywords*—Mel Frequency Cepstral Coefficients (MFCC); Multitapering; Multiclass SVM; Stress recognition**

I. Introduction

Automatic stress speech recognition have been used in several applications such as human to machines communications, craft voice communications, medicine and psychology. Stress refers to human response to different factors such as workload task, environmental condition and health. Stress has an impact in the performance of a person in his daily life. It affects the brain, the muscles, the eyes, the cardiovascular system and especially the speech production system.

Stress recognition systems are composed of two important steps which are feature extraction and feature classification. In literature, different classifiers have been used including Hidden Markov Model(HMM) [1] , Artificial Neural Network systems(ANN) [2], Gaussian Mixture Model(GMM) [3] and Support Vector Machines (SVM) [4].

In last years, extracting the most suitable features set for stressed speech recognition has been an important subject in many researches. Feature extraction aims to obtain a compact representation of speech signals. Studies have proposed different acoustic features to represent the speech under stress signals. These features are essentially Pitch, energy, [5], Linear Predictive Cepstral Coefficients(LPCC) [6] and Mel Frequency Cepstral Coefficients (MFCC). Different features have been extracted in [7] including pitch, energy, formants and MFCC from the stressed speech.

The MFCC features are the most common used features in speech processing because they are based on human auditory system. Usually, MFCCs are computed using a windowed periodogram via the Discrete Fourier Transform (DFT) [8]. It has been demonstrated that the spectrum estimate obtained has a high variance despite it low bias. A solution was proposed to reduce the spectral variance using multitaper spectrum estimate instead of the single windowed periodogram [9], [10]. In order to have a low variance spectrum estimate, the multitaper method applied a set of orthogonal tapers to the speech signal and an average sum of the sub-spectra are then calculated.

The multitaper approach have been used in several domains including geophysical applications [11], speaker verification [12], [13] and emotion recognition [14], [15] and it has been shown to improve the performance and robustness of different systems. However, this method has not been used in stress speech recognition applications. So, the aim of this work is to investigate multitaper MFCC features (MMFCC) in order to improve the performances of the stressed speech recognition system. We are also interested to compare different methods of multitapering including Thomson method, Multipeak method and SWCE (Sinusoidal Weighted Cepstrum Estimator) method.

This paper is structured as follows: Section II present the stressed speech recognition system proposed in this work. Section III presents the process of multitaper MFCC extraction and Section IV describes the multitaper spectrum estimate method. Section V deals with multiclass Support Vector Machines approaches. Results and experiments are given in Section VI. Finally, conclusion is presented in Section VII.

II. System Framework

The system proposed for stress speech recognition is illustrated in figure 1. First, we extract MFCCs and MMFCCs from the stressed speech signals of the SUSAS database. These features are then divided into training and test sets. Second, the classification is realized based on multiclass SVM methods which are One-Versus-Rest(1vR), One-Versus-One(1v1) and Directed Acyclic Graph (DAG). The performance of the system is evaluated with accuracy rate using the test set.

III. Multitaper MFCCs Extraction

In this section, we describe the extraction procedure of the multitaper MFCC features in order to compare it to the traditional extraction technique of MFCC features [8].
The traditional MFCC coefficients can be obtained following the same steps described for MMFCCs and using a special case

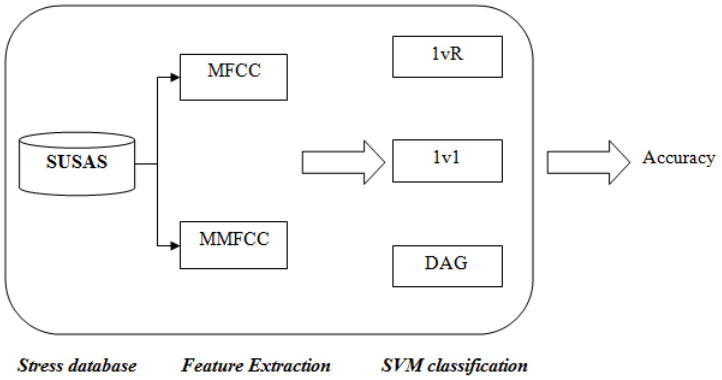


Fig. 1: System framework

of multitaper spectrum estimate which leads to a Hamming windowed spectrum.

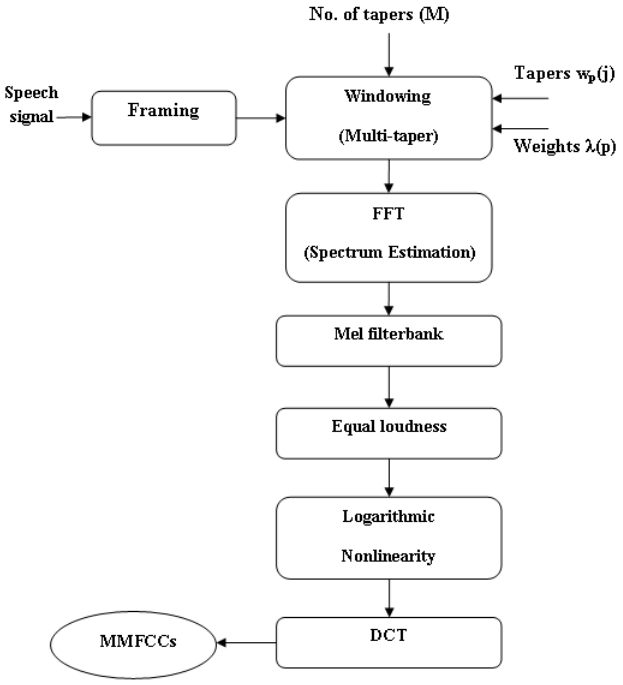


Fig. 2: Extraction procedure of multitaper MFCC

IV. Multitaper Spectrum Estimate

The most used method for spectrum estimation in speech processing is the windowed periodogram known also as Hamming windowed DFT spectrum [17]. This spectrum estimate can be formulated as:

$$\hat{S}(f) = |\sum_{j=0}^{N-1} w(j)s(j)e^{\frac{-2i\pi f}{N}}|^2 \quad (1)$$

where $f \in \{0, 1, ..., N-1\}$ is the frequency bin index, $[s(0), s(1), ..., s(N-1)]$ is a speech frame with length N and $w(j)$ denotes a window function. Equation 1 is also called a single-taper periodogram.

The use of a single taper for spectrum estimate reduces the bias (the difference between the estimate spectrum and the real spectrum) but causes a problem of discarding a significant part of the signal. Indeed, the spectral estimate will have high variance. This variance can be reduced by using multitaper method for spectrum estimate. The multitaper spectrum estimate is calculated by:

$$\hat{S}_{MT}(f) = \sum_{p=1}^{M} \lambda(p)|\sum_{j=0}^{N-1} w_p(j)s(j)e^{-\frac{2i\pi if}{N}}|^2 \quad (2)$$

where w_p is the p^{th} data taper ($p = 1, 2, ..., M$), $\lambda(p)$ is the weight of the p^{th} taper and N is the frame length.
The tapers w_p are chosen to be orthonormal such as:

$$\sum_j w_p(j)w_q(j) = \delta_{pq} = \begin{cases} 1 & \text{if } p = q \\ 0 & \text{otherwise.} \end{cases} \quad (3)$$

So, the multitaper spectrum is the weighted sum of sub-spectra. The single taper power spectrum estimate can be obtained when we set $p = M = 1$ and $\lambda(p) = 1$. The windows functions (tapers) of the multitaper approaches are taken so that the sub-spectra have uncorrelated estimation errors. Indeed, a low variance estimate is obtained when we average these uncorrelated sub-spectra. In literature, different multitaper methods have been proposed: Thomson multitaper [18], Multipeak multitaper [19] and SWCE (Sinusoidal Weighted Cepstrum Estimator) multitaper [20]. The choice of tapers affect significantly the calculated spectrum estimate. These tapers should be resistant to spectral leakage.

V. Classification

Different techniques have been used in the literature to classify stressed states in speech such as Artificial Neural Networks [2], Gaussian Mixture Models [3], Hidden Markov Model [1], and Support Vector Machines [4].

Support vector machines (SVM) are widely used as learning machine classifiers in the past decade due to their best performances compared to traditional mythologies used in

solving signal processing problems. The SVM have showed to perform other well-known classifiers in stress and emotion recognition and was used in many studies [21], [22].

In this work, we are interested to SVM to classify multitaper MFCCs into appropriate stress classes. The principle of SVM is to find an optimal hyperplane that separates two classes using the maximized margin criteria. The SVM were originally implemented to solve problems of binary classification. But, real applications oblige the researchers to extend SVMs to multiclass approaches [23]. Different methods have been proposed including one-versus-rest (1vR), one-versus-one (1v1) and directed acyclic graph (DAG).

The 1vR approach is the simplest and the oldest one [24]. It builds k SVM (one per class). The 1vR SVM is based on training the j^{th} SVM with all the examples of this j^{th} class considered as positive ones and all other examples as negative ones. It can also be used to discover the reject of example which does not belong to any of the k classes. However, this approach is almost criticized because of its asymmetry due to the fact that each hyperplane is training with a number of negative examples more important than the number of positive ones. This problem can be resolved by the use of the 1v1 method which constructs $k(k-1)/2$ binary classifiers (k the number of classes) [25]. Despite that the 1v1 uses a bigger number of hyperplane in the training phase than the 1vR, this method is often faster. The DAG SVM is trained in the same way of training the 1v1 method [26]. However, a rooted binary directed acyclic graph is used during the test phase. This graph has $k(k-1)/2$ nodes where each node is a binary SVM. This method has been developed in order to resolve the problem of areas of indecision caused by the OAO approach. Moreover, it has been demonstrated that the DAG SVM is faster than the 1v1 and the 1vR methods.

The SVM approaches are based on kernels. In literature, there are some kernels which are widely used and are considered as standard kernels. These kernels are linear, polynomial and gaussian. In this work, we used a polynomial and a Gaussian kernel which are defined as follows:

$$K_{Gaussian}(x, x_i) = \exp(-\frac{\|x - x_i\|^2}{2\sigma^2}) \quad (4)$$

$$K_{Poly}(x, x_i) = (a* < x, x_i > +b)^d \quad (5)$$

VI. Experiments

A. Stress Corpus

The stressed database SUSAS (Speech under Simulated and Acted Stress) [27] consists of stressed speech samples recorded under simulated environment and acted environment. In this study, we used only the speech utterances recorded under simulated stress conditions. This domain consists of speech uttered by 9 speakers having 3 different dialects and contains 10 different stressed styles. SUSAS comprises a set of 35 aircraft communication words. Each speaker is reading these words twice. The sampling frequency is 8 Khz.

Four states of speech under stress: Neutral(N), Angry(A), Loud(Ld) and Lombard(Lb) are considered. In our experiments, speech utterances of 8 speakers are considered that to say that 2240 isolated words are used in the training and the test phases. The two thirds of data are used as training set and the third of this data is used for test.

B. Experimental setup

The performance of multitaper MFCCs for speech under stress recognition is evaluated using the SUSAS (Speech Under Simulated and Actual Stress) database. The MFCCs and MMFCCs features are extracted from the data collected from the isolated words which represent the four stress states examined.

For a comparative study, we implement the multitaper approach with different types of tapers (described in Table I) in order to extract multitaper MFCC features. These methods are Thomson, Multipeak and SWCE. The classical MFCC features are computed using the Hamming window. The multitaper methods were used by varying the number of tapers ($2 \leq p \leq 8$). In the experiments, the speech signals were segmented into frames of 10ms lengths. After that, each frame was weighted by a single taper or multitaper method. The different multitaper methods Thomson, Multipeak and SWCE were generated as described in [13].

For evaluation, we have used multiclass SVM approaches including 1vR, 1v1 and DAG. The multiclass SVMs approaches were implemented using the LIBSVM toolbox for Matlab. The different methods are evaluated with two kernels: polynomial and gaussian. The RBF kernel parameters (c, σ) are calculated using the cross-validation procedure [28] where:

$$c = [2^{-15}, 2^{-14}, ..., 2^{14}, 2^{15}] \quad (6)$$

and

$$\sigma = [2^{-15}, 2^{-14}, ..., 2^{14}, 2^{15}] \quad (7)$$

C. Results

In this study, a recognition system for stressed speech is implemented using MFCCs and MMFCCs features. Stress states are recognized with three multiclass SVM classifiers including 1vR, 1v1 and DAG approaches applied with polynomial and gaussian kernels.

The results of the first classification approach are presented in table II. Comparing the classification rates obtained with traditional MFCCs to those obtained with the multitaper MFCC features computed with the three multitaper methods, we can notice that when we use the polynomial kernel, the classification accuracies are improved ranging from 80.05% to 93.44%. These results depend on the multitaper method used and the number of tapers. However, the use of the RBF kernel with MMFCC ameliorate the performance of the stressed speech recognition system in same cases. The best classification rate of 99.44% is obtained with the application of Thomson multitaper method with a number of tapers $p = 3$.

The same experiments were conducted with the two other multiclass SVM approaches: 1v1 and DAG. Results of the second and the third SVM approaches are illustrated in table III and table IV. Indeed, for the 1v1 SVM method, the classification accuracies obtained with the polynomial kernel

TABLE I: Stress Speech Recognition Systems based on Single Taper and Multitaper MFCCs

Appraoch	Description
Hamming	single taper MFCCs using Hamming window
Thomson	Multitaper MFCCs using dpss tapering
Multipeak	Multitaper MFCCs using multipeak tapering
SWCE	MFCC are computed from sinusoidal weighted (i.e., sine tapered) spectrum estimate

TABLE II: Classification Accuracy using 1vR/SVM

Feature	Number of tapers	Polynomial	Gaussian
MFCC	1	76.94	99.06
MMFCC-Thomson	2	93.30	98.92
	3	92.40	99.44
	4	88.62	97.72
	5	89.42	99.19
	6	89.95	98.52
	7	90.46	98.61
	8	87.95	99.19
MMFCC-Multipeak	2	92.10	99.33
	3	84.47	97.32
	4	92.36	99.33
	5	80.05	98.12
	6	91.70	98.52
	7	81.84	98.96
	8	86.63	98.26
MMFCC-SWCE	2	93.44	98.79
	3	92.77	99.19
	4	91.29	98.79
	5	91.16	77.64
	6	88.89	99.19
	7	91.03	98.25
	8	87.81	98.39

are in [74.44%, 95.04%] and with the RBF kernel in [97.99%, 99.30%].

We can remark that both kernels gives important results but the improvement is very remarkable with the polynomial kernel. The accuracy has passed from 50.88% for classical MFCC to 95.05% with the Multipeak multitaper method used with a number of tapers $p = 2$. This amelioration is also obtained when we applied the DAG SVM method. The best results are obtained when we use the gaussian kernel associated to MMFCC computed with Multipeak approach ($p = 7$).

From the three tables representing the results, we can conclude that the use of multitaper methods to extract MFCC features improve the performances of the stressed speech recognition system. An important improvement exceeding 45% is achieved with the polynomial kernel but the best accuracies are often obtained by the application of the gaussian kernel with the three multiclass SVM approaches.

VII. Conclusion

In this paper, we have used the multitaper method in order to extract MFCC features for stressed speech recognition. The windowed DFT used in the traditional extraction process is replaced by multitaper spectrum estimation. The evaluation of the stressed speech recognition system is realized on SUSAS database using multiclass SVM approaches. The results show that there is an improvement in the performances of the implemented stressed speech recognition systems with multitaper MFCCs.

For future work, multitaper approach can be useful in extraction of other features from the stressed speech such as multitaper gammatone frequency cepstral coefficients and multitaper PLP in order to improve the recognition accuracy. Also, we can test the multitaper MFCC with other classification approaches.

Acknowledgment

This work has been supported by the Research Laboratory LR-SITI-ENIT (Signal, Images and Information Technolo-

TABLE III: Classification Accuracy using 1v1/SVM

Feature	Numbre of tapers	Polynomial	Gaussian
MFCC	1	50.80	98.79
MMFCC-Thomson	2	92.50	98.66
	3	91.29	99.30
	4	91.96	99.19
	5	90.62	98.39
	6	89.02	98.79
	7	88.53	99.03
	8	86.74	98.39
MMFCC-Multipeak	2	95.04	98.52
	3	75.36	97.99
	4	92.77	99.19
	5	75.10	98.92
	6	92.10	98.92
	7	74.44	99.48
	8	89.17	98.93
MMFCC-SWCE	2	92.77	99.19
	3	92.77	98.92
	4	93.03	98.92
	5	91.56	98.52
	6	89.29	99.06
	7	88.62	98.66
	8	90.36	98.92

TABLE IV: Classification Accuracy using DAG/SVM

Feature	Number of tapers	Polynomial	Gaussian
MFCC	1	49.73	99.19
MMFCC-Thomson	2	93.97	98.92
	3	93.78	99.30
	4	91.29	98.66
	5	90.36	98.39
	6	89.15	98.79
	7	87.70	99.30
	8	86.61	98.52
MMFCC-Multipeak	2	95.04	98.25
	3	78.44	99.19
	4	92.36	98.52
	5	74.96	97.99
	6	91.03	98.79
	7	75.87	99.61
	8	90.77	98.26
MMFCC-SWCE	2	93.03	98.92
	3	91.83	99.19
	4	92.23	98.52
	5	90.89	99.19
	6	91.03	99.46
	7	90.62	99.19
	8	90.49	99.06

gies), ENIT.

REFERENCES

[1] B.D. Womack , J.H.L. Hansen , N-channel hidden Markov models for combined stressed speech classification and recognition Speech and Audio Processing, IEEE Transactions on, VOL. 7, NO. 7, pp. 668-677, 1999.

[2] B. D.Womack and J. H. L. Hansen, Classification of speech under stress using target driven features, Speech Commun: Speech Under Stress, VOL. 20, pp. 131150, November 1996.

[3] D. Ververidis , C. Kotropoulos , Emotional speech classification using Gaussian mixture models, IEEE International Symposium on, VOL. 3, pp. 2871- 2874, 2005.

[4] T. Nguyen and I. Bass, Investigation of Combining SVM and Decision Tree for Emotion Classification, Proceedings of the Seventh IEEE International Symposium on Multimedia (ISM05), December 2005.

[5] P Boersma, Accurate Short-Term Analysis of the Fundamental Frequency and the Harmonics to-Noise Ratio of a Sampled Sound, Proceedings of the Institute of Phonetic Sciences, VOL. 17, pp. 97110, 1993.

[6] S. E. Bou-Ghazale and J. H. L. Hansen, A Comparative Study of Traditional and Newly Proposed Features for Recognition of Speech Under Stress, IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING, VOL. 8, NO. 4, pp. 429-442, JULY 2000.

[7] S. Besbes and Z. Lachiri, Multi-class SVM for stressed speech recognition, 2nd International Conference on Advanced Technologies for Signal and Image Processing (ATSIP), pp. 782-787, March 2016.

[8] C. Ittichaichareon, S. Suksri and T. Yingthawornsuk, Speech Recognition using MFCC, International Conference on Computer Graphics, Simulation and Modeling (ICGSM2012),Pattaya (Thailand), July 28-29, 2012.

[9] T. Kinnunen, R. Saeidi, J. Sandberg, M. Hansson-Sandston, What Else is New than the Hamming Window? Robust MFCCs for Speaker Recognition via Multitapering, International Speech Communication Association, INTERSPEECH, pp. 2734-2737, 2010.

[10] M. Hansson-Sandsten and J. Sandberg, Optimal cepstrum estimation using multiple windows, IEEE ICASSP, Taipei, Taiwan, pp. 30773080, 2009.

[11] M. A. Wieczorek and F. J. Simons, Minimum-variance multitaper spectral estimation on the sphere, The Journal of Fourier Analysis and Applications, vol. 13, no. 6, pp. 665692, 2007.

[12] J. Sandberg, M. Hansson-Sandsten, T. Kinnunen, R. Saeidi, P.Flandrin, and P. Borgnat, Multitaper estimation of frequency warped cepstra with application to speaker verification, IEEE Signal Processing Letters, vol. 17, no. 4, pp. 343346, 2010.

[13] T. Kinnunen, R. Saeidi, F. Sedlk, K. A. Lee, J. Sandberg, M. Hansson-Sandsten, and Haizhou Li, Low-Variance Multitaper MFCC Features: A Case Study in Robust Speaker Verification, IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 20, NO. 7, SEPTEMBER 2012.

[14] Y. Attabi, M. J. Alam, P. Dumouchel, P. Kenny, and D. OShaughnessy, Multiple windowed spectral features for emotion recognition, in Proceedings of the 38th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP13), pp. 75277531, IEEE, Vancouver, Canada, May 2013.

[15] S. V Chapaneri, D. D. Jayaswal, Multi-Taper Spectral Features for Emotion Recognition from Speech, International Conference on Industrial Instrumentation and Control (ICIC), Pune India, May 28-30, 2015.

[16] J. Trangol and A. Herrera, Traditional Method and Multi-Taper to Feature Extraction Using Mel Frequency Cepstral Coefficients, International Journal of Information and Electronics Engineering, Vol. 5, No. 1, January 2015

[17] F. J. Harris, On the use of windows for harmonic analysis with the discrete Fourier transform, Proc. IEEE, vol. 66, no. 1, pp. 5184, Jan. 1978.

[18] D. J. Thomson, Spectrum estimation and harmonic analysis, IEEE proceeding, vol. 70(9), pp. 10551096, 1982.

[19] M. Hansson and G. Salomonsson, A multiple window method for estimation of peaked spectra, IEEE Trans. on Sign. Proc., vol. 45(3), pp. 778781, 1997.

[20] K. S. Riedel and A. Sidorenko, Minimum bias multiple taper spectral estimation, IEEE Trans. Signal Process., vol. 43, no. 1, pp. 188195, Jan 1995.

[21] C.W. Hsu, C.C. Chang and C. J. Lin, A practical guide to support vector classification, Department of Computer Science and Information Engineering National Taiwan University, Taipei, Taiwan, 2009. Available: www.csie.ntu.edu.tw/ cjlin/

[22] A. Hassan and R. I. Damper, Multi-class and hierarchical SVMs for emotion recognition, In Proc. Interspeech, 2010.

[23] C. Hsu and C. Lin, A comparison of methods for multiclass support vector machines, IEEE Transactions on Neural Networks, VOL. 13, NO. 2, pp. 415-425, 2001.

[24] N. Cristianini and J. Shawe-Taylor, An Introduction to Support Vector Machines and other Kernel-based Learning Methods, Cambridge, UK: Cambridge University Press, 2000.

[25] C. Hsu and C. Lin, A comparison of methods for multiclass support vector machines, IEEE Transactions on Neural Networks, Vol. 13, No. 2, pp. 415425, 2001.

[26] J. Platt, N. Cristianini, and J. Shawe-Taylor, Large margin DAGs for multiclass classification, Proceedings of Neural Information Processing Systems, NIPS99, Denver, CO, pp. 547553, 2000.

[27] J. H. L. Hansen and S. E. Bou-Ghazale, Getting Started with SUSAS: A Speech Under Simulated and Actual Stress Database, EUROSPEECH 1997.

[28] L. I. Kuncheva, Combining pattern classifiers methods and algorithms. New York: Wiley, 2004.

Rule Adaptation in Collaborative Working Environments using RBAC Model

Ahmad Kamran Malik
Department of Computer Science
COMSATS Institute of Information Technology
Islamabad, Pakistan

Muhammad Anwar
Department of Computer Science
Federal Urdu University of Arts, Science & Technology (FUUAST)
Islamabad, Pakistan

Abdul Mateen
Department of Computer Science
Federal Urdu University of Arts, Science & Technology (FUUAST)
Islamabad, Pakistan

Wajeeha Naeem
Department of Computer Science
COMSATS Institute of Information Technology
Islamabad, Pakistan

Yousra Asim
Department of Computer Science
COMSATS Institute of Information Technology
Islamabad, Pakistan

Malik Ahsan Ali
Department of Computer Science
Federal Urdu University of Arts, Science & Technology (FUUAST)
Islamabad, Pakistan

Basit Raza
Department of Computer Science
COMSATS Institute of Information Technology
Islamabad, Pakistan

***Abstract*—Collaborative Working Environments (CWEs) are getting prominence these days. With the increase in the use of collaboration tools and technologies, a lot of sharing and privacy issues have also emerged. Due to its dynamic nature, a CWE needs to adapt the changes into accordingly. In this paper, we have implemented the Adaptive Dynamic Sharing and Privacy-aware Role Based Access Control (Adaptive DySP-RBAC) model which provides user's information privacy to dynamically adapt the changes occurring in the system at any time. The proposed model has been implemented as a prototype and tested. Results have shown that our system efficiently and effectively adapts access rules according to the changes happening in a CWE along with preserving the user's information privacy in the system.**

***Keywords*—Dynamic Adaptation; RBAC; Privacy; Collaboration**

I. Introduction

Today, mutual teamwork-based working environments and cooperation among the members have gained prominence. Now, people without working together at the same place can also collaborate with each other. This type of working environments is known as Collaborative Working Environment (CWE). With the help of CWE, people can share their ideas, efforts, results, inventions etc. and at the same time, they can change their locations as well. A lot of work has been done in order to improve CWEs. In [1], seven different collaboration factors related to collaboration have been discussed. Along with that, a collaborative working model has been presented by summarizing all those factors. With the increase in the use of CWEs, the need for privacy preservation and access control has also increased.

Dynamic Adaptation is yet another important characteristic of a CWE. In most of the collaborating environments, participating users may join or leave the group at any time. This causes the nature of groups to be dynamic. Due to which, rules and policies relating to the collaborating groups need to change as well. In [2], the concept of dynamic adaptation has been presented in the form of virtual teams. It advocates that virtual teams contain members who can geographically be located anywhere in the world. They can not only be linked but also participate equally with the help of emerging telecommunication technologies. In [3], an intelligent system for dynamic collaborations has been presented. In this system, changes frequently occur due to change in the users participating in the collaborative environment, which a system adapts effectively.

We aim to implement a CWE in which system adapt the related variations as and when any change in the system occurs. We have implemented the Adaptive Dynamic Sharing and Privacy-Aware Role-Based Access Control (Adaptive DySP-RBAC) model by extending the DySP-RBAC model [4] to incorporate dynamic adaptation of access control rules. We have monitored dynamic adaptation in different scenarios and recorded results to show the level of adaptation in the respective system. Since information sharing is inevitable in a CWE, that's why it has been considered a crucial step to

provide privacy preservation of information. In [4], privacy and access control has been provided by creating the different type of policies. These policies have also been evaluated and compared with other access control models in [5].

Since CWEs usually have dynamic nature, user's personal and shared data or resources may also change with the change of user's participation in the collaborating environment. Context-awareness is yet another important feature of a CWE, so immense work has been done towards the improvement of context-awareness especially in collaborative environments. Such as in [6], context-aware computing has been defined along with its different categories. Cases of these categories have been prototyped and evaluated for results. Our system monitors user's information and performs adaptation more effectively according to the change in user's personal and shared resources information.

Rest of the paper has been organized as follows. Related work has been presented in Section II. Section III comprises of the architecture details. Results have been discussed in section IV. Section V concludes the paper and also describes future work.

II. Related Work

A lot of work has been done in order to improve CWEs with respect to performance aspects as well as security. A framework named as "Distribute Cognition" has been presented in [7] to explain and analyze collaborative working. In this paper, some theoretical and practical issues regarding collaboration have been discussed. In [8], interoperability issues in CWEs have been focused. For this purpose, a generic CWE has been proposed which allows different types of groupware to collaborate easily and effectively using different Web services technologies. Similarly, in order to deal with new emerging challenges in CWEs, an approach named inContext has been proposed in [9]. This approach has been used to combine some collaboration services which are considered dissimilar with the help of web services. It also handles the CWEs that are considered of dynamic nature. In [10], privacy issues faced due to collaboration has been discussed and a privacy framework has also been given to improve privacy issues. This framework can be adapted for any type of domain since it contains generic privacy ontology. It contains privacy rules also, through which information access has been defined; this part has been named as reasoning engine. In [5], different collaborating environments such as RBAC, Team-based Access Control model (TMAC) and Extended RBAC model has been implemented and evaluated for sharing and privacy preserving rules and metrics. In [11], different challenges related to collaborative environments have been discussed. In the light of those issues, related solutions have also been proposed.

Dynamic adaptation in CWE has immensely been focused such as in [12], a REal-time Software Adaptation System (RESAS) has been presented. In this framework, a tool has been provided to programmers in order to adapt the changes in real time. Similarly, in [13], a policy-driven and context-aware dynamic adaptation framework named Chisel has been proposed. In this proposed system, with the change of user and application context, behavior of different service objects automatically adapted by the system. With respect to different context, a number of policies have also been associated. In [14], a model has been presented in which a user has been provided related policies whenever a change has occurred in the system. The proposed access control system has a feedback component which has been named as "know". Policy protection and level of feedback have been provided by this feedback component. Rules or policies have been efficiently implemented through Ordered Binary Decision Diagrams (OBDDs). In [15], an interactive access control model has been proposed to further improve autonomous computing systems. The idea is based on the interaction between the clients and servers in order to provide access to any resource. On the basis of credentials provided by the clients, servers grant or deny access upon evaluating predefined policies.

In [16], a context-aware access control mechanism has been proposed for ubiquitous applications, for this purpose standard RBAC model has been extended. In this mechanism, the system dynamically adapts changes and grants permissions accordingly as and when a change occurs in the context. Another access control policy model has been given in [17], which has focused context awareness for the sake of resource access and dynamic adaptation for accommodating changes caused in context. Along with that, semantic technologies have been used to specify context/policies in the system. In [3], an intelligent information sharing control system has been presented in which sharing and control policies have been dynamically adapted as and when a change occur in user context, relationships, activities, and interactions. In [18], definitions of context-awareness in Internet of Things (IoT) and Internet of Everything (IoE) along with their architecture have been presented. Similarly, current context-aware approaches in systems such as IoT and IoE have also been analyzed.

In next section, details of proposed model along with some scenarios have been explained.

III. Architecture

As mentioned earlier, we have implemented an extension of DySP-RBAC model and evaluated it how the model dynamically adapts the changes occurred in the system. Details of the model have been given below:

A. DySP-RBAC Model

DySP-RBAC is an extension of core RBAC model. Along with user roles, it focuses on teams and tasks as well, including other data elements such as user, session, and permissions called sharing and privacy aware permissions. This is because permissions have also been created with the help of sharing and privacy elements. Sharing elements are used to enhance sharing among collaborating users. Sharing elements include Collaborative Relationships (CR) and Access Level (AL). Privacy elements are used to preserve the privacy of user's personal and sharing resources. Privacy elements include Purpose (Pur), Condition (Con) and Obligation (Obl). In core RBAC permissions were based on only objects and operations. In Fig. 1, DySP-RBAC model is shown. Any person who is participating in the system is termed as a user while that user may be a member of one or more teams. A user can be assigned multiple tasks in each team he is participating in. Similarly, a user can have multiple roles according to which

he will be assigned appropriate team and task. Objects contain user related information such as his teams, tasks, as well as personal information.

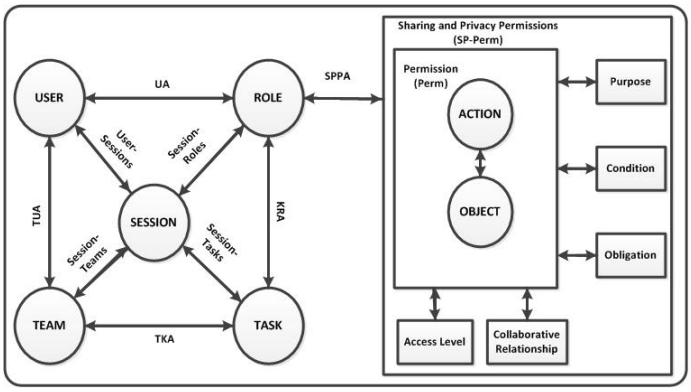


Fig. 1. DySP-RBAC Model [4]

B. Sharing and privacy

In order to provide access control, sharing and privacy based permissions have been used. It means that whenever a user wants to access other user's resource, the request will be evaluated according to the permissions created by its owner. Similarly, a user can create conditions to allow or restrict access to his resource. The level of information sharing has been determined through CR and AL. There are three types of collaborative relationships such as Mutual, Member, and Colleague. If some users are participating in the same team and they have been assigned the same task as well, their CR will be considered as Mutual, while users who are team members but they do not share the same task termed as Member, likewise users who neither have been on the same team nor assigned same tasks are considered as colleagues. Theses CR levels determine the level of information sharing at the different type of collaborative relationships. For example, users having CR as "Mutual" will have more resource access as compared to "Colleague". Similarly, for users having different collaborative relationships or roles, three type of access levels have been used i.e. Level 1 (L1), Level 2 (L2) and Level 3 (L3). So, users assigned AL as L1 will have higher access as compared L3, which will have lowest access level.

Whenever a user wants to create permissions, he will select one or more elements, for example, a CR, an AL, a Team, a Task, a Role and a resource for which he will create permissions. For example, if a user has created a permission having Mutual (as CR), L2 (as AL), Team A (as Team), T1 (as Task), R1 (as Role) and Location (as Resource). This means when any other user will request to access this users "Location", it will be checked that requesting user must be his Mutual (in CR), he must have an L2 (in AL), he must be a member of the specified team (i.e. Team A in this case), he must be assigned the specified task (i.e. T1 in this case), and he must have a particular role (such as R1 in this case). If the requesting user fulfills any of the mentioned conditions he will be allowed access to the requesting resource.

C. Dynamic Adaptation

Our scenario is an enterprise-based system in which different users perform their assigned tasks in the form of teams. Each user will have assigned tasks on the basis of roles that they have been allocated. Users, roles, teams, and tasks have dynamic nature in the system. This means, when a user leaves a team, all his related information and permissions should be dynamically adapted according to the new situation. Similarly, when a task is finished within a team, this will cause the task related information to change. Our system is capable of accommodating all changes taking place as a result of dynamic adaptation. The whole adaptation process has been shown in Fig. 2. Four main steps are Monitor, Analyze, Plan and Execute. Our system continuously monitors the changes on the basis of its knowledge. System knowledge includes collaborative relationships among users, their access levels, permissions/policies and system entities such as teams, tasks, and roles etc. It then analyzes them so that it can be determined that how the changes, which have already being monitored, can be adapted to the new situation. The system plans accordingly and executes the adaptation of changes occurred.

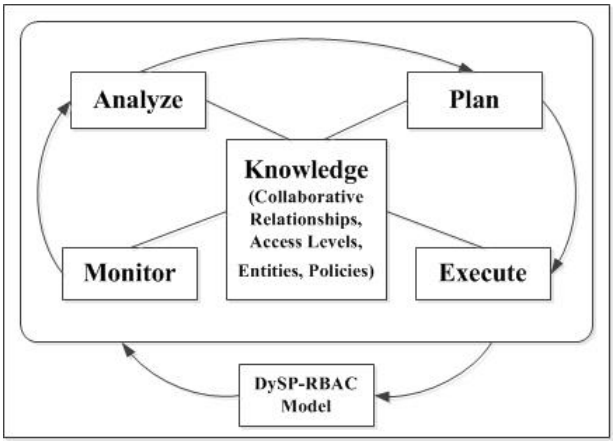


Fig. 2. Adaptive DySP-RBAC Model

1) Scenario: In this part, we have presented a scenario so that dynamic adaption in our system can be explained and understood more effectively. We have taken an enterprise-based CWE in which people collaborate in different teams and perform different tasks. These teams can be overlapping in nature because users can participate in more than one tasks at a time. Similarly, users can join or leave any team or task any time, making the whole scenario of dynamic nature. This can also happen due to the finishing/completion of a task within a team or any team/task can also be revoked from a user at any time. In order to perform a task, users may share and request to share each others resource information. It may be related to a user's personal information i.e. location etc. or team task related information. There are two types of sharing control policies which have been used in our proposed model; User-defined policy and Enterprise-defined policy. Users create permissions/policies to allow or restrict access to their personal or shared data. This is called user-defined policy. In order to control sharing of team related or task related information, enterprise-defined policies will be used [3].

In Fig. 3, we have presented an example of proposed

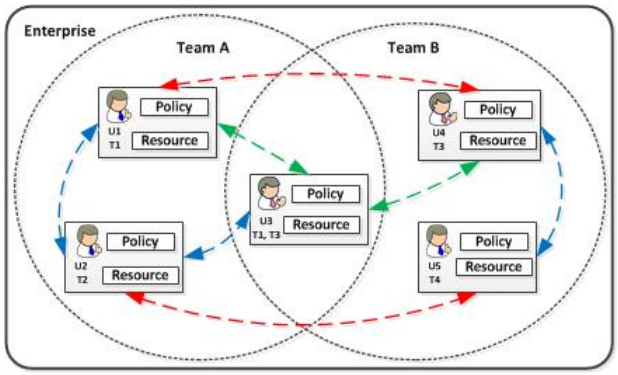


Fig. 3. Dynamic adaptation in a CWE (before) [3]

scenario. We have an enterprise in which users interact with each other in the form of teams and work on different tasks, which may be mutual or individual as well. We have two teams "Team A" and "Team B". Tasks include T1, T2, T3, and T4, while five users are named as U1, U2, U3, U4, and U5. There are policies or permissions associated with each user. The users U1 and U2 are part of the same team, which is team A, but they do not share same tasks. Whereas users U4 and U5 are the member of the same team, which is team B, but they do not share tasks. The user U3 is participating in both teams (A and B) at the same time, since U3 has been assigned tasks common in both teams. Users U1 and U3 have been working on a mutual task T1 while users U3 and U4 have been assigned another same task which is T3. Users U1 and U2 are part of the same team but they do not share any task of the team. This also shows the level of a collaborative relationship among users i.e. Mutual, Member, and Colleague. Several colors have been used in Fig.3 to represent different collaborative relationships. Such as, green color represents a mutual relationship, blue color represents a member relationship and red color represents a colleague relationship.

The said collaborative relationships also determine the level of sharing information among users. As mentioned earlier, the highest level of collaboration is "Mutual" which is being a member of the same team and being assigned the same task. In Fig. 3, we can see that users U1, U3, and U4 have been connected with green arrows because they have a mutual relationship. Medium level of collaboration is "Member" in which users may be a part of the same team but do not share the same task. It can also be seen in the figure that users U1 and U2 have a relationship as a member as well as users U2, U3, and U4, U5 have been assigned member relationship. Moreover, the lowest level of collaboration is "Colleague" which is neither being a member of the same team nor having assigned the same task. According to Fig. 3, the user U1 of team A and the user U4 of team B have colleague relationship, similar relationship exists between users U2 and U5.

Since we have a dynamic CWE, changes can occur anytime. According to the Fig. 3, Team A and B were having a mutual task T1. When the task T1 is finished, the collaborative relationships of participating users are also changed. This has been shown in Fig. 4. Now, users U1 and U3 do not have mutual relationship so the green arrow joining them before has been removed. But, users U3 and U4 still share the same task (T3) and are members of the same team (B) they still have a mutual relationship that is why they are still connected with a green arrow. Likewise, users U2 and U3 were sharing member relationship in Fig. 3, when the task T1 is finished user U3 is not a member of the "team A" anymore so users U2 and U3 are not sharing member relationship anymore. This is how dynamic adaptation takes place in our system. As and when a change occurs in CWE, related policies are changed accordingly by the system. The scenario explained above has also been tested on a prototype model of the system. The results have been explained in next section.

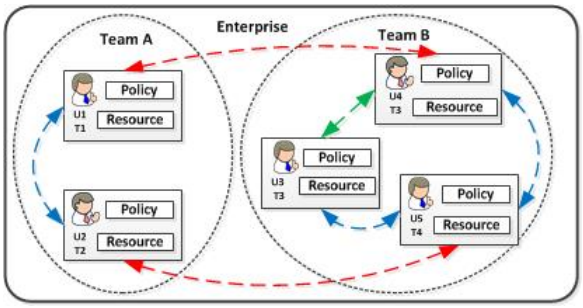


Fig. 4. Dynamic adaptatin in a CWE (After) [3]

IV. Results and Discussion

In order to test the proposed model, we have taken some empirical data set in which a different number of teams, tasks, users, roles have been taken to evaluate different scenarios such as finish team, finish task, task revocation, team revocation. There are 16 users and 20 roles, each user has been assigned one or many roles at a time and on the basis of his roles, he has assigned related team or task. There are 8 teams and 9 tasks. There are 4500 access control permissions which are used in sets of 1500, 3000, and 4500 permissions to test the system with increasing number of permissions. Each aforementioned scenario has been evaluated through these numbers of permissions separately. Details of said scenarios are provided as follows.

Finish Team:

In the proposed scenario, there are a number of teams in which different users are participating to accomplish different tasks. So, whenever any team is finished, its related permissions and information are also removed using the dynamic adaptation system i.e. whenever a team finish will occur the system will automatically accommodate related changes. For 1500 permissions, upon finishing a team, there are 426 permissions that have been changed. For 3000 permissions, in a finish team scenario, there are 576 permissions that have been changed or removed. Similarly, in the case of 4500 permissions, there are 720 permissions which have been changed. This has also been shown in the given Fig. 5. We have compared the number of permissions affected when a team has been finished in three different number of permissions sets. We can see that as the number of permissions increases the system adapts the change accordingly.

Finish Task:

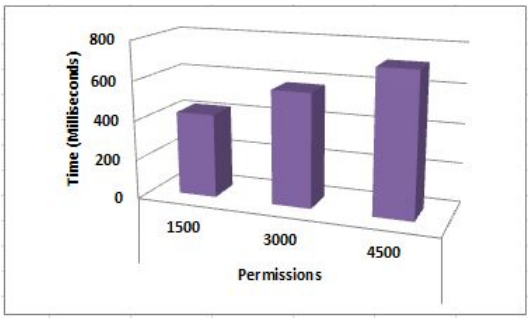


Fig. 5. Adaptation in Finish Team

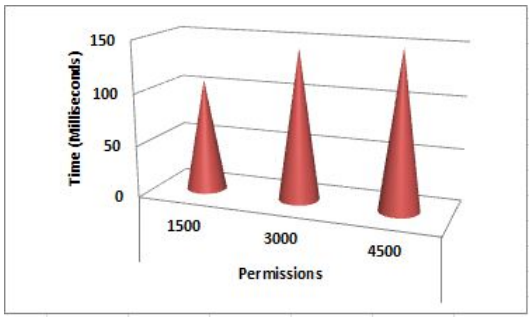


Fig. 6. Adaptation in Finish Task

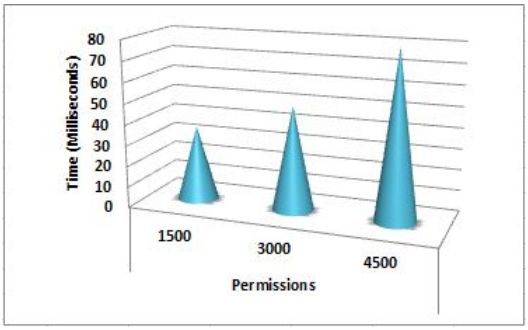


Fig. 7. Adaptation in Revoke Task

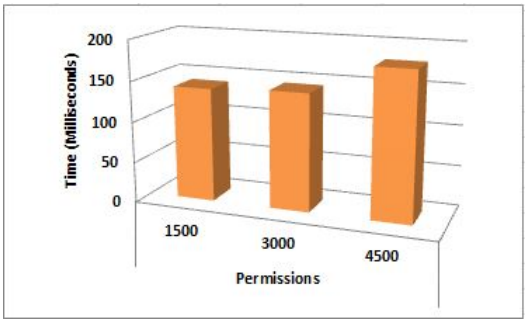


Fig. 8. Adaptation in Revoke Team

In this scenario, the case of finishing a task is discussed. In other words, whenever any task assigned to any team will be completed/finished, its related access permissions will also be removed. A single task may be assigned two multiple users but when the task has been completed all associated changes will be accommodated for each and every concerned user. This is how adaptation takes place in this scenario. In Fig. 6, we have shown how adaptation has taken place with different number of permissions by outlining change in the number of permissions due to task finish. We can see in the Fig. 6, for 1500 permissions, the number of permissions that have been affected are 108, for 3000 permissions the number is 144 and for 4500 permissions it reaches to 150.

Revoke Task:

Task revocation means a task which has previously been assigned to a user is withdrawn from that user. There are some cases in which some users might not be able to produce expected results in a task. In this situation, tasks can be revoked from users. Our system is capable of acclimatizing changes associated with that task. In Fig. 7, we have shown how many numbers of permissions have been affected due to revocation of user assigned task. It shows that, whenever a user is revoked his assigned task, his related permissions are also removed from the system, regardless of what the number of permissions is. Such as for 1500 permissions, there are total 36 permissions that have been affected. For 3000 permissions, there are 50 permissions that have been changed or removed. Similarly, for 4500 permissions, we have 80 permissions that have been removed.

Revoke Team:

Another scenario is revocation of a team from a user. This means, due to any circumstances a user can be considered incapable of being a part of a team. So the user will be removed from corresponding team. This scenario is different from finish task because in that one, a user will be finishing task assigned to him. While in this scenario, it's not the case, a user may not be able to finish his task and he may be released. In this situation the permission which he has already created will also be removed by the system automatically. The Fig. 8 shows the numbers of permissions changed due to revocation of a user assigned team. This is shown for three different sets of permissions. For 1500 permissions, there are 140 permissions that are changed. For 3000 and 4500 permissions there are 144 and 180 permissions that are changed, respectively.

Running Time Comparison:

We have also compared run time taken by our system for all said scenarios. Fig. 9 shows the comparison graph displaying time taken by each scenario.

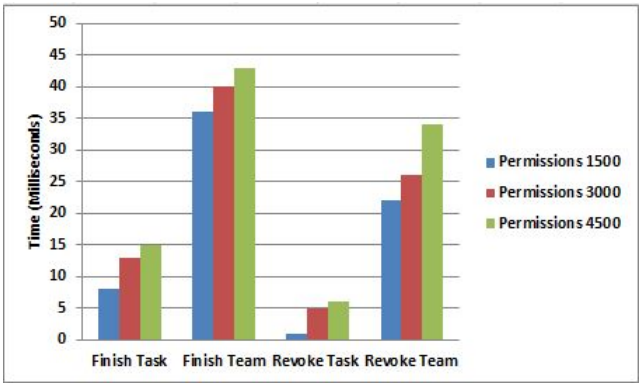


Fig. 9. Comparison of running time for different scenarios

In case of 1500 permissions, time taken by Finish Task is 8 milliseconds, while for Finish Team, Revoke Task and Revoke Team it is 36, 1 and 22 milliseconds respectively. In case of 300 permissions Finish Team has taken maximum time which is 40 milliseconds. While, time taken in Finish Task, Revoke Task and Revoke Team is 13, 5 and 26 milliseconds respectively. In case of 4500 permissions, maximum time has been taken by Finish Task which is 43 milliseconds. However, Finish Task, Revoke Task and Remove Team has been taken time as 15, 6 and 34 milliseconds respectively. We can see that "Finish team" has taken the maximum time among others while "Revoke task from user" has taken the less time among all. This is because a team may contain a good number of users and each user will create different number of permissions, so when that team will be finished all related permissions will be removed and that will be a large number. Similarly, a user will individually be assigned any task and he will create his task related permissions accordingly. So for that task, only those permissions will be removed which are related to that particular user only.

With the help of all results stated above, we have presented how the proposed system efficiently and effectively implements dynamic adaptation in a CWE so that users as well as administrators do not have to worry about managing their large number of access permissions manually.

V. Conclusion

We have implemented an Adaptive DySP-RBAC model in which sharing and privacy of user information is presented using two types of sharing control policies. The Enterprise-defined policy is used to prevent user's shared information (such as their teams, tasks, roles etc.) from being revealed to unauthorized users. Similarly, the User-defined policy is used to prevent user's personal information (i.e. his location, personal details). Users create permissions to control access to their resources. A large number of such access permissions are hard to manage. For this purpose, our system executes in a dynamic working environment in which rules are dynamically adapted at runtime as and when a relevant change is detected. Our system successfully adapts the changes caused by such cases, for example, when a team or task is finished, a user may be revoked a team or task which has been assigned him previously. A prototype of this model is implemented and it has been evaluated for dynamic adaptation of access control rules. In future, this work is to be extended for more than one enterprise, since we have focused on within a single enterprise scenario.

Acknowledgment

This research is funded by Higher Education Commission (HEC), Pakistan, through the project "Information Privacy in Collaborative Environments" approved for CIIT, Islamabad.

References

[1] H. Patel, M. Pettitt, and J. R. Wilson, "Factors of collaborative working: A framework for a collaboration model," Appl. Ergon., vol. 43, no. 1, pp. 1-26, 2012.

[2] A. M. Townsend, S. M. DeMarie, and A. R. Hendrickson, "Virtual teams: Technology and the workplace of the future.," Acad. Manag. Perspect., vol. 12, no. 3, pp. 17-29, Aug. 1998.

[3] A. K. Malik and S. Dustdar, "An intelligent information sharing control system for dynamic collaborations," Proc. 8th Int. Conf. Front. Inf. Technol., p. 30:1-30:6, 2010.

[4] A. K. Malik and S. Dustdar, "Enhanced sharing and privacy in distributed information sharing environments," Proc. 2011 7th Int. Conf. Inf. Assur. Secur. IAS 2011, pp. 286-291, 2011.

[5] W. Naeem, M. A. Shah, and A. K. Malik, "Privacy-Preserving in Collaborative Working Environments," in IOARP, 2015, pp. 18-19, March 2016.

[6] B. Schilit, N. Adams, and R. Want, "Context-Aware Computing Applications," in 1994 First Workshop on Mobile Computing Systems and Applications, pp. 85-90, 1994.

[7] Y. Rogers and J. Ellis, "Distributed cognition: an alternative framework for analysing and explaining collaborative working," J. Inf. Technol., vol. 9, no. 2, pp. 119-128, June 1994.

[8] M. A. Martinez-Carreras, A. Ruiz-Martinez, F. Gomez-Skarmeta, and W. Prinz, "Designing a Generic Collaborative Working Environment," in IEEE International Conference on Web Services (ICWS 2007), pp. 1080-1087, 2007.

[9] H.-L. Truong, S. Dustdar, D. Baggio, S. Corlosquet, C. Dorn, G. Giuliani, R. Gombotz, Y. Hong, P. Kendal, C. Melchiorre, S. Moretzky, S. Peray, A. Polleres, S. Reiff-Marganiec, D. Schall, S. Stringa, M. Tilly, and H. Yu, "inContext: A Pervasive and Collaborative Working Environment for Emerging Team Forms," in 2008 Intl. Symp. Applications and the Internet, pp. 118-125, 2008.

[10] D. S. Allison, A. Kamoun, M. A. M. Capretz, S. Tazi, K. Drira, and H. F. ElYamany, "An ontology driven privacy framework for collaborative working environments," Int. J. Auton. Adapt. Commun. Syst., vol. 9, no. 3/4, p. 243, 2016.

[11] A. Mohamad, N. Yusoff, S. Aris, D. Bismo, and M. Regan, "Dare to Change: An Approach to Implement Enterprise Level of Real Time Well Solution for Collaborative Working Environment," in IADC/SPE Asia Pacific Drilling Technology Conference, 2016.

[12] T. E. Bihari and K. Schwan, Dynamic adaptation of real-time software, ACM Trans. Comput. Syst., vol. 9, no. 2, pp. 143-174, May 1991.

[13] J. Keeney and V. Cahill, "Chisel: a policy-driven, context-aware, dynamic adaptation framework," in Proc. POLICY 2003. IEEE 4th International Workshop on Policies for Distributed Systems and Networks, pp. 3-14, 2003.

[14] A. Kapadia, G. Sampemane, and R. H. Campbell, "Know why your access was denied: Regulating feedback for usable security," Proc. ACM Conf. Computer and Communications Security. pp. 52-61, 2004.

[15] H. Koshutanski and F. Massacci, "Interactive access control for web services," IFIP Intl. Inf. Secur. Conf. , vol. 3, no. 3, pp. 1-16, 2004.

[16] Y.-G. Kim, C.-J. Mon, D. Jeong, J.-O. Lee, C.-Y. Song, and D.-K. Baik, "Context-Aware Access Control Mechanism for Ubiquitous Applications," pp. 236-242, 2005.

[17] A. Toninelli, R. Montanari, L. Kagal, and O. Lassila, "A semantic context-aware access control framework for secure collaborations in pervasive computing environments," ISWC'06 Proc. 5th Int. Conf. Semant. Web, p. 14, 2006.

[18] E. de Matos, L. A. Amaral, and F. Hessel, "Context-Aware Systems: Technologies and Challenges in Internet of Everything Environments," Springer Intl. Pub., pp. 1-25, 2017.

Autonomous Software Installation using a Sequence of Predictions from Bayesian Networks

Behraj Khan, Umar Manzoor, Tahir Syed
National University of Computer and Emerging Sciences, Karachi, Pakistan

***Abstract*—The idea of automated installation/un-installation is a direct consequence of the tedious and time consuming manual efforts put into installing or uninstalling multiple software over hundreds of machines. In this work we propose what is to the best of our knowledge the first learnable method of autonomous software installation/un-installation. The method leverages text classification using as data textual guidelines given for users on the installation window. This is used to arrive at the $Next/Pause/Abort$ decisions for each installation window using multiple classifier schemes. We report the best results using a full Bayesian Network with accuracy level of 94%, while Naïve Bayes and rule-based inference accuracy was 42% and 88%. We attribute this to the sequential nature of the Bayesian network that corresponds to the sequential nature of natural language data.**

***Keywords*—Multiagent System; Machine Learning; Software installation/un-installation**

I. Introduction

With the evolution in distributed environments both network size and complexity have increased substantially. The task of maintaining and improving a network generally requires multiple software installation and un-installation processes. This is greatly dependent on manual effort and therefore scales poorly. Consider the whole process of installation/un-installation which starts with initializing and running the setup wizard, at every screen/step reading the message/text displayed, analyze that text and then choose the appropriate action. This process is repeated till the installation/un-installation finishes successfully. Some software provides the option of silent or unattended installation in the form of set of switches [4]. *Silent installation* is the one which does not require interaction with the user at every screen/step to proceed further. Rather, it will be initialized by the user in the beginning and rest of the action sequence will be performed by the application itself.

However, silent installation proves to be a non-desirable choice for a distributed environment. This kind of silent installers are software- and vendor-specific. The degree of required human computer interaction for accessing every machine and initializing the setup is also very high. In addition to that this silent installation feature is provided for installing software only; to uninstall software no such facility is supported. Not all the software provides the silent installation option. The focus had majorly been upon automated installation on a standalone system within a non-distributed environment.

Automating installation/un-installation in a distributed environment where there might be several sub networks interlinked to each other is a complex and difficult goal to achieve as it increases the size and complexity of the autonomous framework. It may also involve the complexity of finding path over the relevant sub network to reach the destination node, than transferring files to that node. The process of safe and reliable transfer of files precedes the verification and initialization steps which are followed by running the setup and finally it concludes with the update in the system directories and sending the acknowledgment.

Some frameworks for silent/automated aid for software installation / un-installation has been proposed by U. Manzoor and S. Nefti (NDMAS [1], ABSAMN [2] and SUIPM [3]). These models were based upon rule based analysis of the text which is not very efficient and effective in unknown environments.

The installation/un-installation procedure is traditionally a resource dependent task and requires much of manual aid. To lessen this manual dependency and effort this task could be assigned to multiple intelligent agents with efficient learning and text classification capabilities. We automate the process of installation through setup wizard into silent installation. In this paper we propose a framework for installing a software without human intervention, i.e. by automating the process of the so-called silent installation/un-installation. *This is done by interpreting the text that appears on installation screens and thereby predicting the action to be taken next.* In our proposed framework the installer agent (the artificially-intelligent agent which will install software autonomously) activates once the particular installation file is run. The agent activated it extract all the information on application window and on UI controls (e.g. text on buttons like "'Next"', "'Back"', "'Cancel"'). That helps the installer agent classify the text on the windows using classifiers such as a Bayesian network and thereby decide to continue installing the particular software or to quit installation. To summarize, our method works in the following manner:

- Access installation package and network nodes' resources,
- Read the text on an installer window,
- Classify the text using a number of classifiers and decide between classes 'Next' versus 'Cancel'.

Therefore, this work represents the use of text classification as a means to address a problem in distributed assisted software installation/un-installation.

The rest of the paper is organized as following: Section 2 presents the related work in the field, Section 3 presents the system architecture that would serve as a guide for the following sections that describe the algorithms that plug into components of this architecture, and in section 4 classification

models are presented. Section 5 and 6 will focus on experimentation and analysis followed by conclusion.

II. Related Work

The idea of intelligent installation/un-installation agents has received attention from a niche group within the machine learning community. Manzoor & Nefti implement multi-agent aided network monitoring and installation application like "SUIPM", in which they proposed silent unattended installation package manager which generates silent unatteneded installation packages before installation. SUIPM supports all kind of software installation over the heterogeneous setting on different nodes. SUIPM does not require any client software for installing a particular software. However the proposed method has no intelligence, requiring initial training for the operator for installing software. SUIPM generates its own packages for installation which may be twice in size of the original setup[1].

Manzoor & Nefti propose a method in Cognitive Agent for Automated Software Installtion "CAASI"[2] that rectifies the shortcomings on [1]. The proposed method is able to install a particular software intelligently and the setup size of particular installing software remain as original setup, but the proposed method still required training before installing a software.

Manzoor & Nefti propose "ABSAMN"[3], which is an agent-based architecture for activity monitoring over the network. The proposed method watch activity monitoring like user activity, node level activity, and internet monitoring autonomously.

Manzoor & Nefti proposed a framework Smart network installer and tester for installing software autonomously "SNIT". SNIT is also agent based and motivated by An agent based system for activity monitoring on Network "ABSAMN". The proposed method supports unattended installation over the network intelligently. The proposed method install a software intelligently without any kind of training. SNIT do not require any specific kind of setup before installation.

Herrick & Tyndall [22] proposed a method Sustainable Automated Software Deployment Practices for automating software installation SASDP. but the proposed method have no intelligence and user have to watch the installation process till completion. SASDP requires a particular MSI software for running the application, and also give manual installation facility to user.

In the above applications, rule-based text classifiers were implemented for the learning of agents. The short comings of rule based classifiers are observed to be large and slow knowledge-base along with the inefficient performance in unknown environments.

Installation agents may learn through text classification schemes. Text classification is a challenging domain because of a vast number of attributes in the form of words. Many of the efficient text classifier models are designed for the domains with relatively short vocabulary set. But the more practical scenarios usually consist of complex and large vocabulary set (more than thousand words). Many text classification schemes have been proposed and implemented for large vocabulary sets (a detailed comparison has been presented in the table 1). For our research we will classify the text using the Naïve Bayes and Bayesian belief networks.

In their work, Domingos and Pazzanihas [4], [5] concluded that Naïve Bayes based text classification promises very optimistic results with the constraints of zero significance level of the probability calculated by the Naïve Bayes . Some literature shows the implementation of Naïve Bayes classifiers using a binary feature vector. The implementation does not capture the total occurrences of a word in the text rather it captures the probability of all attribute values (both present and absent). A binary feature vectors represents absence or presence of every word.

Therefore, the text is modeled as an event with related binary attributes. This category of implementation is called multivariate Bernoulli Naïve Bayes. This model is closest to the traditional Naïve Bayes. The shortcomings of this implementation is its inability to exploit word frequencies in the text and its suitability for tasks with fixed number of attributes (small documents).This approach has been applied for various text classifiers.

The other Naïve Bayes text classifier implemented is the multinomial model or the standard Naïve Bayes text classifier [6], [7], [8], [9].This model focuses on the number of occurrences of a word, showing the words as events. The order of appearance becomes insignificant and the probability of a particular word becomes significant.This approach has also been applied to multiple domains of text classification like speech recognition and spam filtering.

The standard Naïve Bayes text classifier does not show better performance in comparison with other statistical learning methods like support vector machines [11], nearest-neighbor classifiers [8], and boosting [10]. However, the shortcomings of this model are its rough parameter estimation.Latest researches have focused on exploiting the efficient and simple implementation scheme in multiple practical domains of text classification like web mining and news article classification.

A Bayesian belief model focuses on relationship among the attributes. It could also be applied with different statistical methods to gain enhanced performance and avoidance of data over fitness. It shows good results even if some data entries are missing as it models the dependencies amongst all attributes. This model could be applied to gain better understanding of the problem domain. It is widely used in different application areas of classification. These networks can be learned automatically on the basis of statistical analysis. Naïve Bayes is a special case of these networks. Many text classifiers based of Bayesian belief networks has been introduced [12][13] for multiple domains like disease and cancer diagnosis.

III. Proposed System

The proposed framework will work in a layered fashion/architecture. The top most layer in the hierarchy is the supervisor layer, that controls the whole framework. The supervisor layer controls the controller layer which controls the verification layer. The lowest layer in the framework is the installation/un-installation layer.

A typical distributed environment could be seen as a network of multiple networks. A supervisor agent is moni-

TABLE I. COMPARISON OF TEXT CLASSIFIERS

Text Classification Model	Application Domains	Advantages and disadvantage
Decision Trees [22]	Hierarchical distribution Skewed class patterns Predicate based, divide and conquer strategy	Simple knowledge representation Fast learning and qualitative analysis Inefficient when there is noise in the training data [5]
Rule based classifiers [24]	Based on simple rules to represent text categories Classification rules are defined manually Decision support systems	Easy to understand and modify Easy incremental update by other machine learning models Use of more than one feature values simultaneously Inefficient with the exponential growth of feature space Large number of training rules [25] Conflicting rules and low coverage Change in rules with the change in an environment
SVM Classifiers [26]	Supervised classification algorithms Partitioning of data space into different classes	No transformed space is required Consistent solutions Focuses on linear separators and robust over fitting applicable to binary classifications only frequent generation of zero values Time consumin [27]
Neural Network Classifiers [28]	Multi-Output Perception Learning algorithm (MOPL) Back-Propagation Neural Network (BPNN).	Quantitative analysis Complex knowledge representation, Slow learning and converges on local minima For long trainings it starts over fitting Multiple training runs are required for assessing the applied model [Yao and Zhi-Min et al., 2011],[Manning et al., 2009]
Naïve Byes classifiers	Attributes are considered as independent Web mining, news group classification	Simple, fast and efficient Cheap implementation cost
Bernoulli Naïve Bayes	Speech recognition, Web mining, Spam filter etc	Restrictive conditional independence, poor performance for strongly correlated data [5],[5]
Multinomial Naïve Bayes		lack of uniformity in training data [24] some attributes might remain less trained in comparison with other
Bayesian Belief networks	Attributes are considered correlated to each other Missing attributes could be computed	Fast and efficient Delivers better understanding of the domain knowledge

toring the whole framework at the master server level and multiple controller agents are assisting and coordinating with it .Whereas every controller agent is monitoring a separate sub network. Multiple file transfer agents will serve the purpose of IS-UIS file transferring over the path on the network from source to destination node under the supervision of their respective controller agent. The main and primary objective of the application is to design classifiers to support the installation/Un-installation tasks and to update the knowledge base of the system. The application/program to be installed/un-installed will be pre-processed by the system agents. This task involves gathering the information on every screen as given in figure 1 below:

The information on the screen is in the given format:

- Screen label
- Text/data
- Buttons/text boxes/ check boxes/ option groups along with their text ...

The data (above mentioned) will then be transferred to the classifiers and then data will be processed to make it ready for the classification task as given in figure 2 below:

The information on the software installation window will be extracted first from that window which contain standard amount of text, after extracting data will be pre-processed means that the words like (a, an, the, are, is, of, to, will) be removed from data and then will be given to classifier. The classifier would have a knowledge base on basis of which

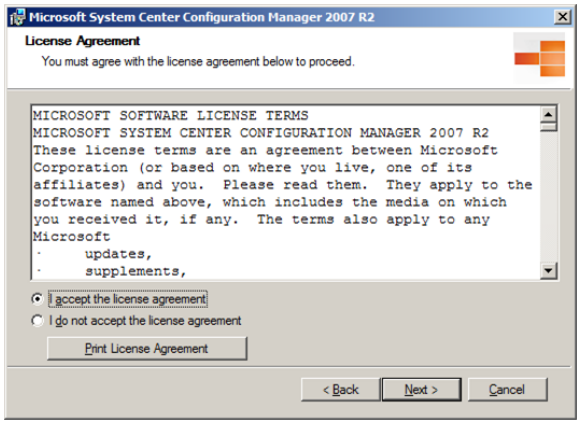


Fig. 1. Gathering Information

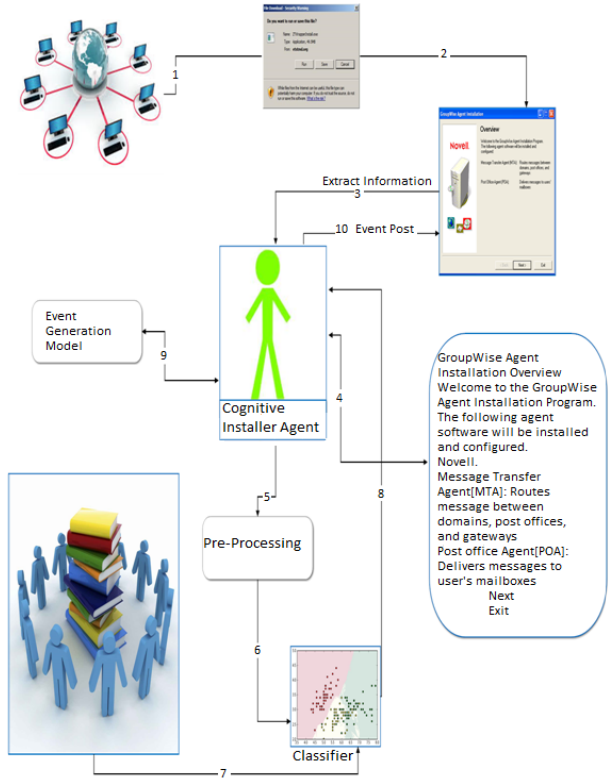


Fig. 2. System srchitecture

it would classify the text and would give the result to event generation model, event generation model would generate an event and then event will be posted.

A feedback module has been added to the application to enhance the accuracy of test cases and to achieve more accurate and promising results for the test data. The feedback module starts working after the completion of classification and it asks the user to help identifying the unidentified cases by giving the class labels as input as given figure 3 below:

The knowledge base is updated according to the user input and whenever new data is tested, the existing knowledge base is also referenced. The processing data will be divided into classes and attributes (tokenized) in order to find the probabili-

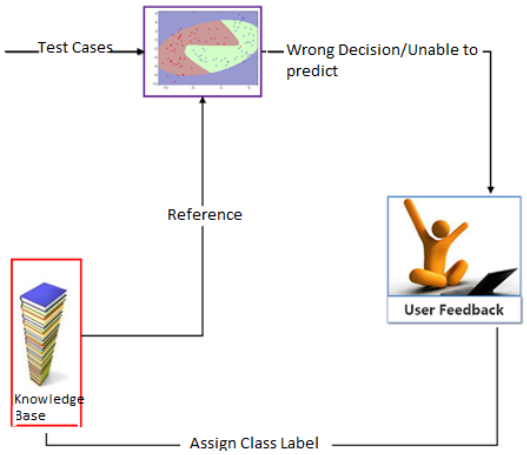


Fig. 3. Feedback Module

ties of relevant attributes according to their expected/classified class labels. A knowledge-base will be maintained to break down the attributes from the data. That knowledge base of data attributes will be used by the three classification algorithms to compute the probabilities and relationships. The classification task comprises of two steps:

- Classification of class label,
- Identification/selection of appropriate action.

Once a screen/attribute has been classified correctly, the objective is to identify the appropriate action/button selection for that particular screen/attribute. Knowledge-base will be maintained by the application to keep track of the expected desired options to be selected/opted by the user against each class to help identifying the action/button to be finalized.

A. Agent Infrastructure

In an autonomous environment the role and behavior of an intelligent mobile agent has always been an ideal choice to represent flexibility and reliability towards achieving the design objective in a distributed environment. The aim of this research is to present an agent based autonomous framework for software installation/un-installation in a distributed environment. The agent based autonomous software installation/un-installation framework will conduct the automated installation/un-installation task over a network. The intelligent software installation/ un-installation agents will be trained using Naïve Bayes and Bayesian Belief classifiers. The system consists of five agents:

B. Server Agent

Server agent is the main agent of the system it initializes the whole system. SA loads the network configurations which contains the IP address and name of the sub server and it also contains the range of the sub server network on which particular software to be installed.

C. Sub-Server Agent

Server Agent initiate sub server agent, then sub sever move to each sub server and perform the following steps. Sub Server

Agent load the configuration file and also load the Knowledge Base. Sub Server Agent is also responsible for creating and initializing File Transfer Agent and Installer Agent.

D. File Transfer Agent

File Transfer Agent is responsible to transfer the software to each node over the network. It distributes the file from a file server to a large number of machines over the network. FTA (File Transfer Agent) transfers the software into chunks and then merges it on each node.

E. Installer Agent

As the FTA completes their task then Installer Agent is initialized. IA contains the list of nodes on which particular software to be installed. Then it moves to first node in the list, loads the Knowledge Base, software profile and also the information passed by the Sub Server Agent. Before starting the installation, Installer Agent will fist check the constraints like space available in the target drive or not.

F. Verifier Agent

After initializing verifier Agent (VA), VA moves to each node and verify that the product installed on each node. After verifying the information about product VA will run the application in back ground and then moves to other node. The method proposed by Umar et al. [19] installed software over the heterogeneous network autonomously. SNIT installs the software(s) over the network autonomously, but the proposed method is rule based and failed in some cases. To overcome this problem we proposed a method in which we extend the SNIT and proposed autonomous installation using Bayesian Belief Network. In this thesis I only modify the Installer Agent of SNIT and named them Cognitive Installer Agent (CIA). CIA installs software over the network autonomously using Bayesian Belief Network. CIA is more intelligent than installer agent, and it can handle all type of software installation and un-installation without any user interaction.

IV. Text Classification Models for the Installer Agent

Different installation packages may have differently-worded text showing on installation wizard screens, but at the same time there are many similarities among the text displayed, beginning with the standard text on UI controls like buttons and keywords that on their own may be sufficient for classification under the iid ssumption that for instance Naïve Bayes takes. So for the task at hand we work with a number of successful classifiers from the text retrieval community like Naïve Bayes, Bayesian Belief Network, and rule-based classifier and compare their performance.

A. Naïve Bayesian classifier

Naïve Bayesian classifier are used commonly for text classification because of its success in this field and its ability to work with limited amounts of data. In Naïve Bayesian classifier we attempt to make a probabilistic classifier which is based on molding the fundamental word features in different classes. The main idea behind is to classify the text on the base of posterior and prior probability. In nave Bayesian classification we trained our classifier in such a way that the target class is known, and then we test the classifier for unknown target class. The classifier has to learn on the input data along with output data.

With the Naïve Bayesian classifier, we first calculate the probability of the target class in the document in the training set and then the probability of each attribute with respect to the target class, after that when the new data is provided to the classifier as test data then it assigns the data target class which probability is higher [3]. The Naïve Bayesian classifier gives the precise result when it is provided a large number of data set. Still it is often difficult that we have a large number of data sets. Beside these when the datasets are large then it required more space and have more time complexity for running. So the case in which we have small data set as is our problem then it give us quick result.

Classification of text is a becomes more challenging with the amount of data present. But classification of large amount of text is much easier than small amount of text document. Our problem is very challenging task because in this particular problem amount of text is very small as text on installation software text window which is very small in amount and consists of standard keywords as given in figure 4 below:

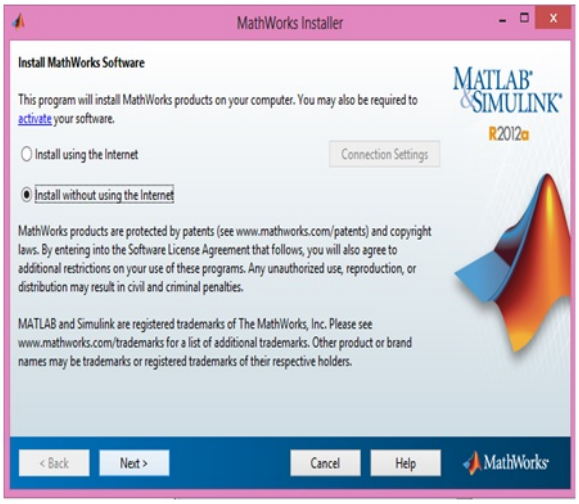


Fig. 4. Installation Window

The text extracted from this window is:

```
Caption
MathWorks Installer
Labels
Install MathWorks Software
This program will install MathWork products on your computer.
You may also be Required to activate your software.
Radio Buttons
Install using Internet
Install without using Internet
Button
Connection Settings
Labels
MathWork Products are protected by patents
(see www.mathwork.com/patents) and copyright laws.
By entering into the Software License Agreement that follows,
you will also agree to additional restrictions on your use
of these programs. Any unauthorized use, reproduction, or
```

```
distribution may result in the civil and criminal penalties.
MATLAB and Simulink are registered trademarks of The
MahthWorks, Inc. Please see www.mathworks.com/trademarks for
a list of additional trademarks. Other product or brand names
may be trademarks or registered trademarks of their respective
holders.
Buttons
Next
Cancel
Help
:
```

As we discussed in previous section that the text on the installation wizard window is small and consists of standard keywords.Now the problem is how to classify this text. Among different types of classifiers which we studied earlier the most suitable classifier for classifying the smaller and standard text is the Bayesian Belief Network which gives us better result than others.

After initialization of co-installer agent, filtered text (standard keyword) of particular window will be passed to classifier. The classifier represents the received information as given in figure 5:

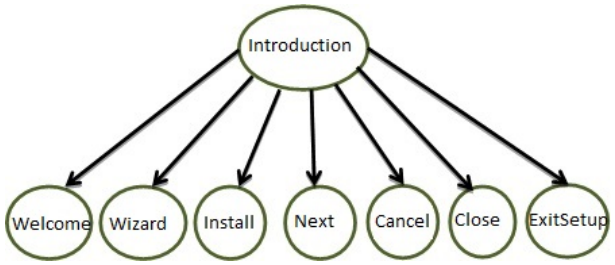


Fig. 5. Naïve Bayes

Above is the model for the very first window of installation process which is Introduction. After classification the result is given to event generation model and then other screen of setup wizard provided to the classifier autonomously and a description similar to the figure is generated for each of them.

B. Bayesian Belief

Bayesian Network is a probabilistic graphical model that represents a problem into set of random variable. In Bayesian Network model a problem is represented in the form of directed acyclic graphical model in which every node have probability. In directed acyclic graph, nodes represent variables and edges represent the probability among nodes. Each of variable have some variables (nodes) on which it depends that is called parent of them. Each of variable have their parents. Besides these Belief Network also has a probability table which shows the probability of each child with their parents. The conditional probability (CPT) of X and its parents is represented by a clique of size (k+1) in the graph and have $d_k(d-1)$ parameters. Learning process in Belief Network consists of two parts one is learning the network graph structure and other is learning the probability [13]. Bayesian Networks have three types of connections serial, diverging and converging.

In Bayesian learning the network is learned on trained data which contain output classes and then data without output classes are given to the network for testing[6].

The Bayesian network based classification is gaining popularity in almost every field of evolutionary sciences. It has been found very useful especially when the data is missing and incomplete. The Bayesian based classification has different forms. The focus of the research is performance measuring of Naïve Bayes, Bayesian Belief and Tree Augmented based classifiers and their comparison on the installation/un-installation data. Following is a brief description of the working of each of the three models along with the design details.

In our proposed design as the co-installer starts initialization it takes the extracted information from the window screen and provide it to the Bayesian belief classifier. Bayesian belief classifier represents the received information in network like structure and calculates the probability for the provided screen and passes it to event generation model. Following figure 6 is a brief model for the very first screen of setup wizard.

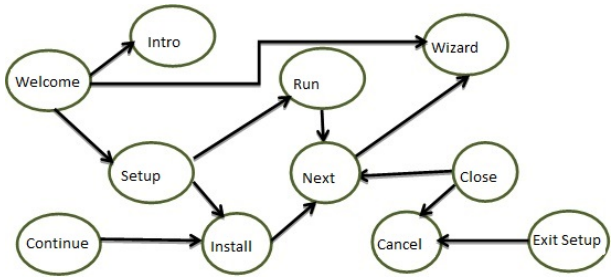


Fig. 6. Bayesian Belief

After classifying first screen, remaining windows of installation process are provided to the classifier in same fashion.

C. Tree augmented Bayesian

As co-installer initialized it takes preprocessed text of the very first window screen of the installation setup wizard window and provide it to augmented tree for classification. Augmented tree takes the received information from co-installer agent and represent it in tree like structure as shown in figure 7 below for classification.

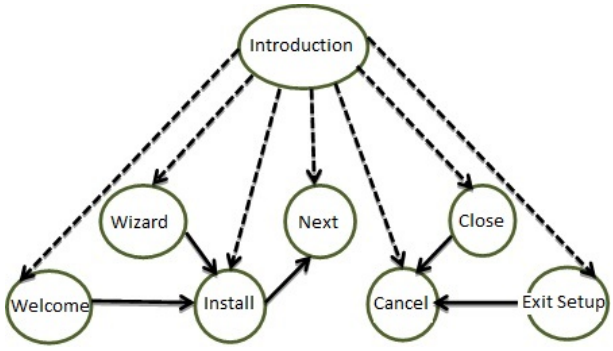


Fig. 7. Tree-augmented Bayesian Belief Network

Above is a brief model for very first model, same is generated for every window screen after receiving from co-installer agent.

V. Experiments and Discussion

The classification models were trained for exemplars taken from 25 different installation software programs and then tested for unknown test data set collected from 15 other software; a ratio of 62% training data set and 38% testing data set. Over all the results of the Bayesian Belief are found very promising. Following is a brief description that illustrates the performance of both the classification techniques with respect to different attributes.

Overall, the Naïve Bayes based classifier shows 42% accuracy rate with 84 exemplars classified correctly, rule-based showed 88% accuracy with 174 exemplars classified correctly, while the Bayesian Belief model classifier outperforms the others with an accuracy level of about 94% by classifying 187 cases correctly. The failure rate in terms of incorrect classification is 58%, 12% and 6% for Naïve Bayesian, rule-based inference and Bayesian Belief based models respectively.

In terms of training and testing data set; the Naïve Bayes based classifiers show 60% and 38% accuracy, rule-based inference shows 87% and 13% accuracy for training and testing data set respectively. Whereas the Bayesian Belief classifier performed very well by correctly classifying 96% and 92% of training and testing dataset respectively. The graph given below represents the overall performance presented by both the classification models.

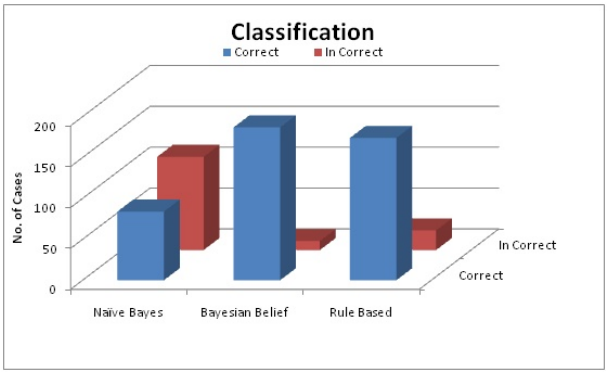


Fig. 8. Performance comparison of Naïve Bayes, Rule-based and Bayesian Belief Network

The dataset is categorized both by classifiers and also by rule-based inference. There were a total of 21 classes; each containing their relevant set of exemplars. We also measure the performance of the rule-based, Naïve Bayesian and Bayesian Belief classifiers for each class label. As discussed above the Bayesian Belief model was found to be more accurate and promising for each class (both training and testing data). The class design was based upon the output action selection attribute; Run, Next, Finish etc. As Bayesian Belief model focuses more upon attribute dependencies it performed well. On the other hand the Naïve Bayesian model assumes no dependencies amongst the data and depicted poor performance.

Given in Fig. 9 below is the graph showing the performance comparison of both the classifiers for 13 classes out of 21 as they demonstrate the good and true representation of all the classes (the similar trend was observed in the remaining classes as well). The total number of exemplars per class varied as the data was taken from installation software screens.

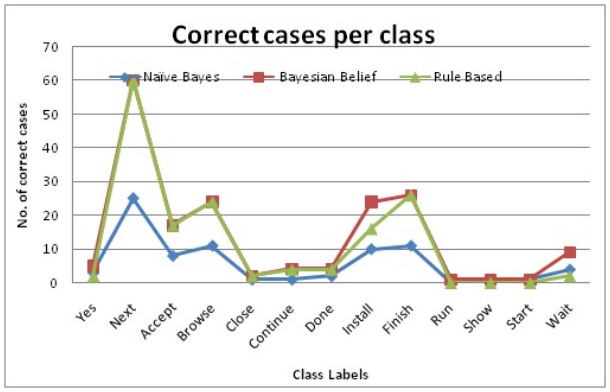


Fig. 9. Analysis of correct cases per class

As discussed above the screens/cases of a total of 40 software were used as the dataset for our research. Each individual screen while running software was considered as an individual and distinct exemplar. We collected a total of 198 exemplars for our experimentation. For a probabilistic classification paradigm this figure/number is considered to be sufficient. To measure the performance of any classification model the accuracy of results are the primary concern followed by the time constraints as the secondary point of focus. The time complexity of a Naïve Bayesian model is simpler and less than that of Bayesian Belief model and Rule Base in terms of time taken as well as the number of steps involved in the classification process.

Another comparison to judge the time based performance of both the classifiers and Rule Base system is to compare the time taken to classify/ predict single independent installation software. A total of 8 distinct software were selected and classified using both the models and also on Rule Base to measure the time taken to classify all the screens/dataset for each. Once again due to its simplicity of relationship and absence of inter dependency amongst the attributes the Bayesian Belief model showed better results.

Fig. 10 shows the graph representing a comparison of time efficiency of both the classification models and Rule Base for 8 different installation software.

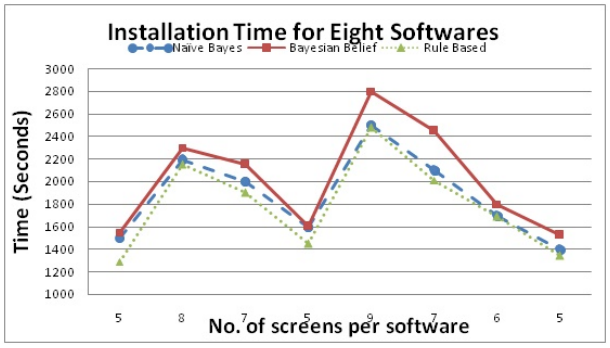


Fig. 10. Analysis of time

The whole research is focused upon the concept of automated installation; the support to this prime task is essential to be discussed. Hence their performance while running software was also judged along with measuring the outcome of both

classifiers, Randomly 5 different installation software were selected and the screens/data set from all of them were classified using both the models. The results of the Bayesian Belief model were amazing; in some software it showed even 100% performance by selecting the correct output class. Whereas the output of Naïve Bayes model and rule-based inference was disappointing as in some software it could not show more than 40% correct classfication. During installation Bayesian Belief Model is found to be resource-hungry in terms of time as compare to Naïve Bayesian and rule-based inference,which may affect system scalability once deployed, but our main focus is on accuracy of the number of softwares that are correctly installed. So Bayesian Belief Model is the method of choice available. We surmise that the major reason for its success is the fact that it does not ignore temporal relationships (i.e. one word following a certain number of others) between variables.

VI. Conclusion

In this chapter we compare the result of different classifiers like Naïve Bayesian, Bayesian Belief Network, augmented tree and rule base classifier. We also generalize in this chapter that which classifier performance is best than others. The proposed frame work is unique to the best of our knowledge as it will guide the installation/un-installation setup on the basis of information retrieved from the knowledge base; designed and modeled upon the Naïve Bayes and Bayesian belief classifiers. The autonomous framework is capable of handling unexpected situations and responds to the changes in the environment during the execution. In addition to this, it is capable of learning by adding new facts to the knowledge base. We also validate an implementable system architecture or framework, based upon a multi-agent environment interact and cooperate with each other in order to meet their design goal [21, 22].

References

[1] Charu C. Aggarwal and ChengXiang Zhai. 2011. A Survey of Text Classification Algorithms, Pattern Recognition, volume 37, Issue 3,(28 May 2011): pp 169-180, DOI: 10.1007/978-1-4614-3223-4-6.

[2] George Tsatsaronis and Vicky Panagiotopoulou. 2011. A Generalized Vector Space Model for Text Retrieval Based on Semantic Relatedness, (2009):Pages 70-78

[3] Gerhard Weiss. 1999. Multiagent Systems: A Modern Approach to Distributed Modern Approach to Artificial Intelligence (USA: MIT Press, 1999) ISBN 0-262-23203-0.

[4] H. Van Dyke Parunak , Paul Nielsen , Sven Brueckner and Rafael Alonso. 2007. Hybrid Multi-Agent Systems: Integrating Swarming and BDI Agents, Volume 4335, (2007), pp1-14.

[5] Han-joon Kim1 and Jae-young Chang. 2003. Improving Naïve Bayes Text Classier with Mo died EM Algorithm, Volume 2871, (2003): pp 326-333, DOI: 10.1007/978-3-540-39592-845.

[6] Ian H. Witten, Eibe Frank and Mark A Hall. 2011. Data Mining: Practical Machine Learning Tools and Techniques.(USA: Elsevier, 2011).

[7] James Allen, Nate Blaylock and George Ferguson. 2002. A Problem Solving Model for Collaborative Agents.,(July 15-19, 2002), DOI:10.1145/544862.544923.

[8] James Ingham. 1997., What is an Agent?, Technical Report 6/99 (1997),.

[9] Kotz, David, Mattern, and Friedsmann. 2000. Mobile agent applications, Volume 1882 (September 13-15, 2000) ISBN 978-3-540-45347-5.

[10] Keith S. Decker and Katia Sycara. 1997. Intelligent Adaptive Information Agents, Volume9, Issue 3, (1997/11/12): pp239-260. Klaus Dorer, Applications of Multi-Agent Systems in Logistic:Lecture 10. Hochschule Offenburg University of Applied Sciences.

[11] Manzoor and Nefti. 2010.,QUIET: A Methodology for Autonomous Software Deployment using Mobile Agents. Volume 33, Issue 6,(November 2010): Pages 696706, DOI: 10.1016/j.jnca.2010.03.015.

[12] Manzoor and Nefti. 2011 " Autonomous agents: Smart network installer and tester (SNIT)." Volume 38, Issue 1,(January 2011): Pages 884893, DOI: 10.1016/j.eswa.2010.07.066.

[13] Manzoor and Nefti. 2009. An agent based system for activity monitoring on network, Volume 36, Issue 8 (October 2009): Pages 10987-10994, doi:10.1016/j.eswa.2009.02.060

[14] Peter Stone and Manuela Veloso. 2000. Multiagent Systems: A Survey from a Machine Learning Perspective, volume 8, Issue 3, (June 2000): pp 345-383.

[15] Christopher D. Manning, Prabhakar Raghavan, Hinrich Schtze: An Introduction to Information Retrieval, page 181. Cambridge University Press, 2009

[16] Stuart J. Russell, Peter Norvig. 1995. Artificial Intelligence: A Modern Approach. (New Jersey: Prentice Hall, 1995), 07632.

[17] Sang-Bum Kim, Hee-CheolSeo and Hae-Chang Rim,Poisson Naive Bayes for Text Classication with Feature Weighting Volume 11,(2003): Pages33-40 doi:10.3115/1118935.1118940.

[18] Songbo Tan , Yuefen Wang and Gaowei Wu. 2011. Adapting centroid classier for document categorization, Volume 38, Issue 8, (August 2011): Pages 10264-10273, DOI: 10.1016/j.eswa.2011.02.114.

[19] Songbo Tan. 2006. , An effective renement strategy for KNN text classier, Volume 30, Issue 2, (February 2006): Pages 290-298 , doi:10.1016/j.eswa.2005.07.019.

[20] Taeho Jo, 2009. NTC (Neural Text Categorizer): Neural Network for Text Categorization, Volume 2, Issue 2, (28 October 2009).

[21] Zhao Yao and Chen Zhi-Min. 2012. ,An Optimized NBC Approach in Text Classification ,Volume 24, Part C (2012): Pages 1910-1914, DOI: 10.1016/j.phpro.2012.02.281.

[22] Dan R. Herrick and John B . Tyndall. 2013. ,Sustainable Automated Software Deployment Practices , (8 Nov 2013), http://dx.doi.org/10.1145/2504776.2504802.

[23] IEEE Intelligent System, 13(4):18-28, 1998. BibTeX entry. [9], Thorsten Joachims. Text categorization with support vector machines: learning with many relevant

[24] Phil Hayes, software that finds names in text, Intelligent Multimedia on Innovative Applications of Artificial Intelligence, p.49-64, May 01-03, 1990.

[25] Mining text data, CC Aggarwal, CX Zhai, Springer Science & Business Media.

[26] Inductive learning algorithms and representations for text categorization S Dumais, J Platt, D Heckerman, M Sahami, Proceedings of the seventh international conference on Information and

[27] Songbo Tan, Yuefen Wang, Gaowei Wu, Expert Systems with Applications: An International Journal archive, Volume 38 Issue 8, August, 2011, Pages 10264-10273, Pergamon Press, Inc. Tarrytown, NY, USA

[28] A re-examination of text categorization methods. Yiming Yang and Xin Liu. School of Computer Science, Carnegie Mellon University. Pittsburgh, PA 15213-3702

Generation of Sokoban Stages using Recurrent Neural Networks

Muhammad Suleman, Farrukh Hasan Syed, Tahir Q. Syed, Saqib Arfeen, Sadaf I. Behlim, Behroz Mirza
Department of Computer Science
National University of Computer and Emerging Sciences, Karachi, Pakistan

***Abstract*—Puzzles and board games represent several important classes of AI problems, but also represent difficult complexity classes. In this paper, we propose a deep learning based alternative to train a neural network model to find solution states of the popular puzzle game Sokoban. The network trains against a classical solver that uses theorem proving as the oracle of valid and invalid games states, in a setup that is similar to the popular adversarial training framework. Using our approach, we have been able to verify the validity of a Sokoban puzzle up to an accuracy of 99% on the test set. We have also been able to train our network to generate the next possible state of the puzzle board up to an accuracy of 99% on the validation set. We hope that through this approach, a trained neural network will be able to replace human experts and classical rule-based AI in generating new instances and solutions for such games.**

***Keywords*—stepwise cooperative training; generative networks; recurrent neural networks; Sokoban; puzzles; deep learning**

I. Introduction

Solving board-games and puzzles is often NP-hard and therefore arriving at a solution is not computationally feasible. Many strategies for solving board games have been developed using AI techniques. One important approach is to change board logic into a constraint satisfaction problem (CSP) and then solve the CSP using a theorem prover. Although most theorem provers perform optimizations for specialized cases, the generation of solution by theorem provers still remain computationally infeasible for higher order problems. In this paper we investigate whether we can speed up this process for solving board games using neural networks. The basic idea is to generate solved stages using the theorem prover and use this stage data to make training and testing data set for neural network. In other words, instead of solving games using a game tree we are interested to discover whether a trainable AI can replace the game tree. We therefore investigate if we could suitably train a neural network which can give us *exact* next stages after learning from training set provided by theorem prover. This is different from all the other generative neural network processes e.g. a generative adversarial network (GAN) in which an approximate result is acceptable such as a generated image that may be similar to an image in the dataset, in our case approximate result is not acceptable since it can result in deformed stages which lead to dead ends. The other crucial difference is in the way training is performed. In GANs, two networks compete against each other, where one generates an example that may or may not be similar to the examples in the training set's distribution, something which is decided by a discriminative network. Our discriminative network is replaced by the puzzle-solving theorem prover which is guaranteed to deterministically generate the next stage. We conclude that with proper sampling we can get an exact next stage of board games with high accuracy.

In our approach, we train a Recurrent Neural Network (RNN) to classify Sokoban board states as valid or invalid, and use one valid state to propose the next valid state, such that each state brings us closer to the solution. The problem of declaring a particular state of the Sokoban board may be easy enough if all valid and invalid states could be enumerated, but to decide a move onward from a particular board state enumerating all following states may not be a combinatorially appealing solution. There are several motivations for the work being presented:

1) the trained network could be used for training new players as well as provide in game help. The network can be used to identify any state as either valid or invalid. This in turn may be used to identify whenever a player reaches a blocking state (from which there is no solution) and guide the player to back-track some steps. We can also use the model to design new levels for the game. If our proposed solution works well for Sokoban, it may be extended for more complicated puzzles and problems.
2) methods proposed in the literature takes exponential time to reach a solution in worst case. For example, the theorem prover we used to generate the solution takes around 5 minutes to solve many 8×8 puzzle. Larger puzzles will take increasingly more time. If we could train neural networks to solve the puzzles, solutions could be reached a lot quicker. If the learning approach works well, the networks may be trained to learn puzzles of any size just by watching human players play. All other methods for solving board games with single player and step by step solution require extensive domain specific knowledge to implement.

Concretely, our objectives are as follows:

1) to create some invalid Sokoban mazes. Invalid maze means any maze which is unsolvable. We needed to do this as although valid mazes are available on the internet [1], invalid mazes are not.
2) to design and train neural network architecture that could detect whether any given maze was valid or invalid
3) to generate step by step solutions for valid mazes using the theorem prover[2]

4) to use the solutions (from objective 3) to train a neural network so it could produce solution steps for similar mazes.

To the best of our knowledge, ours is the first work that:

1) applies a deep learning method to general puzzle-solving, and
2) uses a hybrid generator-generator training where the generator network is trained against a game-tree generator.

II. Literature Review

Sokoban is a popular puzzle game. It was developed in the 1980s in Japan. The player is given a maze. The maze consists of walls, boxes, floor, holes and one player. The player has to push the boxes onto the holes. Any maze is valid if it satisfies certain conditions. For example, if there are an equal number of holes and boxes and some way for the player to push the boxes so they cover all the holes, the maze is valid. A number of approaches have been used to find optimal solutions to Sokoban puzzles. Some solutions have used graph algorithms such as breadth first and depth first search algorithms.

[3] has proposed solutions using Best-FS (Best-First Search),Iterative-Deepening A* search, and Genetic algorithm. Sokoban solutions can be viewed as expansion of trees of possible actions based on a certain state of the puzzle. Both Breadth first and Depth first graph algorithms will blindly traverse the trees hoping to find the solution. The Best-FS algorithm uses heuristics to decide which tree branch to take based on which path will move a box closer to a hole.

In [4], a hierarchical decomposition approach has been proposed where the problem is divided into a sequence of higher actions and elementary actions. Secondly, a database is maintained which keeps track of the mistake made by the algorithm giving it the ability to learn.

[5] propose a method for finding optimal solutions using Instance Dependent Pattern Databases. Using this method, the puzzle is decomposed into a goal zone, entrance and maze zone. A distance database is created from each box to a space called an entrance which is any square from which the goal may be reached.

In [6], an algorithm to generate Sokoban levels automatically has been described where they create an empty room, place goals in the room and find states farthest from the goal state, i.e. go from the end state back to a start state.

In all the papers discussed, knowledge about Sokoban and its rules is necessary to implement the solver or maze generator. We propose to circumvent the need to know the rules of a game for actual game-playing by the use of a generative neural network.

III. Methodology

We constructed a dataset of 700 valid Sokoban puzzles from [1]. We divided the dataset into halves so that one half $(i.e.350)$ represent the valid Sokoban puzzles and the other 350 are modified in a way that it leads to an invalid maze stage. Each puzzle has a dimension of 8×8. For puzzles which are smaller in any dimension, we pad with walls.

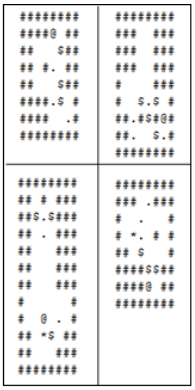


Fig. 1. Some valid maze configurations[8]

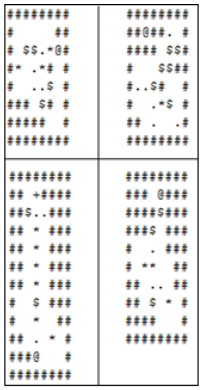


Fig. 2. Some invalid maze configurations

Each level of Sokoban contains 7 elements shown in Fig. 3. This results in a state space of 7^{64} stages. We then assign each maze element an integer value as shown in Fig. 4. We then convert all the elements to $1D$ as shown in Fig. 5.

Level element	Character
Wall	#
Player	@
Player on goal square	+
Box	$
Box on goal square	*
Goal square	.
Floor	(Space)

Fig. 3. Maze elements

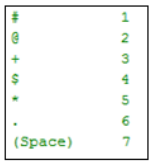


Fig. 4. Convert to integer values

There need to be a certain number of holes and a certain number of boxes in the maze, and the elements of the maze, i.e. the Walls, Boxes, Holes, and the Player need to be in a particular pattern for the maze to be valid. Some of the valid and invalid configurations of maze are shown in Fig. 1 and Fig. 2 respectively. Furthermore, since the player can only move in four directions (up, down, left, right), each state is the result

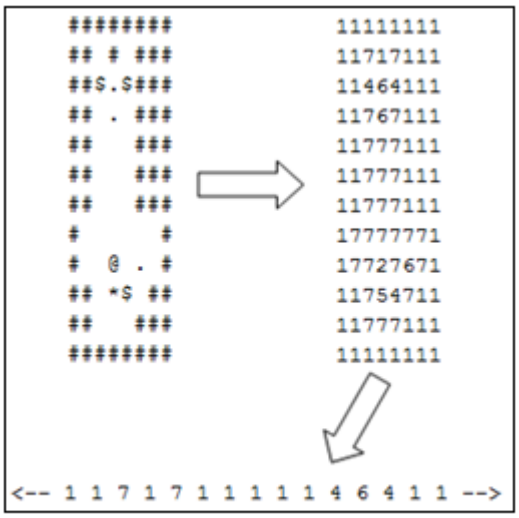


Fig. 5. Conversion of data to 1-D

TABLE I. MODEL LAYERS EXPERIMENT 1

Layer no.	Layer type	No. of Neurons
1	LSTM	128
2	Fully connected feed forward	2

of simple changes to the previous state, and is dependent on the layout of the maze elements shown in Fig. 3.

Our experiments are done on Keras with a Tensorflow backend. We used a special type of RNN called Long-Short Term Memory (LSTM)[7]. The LSTM provides a mapping betweennsequences of length 1. The input to the network at a given time stamp is a board state, and the predicted output is the next board state. For adversarial training, we limit the use case to Sokoban and the theorem proven in [2] which outputs a unique board state given the previous one.

IV. EXPERIMENTS

A. Experiment 1 - Prediction of Valid / Invalid state

There are many constraints which need to be fulfilled for a Sokoban puzzle to be valid. E.g. The number of boxes should be equal to number of holes. In other solvers these constraints are explicitly checked using if else statements. In deep learning approach, the network learns all of these just by viewing examples, i.e. it frees programmer from writing explicit constraints as if/else checks. Before coming to final experiment for prediction of valid / invalid state, we did a series of experiments to verify that our network can actually learn all the constraints for finding valid / invalid state. Input and Output of this experiment are shown in Fig. 6. Following are brief descriptions of our prior experiments.

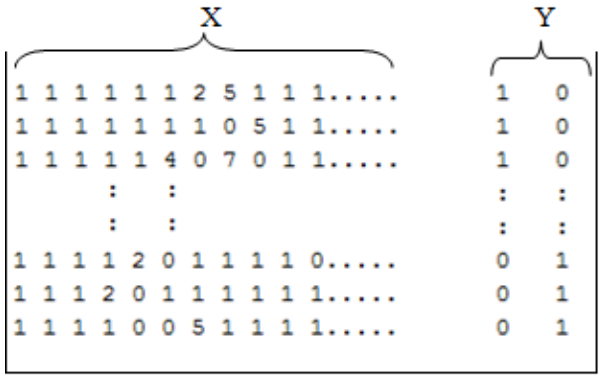


Fig. 6. Input and Output to the model

In our first experiment we made a Sokoban board with only three symbols, Box = '$', Goal square '.', Floor '(space)', and 16 squares. Each square can contain any of the three symbols. We marked a board as invalid if it contained unequal number of '$' and '.'. Our state space contained $3^{16} = 43046721$ examples. Out of these examples we randomly selected 2000 examples for our training data. After training our network up to 99.99% accuracy we tested our model on 50 unseen examples. In almost all experiments, (49 out of 50) unseen examples were correctly classified.

We subsequently experimented on an increasing number of squares and training examples:

TABLE II. RESULTS: B=BOX, G=GOAL, SQ=SQUARE, SP=SPACE, F=FLOOR, W=WALL

No.	Size	Symbols	Exp.	State space	Acc.	Cor. classified
2	64	B, G, Sq, Sp	8000	$3exp(16)$	99.99	$49/50$
3	128	B, G, Sq, Sp	8000	$3exp(128)$	99.99	$46-48/50$
4	16	B, G, Sq, F, W	2000	$4exp(16)$	99.99	$49/50$
5	64	B, G, Sq, F, W	4000	$4exp(64)$	99.99	$49/50$
6	128	B, G, Sq, F, W	8000	$4exp(128)$	99	$46-48/50$

TABLE III. MODEL LAYERS EXPERIMENT 2

Layer no.	Layer type	No. of Neurons
1	LSTM	128
2	Repeat Vector	
3	LSTM	128
4	LSTM	8

We repeated above prior experiments for 3, 4 and 5 relations and results are same. This showed that our network can learn up to 5 relations of length 128. More results can be produced by performing further experiments.

In our final experiment we have 7 symbols Wall '#', Player '', Player on goal square '+', Box '$', Box on goal square '*', Goal square '.', Floor '(Space)'. We downloaded 696.txt from [1]. This file contains 696 valid sokoban puzzles. We randomly selected nearly 350 sokoban puzzles out of the 696 puzzles. To make invalid puzzles we made little changes in each of 350 valid puzzles. Finally, we gave these 700 examples to our network along with their respective labels. Around 27-29 out of 30 unseen examples are correctly classified after training our network up to 99 percent.

B. Experiment 2 - Next-state predictor

In this experiment we downloaded 54 puzzles of 8×8 from $696.txt$ [1]. We gave every puzzle to the theorem prover[2]. The input and output of theorem prover is shown in Figs. 7 and 8 respectively. For valid mazes, the theorem prover generated step by step solutions. Fig. 8 demonstrates how the theorem prover generated step by step solutions.

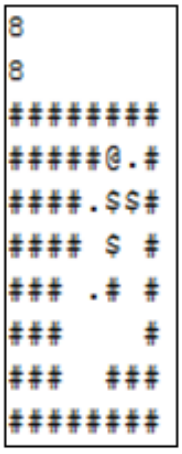


Fig. 7. Theorem Prover input

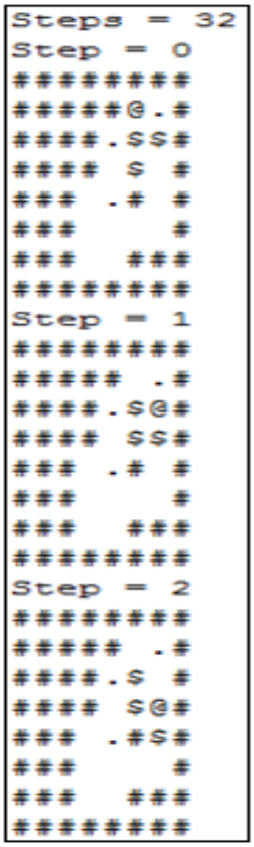


Fig. 8. Theorem Prover output

Our basic idea is to get board configuration in step 2 as the output of board configuration 1. For example, the output in step 2 becomes input for step 1. If our puzzle is solved in 32 steps than we will get 31 training rows for our network {1-2, 2-3, 3-4, ... 31-32}. Both X and Y contain 64 characters as shown in Fig. 6. The model of our network is shown in Table III. Each puzzle out of 54 puzzles gets solved in 16 to 64 steps. As a result we get nearly 3000 rows for our initial experiment. After achieving 99 percent accuracy on training data, we tried to solve whole puzzles using our network. We randomly chose a single puzzle from the 54 puzzles that we used for training set. This was a puzzle that was solved by theorem prover in 64 steps. 57 out of these 64 steps are correctly guessed by our network. For the other7 steps it gives an error of maximum 2 elements out of 64 elements in the maze. Next we fed an unseen puzzle to the network. The theorem prover had solved this puzzle in 64 steps as well. Although our trained network was unable to guess any step up to complete 64 places correctly, it guessed a maximum of 4 elements incorrectly out of the total 64 elements. Most of the time, the error was only at 2 places out of 64 places. We think this shows that 54 puzzles were not enough for solving Sokoban 8x8 puzzles. We will discuss how to increase problem instances for our problem in sections 5 and 6.

V. Results and Discussion

Experiment 1 shows that we can find valid/invalid configuration for many board game with very few examples. Many Sokoban solvers which use A*s search algorithm get stuck when given an invalid sokoban puzzle with just 1 unmoveable box, they keep on searching the whole state space for solution. Our proposed solution is very simple. We train a network which just guesses whether a given problem is valid or invalid. If it is invalid, we won't waste time in trying to find a solution.

We conduct our Experiment 2 to provide an efficient solver for board games like Sokoban, Colour Bridge and other similar games. The basic motivation for Experiment 2 is that once trained properly, neural network can guess solution to any Sokoban puzzle within seconds. In comparison, when we give random examples of 8x8 sokoban puzzle to any current solver which uses A* algorithm or theorem proving technique, it normally takes atleast 60 seconds to solve a puzzle, and for many 8x8 puzzles current solvers get stuck altogether. We have not experimented on 9x9 or bigger instances of the puzzle because most current solvers get stuck on these problems. As a consequence, we are unable to generate the required amount of train data.

There is a direct analogy between all the sequence to sequence learning tasks. For example puzzle solving by our method has direct analogy with question-answering task from supervised learning. Both are sequence to sequence learning, both have vocabulary. The difference is that the vocabulary is very small in case of a Sokoban puzzle, i.e. just 7 words, while in question-answering context, it is atleast 32 (as in the case of babi_rnn (facebook data)) for which an RNN gives 99% accuracy for 1 word answers. Here we are trying to extract a 49 or 64 word answer, so the error is expected. Bigger dataset which expect answers of 10 or more length are not extracted exactly i.e. giving 81% or less accuracy even by state of the art results [squad references]. what we are trying to discover is that if vocabulary size is small, is it possible to extract exact answers of size 49 or greater. We conclude that for the sample size we used in our experiments, we are able to extract answers with an error of 2 characters most of the time. For question- answering task, an answer with an error of only two characters can be a most ideal answer, but for puzzle solving we will need exact answers. We keep proper method of sampling from puzzle state space as future work since solution to the puzzle solving problem lies in proper sampling according to our analysis. We keep implementation and result enhancement using proper sampling method as future work.

VI. Conclusion and Future Work

In this work we propose a variant of adversarial training for the application of puzzle solving, where one generative model is used to train another in a way similar to a human oracle providing labels at each time step to a recurrent neural netwrok. We obtain 99% accuracy on validation set which implies network has learned whatever structure has been given to it in training and validation set. Now if test data is similar to training and testing set, the network will be able to extract correct answer. Since this is not happening, it implies that sample size is small. Several reasons mainly related to the size of the possible state space of the puzzle problem suggest that the problem lies in small size of training set. In other words the sample size is not large or varied enough to train the network properly.

There are many more reasons for comparing our work with other sequence to sequence learning tasks. There are non-trivial differences between Question-Answering (QA) and Puzzle Solving. For example, QA can be based on comprehension where answer spans searching through multiple interrelated sentences in a monotonic order, while in puzzles a change of just 1 position can result in new state. Variable-length answers (which make the QA problem difficult) and a fixed-length answer is also another key difference between QA and puzzle solving task. Due to these differences we can expect that in modeling sequence to sequence learning, the easiest task can be modeling of state to state for puzzle solving (fixed length answers, not dependent on many positions/sentences, with a very small vocabulary size). This argument motivates us to expect long exact answers for puzzle solving using sequence to sequence modeling.

REFERENCES

[1] http://www.sourcecode.se/sokoban/levels/

[2] http://bach.istc.kobe-u.ac.jp/copris/puzzles/sokoban/

[3] M. Dorst, M. Gerontini et al, "Solving the Sokoban Problem", 2011

[4] Jean-Noel Demaret, Francois Van Lishout et al, "Hierarchical Planning and Learningfor Automatic Solving of Sokoban Problems"

[5] Andre Grahl Pereira et al, "Finding Optimal Solutions to Sokoban using Instance Dependent Pattern Databases", Institute of Informations Federal Univerity of Rio Grande do Sul, Brazil, Proceedings of the Sixth International Symposium on Combinatorial Search, 2013

[6] J. Taylor, I. Parberry, "Procedural Generation of Sokoban Levels", Dept. of Computer Science and Engineering, University of North Texas

[7] Hochreiter, Sepp and Schmidhuber, Jurgen, "Long Short-Term Memory", Neural Computation, 9(8):1735-1780,1997

Automatic Conditional Switching (ACS), an Incremental Enhancement to TCP-Reno/RTP to Improve the VoIPv6 Performance

Asaad Abdallah Yousif Malik Abusin
Faculty of Computing and Informatics,
Multimedia University,
Cyberjaya, Selangor, Malaysia

Junaidi Abdullah
Faculty of Computing and Informatics,
Multimedia University,
Cyberjaya, Selangor, Malaysia

Tan Saw Chin
Faculty of Computing and Informatics
Multimedia University
Cyberjaya, Selangor, Malaysia

***Abstract*—In this research work an Automatic Conditional Switching Protocol (ACSP) is proposed, which is a conditional switching method between Delay based TCP-Reno and RTP (Real-Time Transport Protocol). It is a delay constrained method based on the VoIPv6 delay limit stated by the Internet Engineering Task Forth (IETF), which is 150 milliseconds in accordance with the ITU G.114 standard recommendation (G.114 International Telecommunication Union, 1994) which recommended that the permissible amount of delay in VoIP to be less than or equal to 150 milliseconds in order to maintain the voice quality. This proposed Automatic Control Switching (ACS) should be implemented in the routers and switches in the protocol layer of the VoIPv6 network to improve VoIPv6 performance. The system can be defined simply as combination of TCP Reno/UDP (RTP) with Automatic Conditional Switching. This Automatic Conditional Switching is based on Delay based Congestion Avoidance (DCA).**

Keywords*—TCP Reno; VoIPv6 Performance; VoIPv6 improvements

I. Introduction

The Automatic Conditional Switching Protocol (ACSP) is a Delay Avoidance, Two Pass method for VoIPv6 which should be implemented in the routers and switches in the protocol layer of the VoIPv6 network in order to improve VoIPv6 performance utilizing DCA (Delay based Congestion Avoidance) as a driver. A report on the performance of the two protocols utilized in ACS is presented in [26], and a proposal to carry TCP traffic over UDP was proposed in [27], which was supported by review on the different TCP variants, and how they can support the voice traffic, these TCP variants were discussed in [29]. [30] compared between the different TCP flavors among which ACS utilized TCP Reno to work conditionally with RTP in order to produce the resultant ACS packets.

II. Research Methodology, Analysis, and Design

A. Key Concepts in Delay-based TCP as Part of The Switching Process in this Proposed ACS

The collection of the TCP packets congestion control techniques that are correlated with TCP/Reno are considered as a performance improvement measures which are mainly (slow-start and congestion avoidance) which can provide network saturation based on improved characteristics. The enhancement method to improve VoIPv6 performance ACS is proposing to use this delay based TCP in one hand and RTP (UDP) on the other hand and applies a conditional real time switching between both based on conditional switching criteria. In order to establish this, there is a need to apply congestion controls in VoIPv6 traffic, these congestion conditions according to [2] are:

- Delay based TCP
- Loss based TCP
- TCP congestion control
- Mixed loss/delay passed congestion control
- Explicit congestion notification
- Prediction of the lost VoIPv6 packets.

Each of the above six congestion control conditions is subjective to enhancement, and can be proposed as a research basement for further enhancements in VoIPv6 performance. This research work made enhancements on the proposal made by [2], and implemented new VoIPv6 improvement technique that is Delay based TCP/RTP (UDP) with conditional switching capability to be implemented as an ACS in the IPv6 network routers and switches. As a step to support this proposal, it is important also to study, test, and evaluate the overhead imposed by this improvement suggestion and then look for ways to reduce it. [28] discussed the challenges in network chatting and Sending Data including voice. A very important statement from [2] is that, in order to attain high level of deployment of new TCP enhancements such as this research work proposed ACS, there must exist sufficient encouraging economic factors in order to adopt such proposed measures, in some cases these encouraging economic factors to deploy new improvements are weak or may result in temporary sacrifice of the available resources for users prepay for a considerable length of time before its deployment become practically popular in the network. The VoIPv6 performance improvement proposal in this research work is of high economic value since the economic value of long distance

calls, business transactions, and electronic transactions in general, besides the real-time transactions and the social media is high. Both the research work carried out by [3], and the proposed ACS in this research work utilizes the benefit of Delay based Congestion Avoidance (DCA) added by studying the performance of Primal Dual Schemes for Congestion Controls in busy networks with dynamic stream such as VoIPv6 network.

In order to test ACS in this research work, a VoIPv6 Client/Server transactions were monitored and the inter-arrival time of the voice packets were captured using FreeBSD which is similar to [4] who studied Client/Server transactions for sensitive traffics such as VoIPv6, in which the scaling of virtual worlds in the age of cloud computing is complicated by the problem of efficiently directing client-server traffic in the service of dynamic compute resources in a different approach from this research work emphasizing on cloud computing proposing a model in which software defined networking and compact encoding of important data in packet headers were combined to make a fast, scalable, high capacity proxy server that can hide the server infrastructure while fitting well with the structure and design of modern network service providers.

A DCA (Delay based Congestion Avoidance) algorithm cannot reliably assess the congestion level at the router; this is due to the Short-term queue fluctuations. The measured data in [2] was collected in the year 1999 and the results were published in the year 2003 provided that the observed dynamics from their conducted tests were representing June 2003 public Internet quality and behavior. This also is pointing clearly to the fact that improvement in VoIPv6 performance is of cumulative nature, and this research work studied these accumulated factors such as VoIPv6 packet delay, and the performance of the network related functions, based on Client/Server tests conducted in the years 2013 and 2014. In order to implement ACS it is important to study the interrelation between packet delay and TCP Round Trip Time for the data packets, this is mainly because the TCP constrained RTT congestion verifier exactly track the level of packets congestion related to packet delay over busy network with high packets stream forwarding including VoIPv6 stream. A DCA (Delay based Congestion Avoidance) algorithm cannot efficiently assess the level of congestion at the router; this is due to the Short-term queue fluctuations, and the performance improvement in any of the network levels will contribute to the overall VoIPv6 network performance improvement; in this point this research work focus on packet delay and packet delay avoidance.

[5] coined the term Delay-based Congestion Avoidance (DCA), he admitted that his DCA algorithm is not sufficient for practical networks, but his work provided the foundation and framework for future research on DCA and interrelated mechanisms including ACS as an enhancement in the protocol level of the VoIPv6 network, introducing TCP/Vegas compared to this research work which utilized TCP/Reno as part of ACS. Their studies were based on Internet paths that existed in the early 1990s which generally involved at least one-To-one speed link, and allowing any given flow to consume a significant fraction of available bandwidth. The network performance improvement methodology proposed by [5] discussed congestion avoidance algorithms, this congestion avoidance is the scientific basement of Delay based TCP, and Delay based TCP is the basement for this research work main contribution (ACS Protocol), supporting the VoIP continuous evaluation, and improvement effort as in [22].

According to [2], a substitutional packets congestion control mechanism ought to be a TCP-compatible resulting in a similar performance to that achieved by a likewise network host across resembling network path to TCP characteristics. As this research work is an incremental enhancement to TCP, and accordingly any proposed incremental performance improvement in TCP that coincide with the following requirements will be competently deployable, and accordingly this research works proposed ACS deployment in the network can be evaluated and assessed according to the following criteria:

- The incremental enhancement ought to increase the number of transmitted TCP packets that adopts the performance enhancement.
- The incremental enhancement ought not to negatively affect the performance of other TCP packets streams competing in the utilization of the same network route through which the enhanced TCP flow is being transmitted, that is in order to attain better utilization of the network available bandwidth without negatively affecting other TCP flows.
- The incremental enhancement need to change the TCP client, and the above two properties must be attained without considering the number of improvements in the TCP connections through the same Client/Server path, consequently these properties must be maintained even if there exists just a single enhanced flow in the client/Server connection. In the wide spectrum of the global Internet for the purpose of supporting global implementation of TCP incremental enhancements, a substantial enough encouraging reasons must be present leading to adapt them, in the case when these encouraging factors are less significant, or causing short term cut of resources for users before its complete deployment, then the resultant final wide deployment of such protocols will be negatively influenced.

[2] described the main features of the Delay based Congestion Avoidance (DCA):

- The DCA strengthen the performance of the TCP/Reno protocol.
- DCA monitors, and trace the TCP packet Round Trip Times (RTTs) allowing the algorithm to reside holistically within the TCP server.
- The Round Tripe Time continuous changes are looked into as a consequence of changes in queuing delays occurred to the timed out packets.
- Based on Round Trip Time changes, the Delay based Congestion Avoidance (DCA) makes the necessary

congestion decisions that reduce the packets transmission rate by adjusting the TCP congestion window.

[5] invented the term delay-based congestion avoidance, he confirmed that his proposed DCA was not substantially applicable for a practical packetized network; but DCA provided the platform, and benchmark for the on-progress research work on Delay based Congestion Avoidance (DCA) towards making it practically applicable. This research work proposed ACS protocol as an improvement in Delay Tolerate Networks (DTN) provided that ACS is governed by the 150 milliseconds delay limit or hit value, as this research try to utilize (R. Jain, 1989) idea in order to produce the Conditional Switching Protocol (CSP).

[25] encouraged improvements during the World IPv6 Day by AMPST of Australia in the different level of the network to support IPv6 applications. Research work in [2] was utilized as part of this research work proposed ACS as it assessed the possibility in TCP controlled congestion sampling algorithm in RTT. Forwarded by utilized it in predicting the possibly coming packet loss occurrence which can be part of performance improvements in packet based real time networks, and for this reason, this research work used RTT-based congestion sampling algorithm as part of DCA in the proposed ACS. Source to destination packet delay and loss attitude over the Internet, as these factors are considered the main two considerations in VoIPv6 performance improvements. A finding result in [2] showed that the occurrence of packet loss was found to be interrelated over packets inter-arrival time values up to 200 milliseconds with loss correlation up to one second ignoring the interrelationship between the rise in RTT and packet loss ignoring the existing interrelationship involving loss conditioned delay that occurs immediately after a packet loss. The motivations of [2] were similar to this research work in the sense that they studied the protocol level locking for possible improvements. When examining the occurrence of packet loss, the result will be that the RTT that tend to increase the surrounding packet loss will live for short time. Loss processes are optimistically interpreted as chunks of data during which there exists a very short-term stops, instead of short chunks of data over which congested router's buffer is congested with packets that could be discarded.

The research outcome from [2] is that DCA working alone is not substantially enough to improve TCP performance and is accordingly conditionally not applicable to be deployed as an increment to TCP unless DCA is supported by other techniques such as this research work proposed ACS since this research work tries to improve DCA by adding it accumulatively in the switching between TCP Reno and RTP, to work together in a joint performance improvement effort. Besides the TCP congestion control aspects, the research led by the internet research bodies of concern including the IETF, and the IF has also been extensively verifying, examining, and applying TCP algorithms for real-time packet traffic applications including voice, this is exactly what this research work try to achieve by proposing ACSP which can be considered a TCP/RTP-Friendly algorithm to improve VoIPv6 performance. The performance of the deployable TCP-friendly algorithms can be similar to TCP when they utilize Additive Increase Multiplicative Decrease (AIMD) behavior with the same features as TCP. The TCP potentially deployable methods, such as this research work proposed ACSP, are subjective to further assessment referring to steady-state or dynamic behavior, an illustrative example here is the TCP-friendly rate control (TFRC) technique which responds to a single packet loss by reducing the packets transmission rate to values smaller than TCP packets transmission rate. The TCP deployable incremental algorithms are targeting to be TCP compatible by emulating the TCP behavior, with the TCP main features such as slow start, congestion avoidance, and exponential timeout back off in addition to self-clocking. All rate-based protocols that fix their rates based on a computation, including this research work proposed ACS incremental enhancement protocols, ought to include a measured RTT in order mainly to add an element of delay-based congestion avoidance (DCA) to their algorithms.

[2] proposed techniques based on hypotheses on implicit packets congestion feedback based on delay instead of packet loss or additional to packet loss, and this can positively improve the performance of the global IP network in case high number of IP traffic adopted the same algorithm emphasizing mainly on packet delay rather than packet loss. The expected outcome would be the same as what inspired the alternative best effort (ABE) proposal, which is a network that provides packets forwarding service that dedicated to attain less amount of delay and less throughput; such service expected to be beneficial for multimedia applications that are very sensitive to delay such as VoIPv6. This encourages studying shortest bath selection for VoIP packets and the hop beer hop behavior in DCA as this DCA is the basement for this research work proposed Automatic Conditional Switching (ACS).

In their experiments [2] used TCP-Dump to trace the TCP connection in a similar manner as in this research work experiments in which Client/Server connections were established, traced, and monitored from the time the connection was established between the VoIPv6 client and the VoIPv6 server to sending VoIPv6 packets stream, the TCP sender experiments as well as the trace point was a 333-MHz PC running FreeBSD. FreeBSD was the same performance testing medium used in this research work. The machine used in [2] consists of 3COM PCI Ethernet adapter and was attached to North Carolina state university (USA) campus network through a 10-Mb/s Ethernet connection. These experiments were similar to this research work experiments conducted at Multimedia University (MMU Malaysia) intranet in the year 2013 in which a developed Client/Server codes were used to capture the arrival and the inter-arrival time of a real time VoIPv6 packet samples, in addition to that the experiments carried out in this research work similarly used FreeBSD as a testing tool. The packets queue management techniques that can be injected in the routers can contribute in rendering enhanced performance, and this research work main contribution ACS is proposed to be applied in the routers and switches of the VoIPv6 network.

[6] presented a simulation scenario describing the performance assessment of the two main transport layer protocols TCP and UDP using two queue management methods; Random Early Detection (RED) and Drop Tail,

mainly to trace the network throughput, queuing delay, packet drop rate and bandwidth utilization studying the effect of packets buffering on each performance measurement by simulating TCP, UDP and shared topology network scenarios by considering different number of client topologies and covering also the effect of performance in TCP and UDP with increasing the number of clients, such simulation work can be beneficial for implementing the packets queue control mechanisms in the router based on the traffic type and the available network bandwidth, an interrelated terms to packets management in the network are TCP, UDP, drop tail, RED queue, bandwidth delay product, throughput, and end-to-end delay, such effects can be considered as the major factors affecting VoIPv6 performance.

B. Existing Measures Contributing to VoIPv6 Performance Improvement

The way is open for new improvements in IPv6 and the accompanying applications, this is according to [24], ACS utilized measures contributing directly to the VoIPv6 performance improvements, which are mainly measured to remedy the collegian, contention, congestion, and queuing of voice packets. Performance measurements are needed to measure the collegian, contention, congestion, and queuing of the voice packets in terms of percentage of the network loss and collegian rate. One technique is needed to reduce or avoid collegian, and in case the technique is good for normal traffic then it can be tested in highly congested traffic in order to develop measures to reduce collegian, and avoid delay. [23] supported research, and discoveries for VoIP supported by ACS. The proposed ACS utilizes User Datagram Protocol (UDP) in its enhanced RTP form, and Delay based TCP which is using DCA jointly as an enhancement for the two commonly used protocols on the Internet (TCP and UDP) for them to work concurrently and conditionally. This is an improvement contribution in this research work because UDP, and its refined form (RTP) used not to send important data such as web pages, and database information, and it is commonly used for packetized voice and video Streaming, such as Windows Media audio files (WMA), Real Player (RP). The constrained packet transmission was used in this research work proposed ACS conditionally, and when the condition satisfied the switching to TCP-Reno will take place, and with the continuous real-time monitoring of the voice speed (less than or equal to 150 Milliseconds) switching takes place back and forth conditionally.

Many VoIPv6 current applications uses UDP because it offers speed, and faster processing for the real time traffic. The reason UDP is faster than TCP is because there is no any form of flow control or error correction. The data sent over the Internet is affected by collisions according to witch errors in packets forwarding will be present. UDP is only concerned with packets speed, sacrificing the quality, and this is the main reason behind the lake of quality in streaming media, which is a hot research aria as this research work concern. In this research work proposed ACS, in the switching process, a consideration for future research work can be to figure and correlate by obtaining a correlation function, between the overall total delay and packet delay, and loss relationship with the routers buffer size and the per hop behavior in VoIPv6 traffic, keeping in mind that ACS avoids packet loss focusing on improving the VoIPv6 performance utilizing packet delay in the form of ACS as this research work main finding.

The Real Time Transport Protocol (RTP, RTCP) is utilized as part of the switching in this research work proposed ACS as illustrated in Fig. 1. In [7], they specified set of extensions to the base Real-Time Transport protocol which was specified in RFC 3550, according to which RTP is a collection of network transport functions that are applicable for applications transmitting real-time data such as VoIPv6 stream transmitted between network client and a server. When comparing this research work proposed ACS with the proposed TCP/UDP router in [8], it is noted that their performance improvement suggestion is based on developing a linear time-delay system model which can be derived by obtaining packets flow stability conditions with elementary router parameters using the 2-D Laplace-Z transform technique leading to deriving parameter setting for TCP/UDP routers in order to improve the network congestion control based on 2-D local stability conditions, and this can be achieved through applying serial of simulations which verified that their proposed router can attain the desirable packets congestion control by utilizing the stability conditions, proving that performance comparison for different TCP/UDP network scenarios with the traffic stability conditions can contribute to performance improvements as in this research work the reference stability condition is the 150 milliseconds overall cumulative delay limit hit value. Up to the year 2014, according to [9], most multimedia applications, including VoIPv6 services, and applications utilizes a combination of Real-Time Transport Protocol (RTP) and User Datagram Protocol (UDP). ACS in this research work is a switching mechanism applied in a combination of TCP Reno and RTP. The developed application programs at the source end format payload data into RTP packets using RTP specifications and dispatch them using unreliable UDP along a single path, multi-path transport, this means that up to 2014 there is no suggestion to switch between TCP and UDP which is considered an important on the efficiency of voice data delivery.

C. The Proposed ACS Compared to New Trends in VoIPv6 Performance Improvement Methods

A step towards VoIPv6 performance improvement is through continuous performance testing trails, and for the improvement in the protocol layer of the VoIPv6 network, one of the improvement methods (as they are many ways in the three VoIPv6 network layers), is the removal of the checksum operation at the network layer which is not enough to recover the long address processing time. The IPv6 header is considered simpler when compared with IPv4, and the simpler IPv6 header does not help in this case because experiments in [10] were conducted within MMU (Multimedia University of Malaysia) intranet in local loop intranet environment. From the theoretical point of view, it will enhance the end-to-end performance if there are many intermediate hosts since there will be less processing in the intermediate hosts and nods due to the distribution of tasks. IPv6 uses smaller send and receive buffer compared to IPv4, this affect the conducted performance tests significantly since small buffer become full faster. IPv6 is still considered improving and in its evolving stage, this is

Identify applicable sponsor/s here. If no sponsors, delete this text box (*sponsors*).

according to the IETF. And ITU G.711, and G.744 of the ITU which expect IPv6 to reach its maturity by the year 2050 welcoming continuous requests for comments to be published as RFCs. The IPv6 KAME stack is under continuous development and frequently being changed and updated confirming the fact that the current release of IPv6 is not the best all the time, and a research work can be carried out in investigating the performance of the implementation of the IPv6 stack. KAME IPv6 stack implementation needed continually to be improved to the most optimal level to give better performance in the associated applications. These tests can be repeated in different network conditions and highlight the differences and similarities. The IPv6 features supporting real-time traffic including VoIPv6 stream can be clearly noticed in the current internet, which are needed due to the evolving enhancements in IPv6 network with the introduction of Broadband internet service which can contribute to performance improvements in the real-time applications under IPv6.

III. Description of the Proposed ACS Model

This research work proposed ACS is to be implemented in the routers and switches of the VoIPv6 network. An illustrative diagram for ACS is shown in Fig. 1. ACS treats VoIPv6 packets under TCP Reno then apply Congestion/Delay based TCP as in [11] from which the resultant packets to be injected to TCP-Dump to trace the TCP connection and then use RTT-based congestion sampling as in [2]. Then use Router scheme with classifier and access control to separate UDP from TCP traffic using linear time delay system model based on stability condition which is the VoIPv6 maximum delay limit which is 150 milliseconds. Then use Delay based (for voice packets)/Loss based TCP congestion Avoidance (DCA) and loss based TCP/UDP router (for the associated data packets). The last step is the switching step, which is to use Linear Time delay system to continuously compare the total delay with 150 Milliseconds, this is the moment of the conditional switching when the total delay is less than or equal 150 milliseconds then to use TCP Reno, else if the total delay is greater than 150 milliseconds then to use the Real-time Transport Protocol (RTP) in order to produce the resultant processed ACS packet.

There are two options to conduct ACS tests in real time:

1) Software Virtual Machine in ready available commercial software

2) Hardware development, which can be done by following the steps in the ACS descriptive diagram as in Fig. 1

In this research work, the ACS switching was tested based on FreeBSD captured VoIPv6 data (Packets Inter-arrivals) as shown in Fig. 2. ACS switching was tested using Mat Lab. In order to accomplish this goal this research followed the following steps while using Mat Lab:

1) Read data which was put in Excel sheets.

2) Count the number of rows and columns

3) Then if conditional statement, in case the incoming tested value overall delay is less than or equal to 150 milliseconds then use TCP Reno, else if the overall delay is greater than 150 milliseconds then switch to RTP and be in continuous monitoring alert mode to take the ACS decision. In case the intention is to make the tests in real time then the steps as in the diagram in Fig. 1 should be followed. For performing the ACS testing in real time then the following equipment's are needed:

- Analog to digital converter to produce the digitized voice in real time.
- Traffic shaper to shape the packets in IPv6
- Perform ACS.

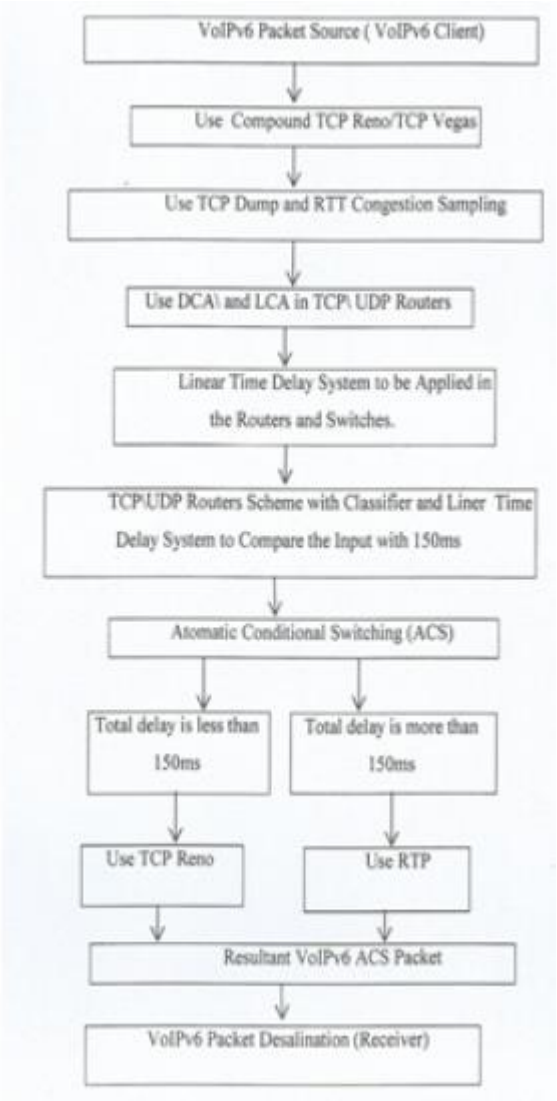


Fig. 1. ACS Descriptive Model Diagram, Mathematical Model Using Laplace Z Transform is used to Represent the Switching in ACS

A. ACS Validation

ACS validation and testing is of concern here as it can be a platform for further improvements towards ACS practical implementation, in this context a future work can be the usage of Laplace Transform to represent the switching mathematically.

B. Mat Lab Code Describing ACS on a Captured FreeBSD VoIPv6 Data

The following Mat Lab code was developed, and used to perform the switching in ACS on captured VoIPv6 inter-arrival time data values captured using FreeBSD. The following Mat Lab code reads the VoIPv6 inter-arrivals data values in milliseconds stored in Excel files:

```
clear all; clc
% Function Name: Asaad Calculator ACS%
% Read the file
Data = xlsread('Asaad_Data_02.xlsx');
[Nr,Nc] = size(Data); % size of the data
count_yes = 0;
count_no = 0;
for i=1:Nr
    if Data(i,2)<= 150;
Ans(i)=1;
count_yes = count_yes +1;
    else
Ans(i)=0;
count_no=count_no + 1;
    end
end

%% Final Answer
Ans=Ans';
Percent_discarded = (count_no/Nr)*100;
Percent_Accepted = (count_yes/Nr)*100;
```

C. Test Results for ACS Validation Using Mat Lab Applied to the VoIPv6 Packets Inter-arrival Data

In Table 1 the number of yes indicates that the Inter-Arrival time for the voice packets is less than or equal to 150 milliseconds. The number of No indicates that the Inter-Arrival time for the VoIPv6 packets is greater than the 150 milliseconds hit value.

TABLE. I. RESULT ANALYSIS USING MAT LAB FOR CAPTURED VOIPV6 PACKETS DATA

Data Name	Table Column Head			
	Number of Yes	*Number of No*	*Percentage Accepted*	*Percentage Discarded*
Test 1	677	323	67.7%	32.3%
Test 2	846	154	84.6%	15.4%
Test 3	996	4	99.6%	0.4%

D. Testing the Switching Possibility in ACS Based on DNS

The performance of the IPv6 addressing system to surf the internet is based on the communication between source and destination IP addresses which is based on spotting the IPv6 address and the port number, besides the Domain Name System (DNS) that is involved in the background applications accompanying the VoIPv6 stream of packets in which the client is continuously monitoring and recording the arrival and inter-Arrival time of the IPv6 packets, and ACS is based on continuous monitoring for the 150 milliseconds voice packets hit value in order to decide performing the switching back and forth conditionally between TCP Reno and RTP. The recorded background information that is required as condition to perform the switching, is lengthy for UDP in order to take the decision to perform ACS, and the DNS steps can be:

1) The server responds with specifying opcode to have client switch to TCP Reno.

2) When the server is not responding, the client re-transmit over TCP.

3) The server respond by opening TCP connection to client Network Address Translation (NAT).

4) The client now that the given query should be run prior over TCP, for this reason RTP cannot be used in the first place.

5) DNS tiny chunks turn UDP into TCP conditionally utilizing DNS and Name server searching trails.

The conditional switching based on DNS is known in the internet used to block the surfing of undesirable web sites, but the automatic conditional switching between protocols is not practically known yet which is proposed as ACS in this research work which needs more verification before its practical deployment. ACS also is not deployed in the current versions of Opnet (current version 17.5) and Netsim (version 10). These WAN and network simulators are mainly dedicated to perform the simulation only for delay, packet loss, and jitter targeting mainly these network impairments, and needed to be upgraded to newer versions that can allow for Automatic Conditional Switching between TCP Reno and RTP.

E. How to do the switching Between TCP Reno and RTP?

According to this research work it can be achieved by encapsulating TCP Reno on RTP, and the encapsulation, can be achieved using Network Address Translation NAT defined in RFC-1631. Originally NAT was designed not to permit the depletion of IPv4 address, it is proposed here to be used in encapsulating TCP Reno into RTP in order to facilitate the switching between them. NAT is a method to remap one IP address space into another by modifying network address information in the Internet Protocol (IP) datagram packet headers during their transit across a network traffic router. The port forwarding is needed in NAT, in addition to the category network address including the Address Space. Some researchers criticize NAT, and think to avoid it in the IP network design and implementation because it breaks the connectivity between hosts especially in packetized VoIPv6 networks. [13] defined NAT when a packet is received from an internal host, the NAT process searches for the matching source address-source port combination in the port-mapping table which contains a list of inner local port numbers that get translated into particular mapped ports, and in case the entry is not found for the received source port, a new port translation entry is created and the new port is allocated for the received source port. The source IP address is replaced with a translated global address from the translation table or NAT address pool and the source port is replaced with port number from the port-mapping table. At the end of the NAT process, the voice packet is sent out with the translated source IP address and source port number, and the time needed for this process is named as practical delay in NAT Router.

F. Practical Delay in NAT Router and its impact on ACS

The delay in NAT router needed to be reduced to the minimum value with a minimum possible contribution to the 150 milliseconds less overall delay in VoIPv6 as recommended by the G.711 of the ITU. The practical network delay with NAT network according to [14], when ping a server which

usually requires 20 to 30 milliseconds for NAT, the Virtual Box will get 20 to 30 milliseconds for the first three pings (acceptable for voice), then 500 milliseconds (unacceptable for voice) for the next 3 pings which is overcome by ACS, then 500 milliseconds for the next 3 pins (unacceptable for voice), and then back to 20-30 milliseconds (acceptable for voice). Because of NAT nature the first switching by ACS after receiving three consecutive NAT packets should be freezed for a predetermined length of time before the next switching. RTP as part of ACS runs over UDP, for this reason a VoIPv6 packet has a 20 bytes IP address distributed as 8 bytes UDP packets, and RTP 12 bytes packets headers added to the encapsulated voice payload. The Digital Signal Processors (DSP) creates a package from each10 milliseconds of Analog voice, and two of those packages are transported within one IP packet. A 20 milliseconds resultant packets of voice traffic is common in one IP packet. The number of bytes resulting from 20 milliseconds from Analog voice depend on the codec slandered used. The G.711 which produces 64 Kbps, produces 160 bytes from 20 milliseconds of Analog voice, and G.729, which generates 8 Kbps, out of which is 20 bytes for 20 milliseconds of Analog voice signal. The RTP, UDP, and IP headers, with total of 40 bytes, are added to the voice bytes (160 bytes for G.711 and 20 bytes for G.729) before the whole group is encapsulated in the Layer 2 frame and transmitted.

In case the Network Layer 2 overhead is not considered, considering only the overhead resulted by RTP, UDP, and IP, it can be notable that the needed bandwidth is more than the bandwidth that is needed for the voice payload, as when using the G.711 codec, the required bandwidth for voice is 64 Kilobytes per Seconds, but with 25 percent added overhead of IP, UDP, and RTP, the required bandwidth increases to 80 Kbps. In case the G.729 is used, the bandwidth required for voice packets is 8 Kbps, but with the added 200 percent overhead imposed by IP, UDP, and RTP, the required bandwidth increases to 24 Kbps. The overhead imposed by the Layer 2 protocol provided that any other techniques such as tunnelling or adding security mechanisms is not included.

In order to reduce the header overhead imposed by IP, UDP, and RTP the Compressed RTP (cRTP) can be used, which is also called RTP header compression, as cRTP reduces the overhead imposed by all IP, UDP, and RTP protocol headers. cRTP should be applied on the sending and receiving ends of the voice packets, and both agree to a hash number that is associated with the 40 bytes of IP, UDP, and TCP headers. In cRTP most of the fields in the IP, UDP, and RTP headers do not change among the packets of a common flow. After the initial NAT packet with all the headers is submitted, some packets that do not carry the 40 bytes of headers, and instead, the packets carry the hash number that is associated with those 40 bytes with their sequence number built in the hash. The main difference among the headers of a packet flow is the header checksum which is UDP checksum, and in case cRTP does not use this checksum, the size of the overhead is reduced from 40 bytes to 2 bytes. In case the checksum is used, the 40 bytes overhead is reduced to 4 bytes, and in case during transmission of packets, a cRTP sender realized that a packet header has changed from the normal form, then all the header instead of the hash is submitted. The benefit of using cRTP with smaller payloads such as digitized voice is more recognized than other large payloads.

Fig. 2 describes the steps in the switching by encapsulating TCP Reno over RTP in order to improve VoIPv6 performance. The starting of the process is the voice packets source exchanging voice packets arrival and inter arrival information within the network mainly with the server. An application program is needed to shape the voice packets in IPv6 followed by using TCP Reno as in RFC 793. A real time bottleneck is applied in order to monitor the arrival and inter arrival time of the voice packets in order to perform the switching conditionally.

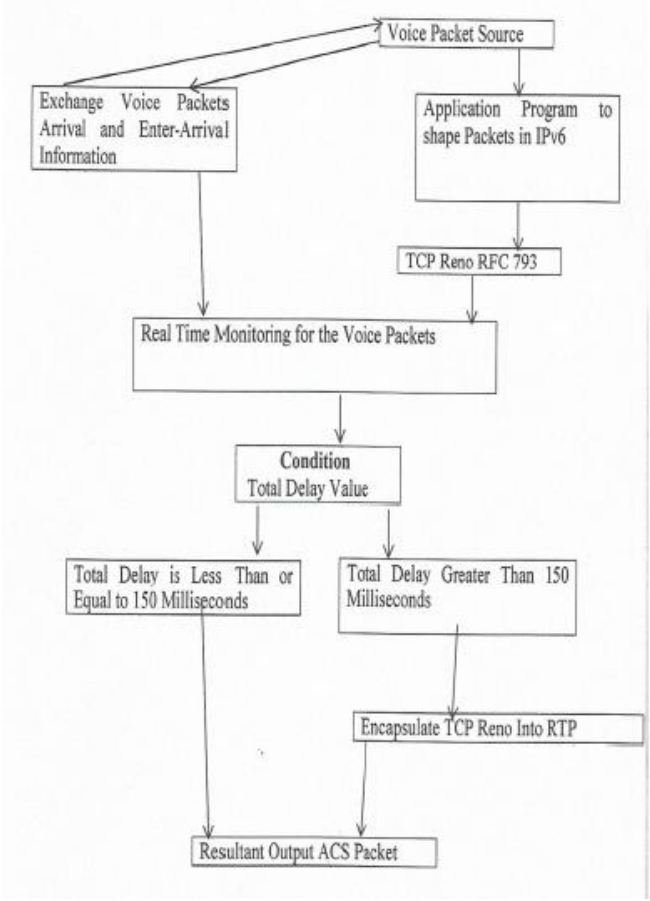


Fig. 2. The steps in the switching by encapsulating TCP Reno over RTP in order to improve VoIPv6 performance

IV. This Research Work Main VoIPv6 Improvement Method (ACS)

This research work's main VoIPv6 improvement idea is a smart conditional switching capability for the VoIPv6 traffic as an incremental enhancement to a combination of TCP Reno and RTP, which can be described precisely as, Delay-based TCP-UDP Selective Dual Stack (TCP Reno/UDP RTP) routers scheme with classifying, selecting and switching capability. This ACS model idea resemble the automatic conditional switching in the automatic gearing system in the automatic car, whereby the continuous monitoring in the automatic gear is for the car speed and in this research work proposed ACS it is for the voice speed in milliseconds. Likewise the existence of dual

stack IPv4/IPv6 idea, this research work proposes the dual existence of Delay-based TCP Reno-RTP Selective Conditional Switching Dual Stack routers to improve VoIPv6 performance.

According to [9], in the year 2014, most multimedia applications utilize a combination of real-time transport protocol (RTP) and user datagram protocol (UDP), this statement support the uniqueness and the genuineness of this research work proposed ACS as a smart back and force conditional switching between RTP and TCP Reno. When comparing this research work proposed ACS with the proposed TCP/UDP router in [8], it can be noted that there proposal is a packets congestion avoidance scheme that can be expressed by a linear time-delay system model after deriving some stability conditions with simple router parameters by utilizing the 2-D Laplace-Z transform technique.

This research work proposed ACS is based on the following:

- A criteria in the VoIPv6 network that contributes to good or bad voice quality, and accordingly finding new switching criteria based on the 150 milliseconds less possible voice delay limit.
- The correct switching decision in the VoIPv6 network with reference to ACS can be able to improve the performance in VoIPv6 utilizing Automatic Conditional Switching Protocol or ACS which is this research work main contribution.
- Formulate a model for the switching process, which is a design to switch back and forth automatically between RTP and TCP Reno.

ACS can be considered as a new contribution when compared with the statement from [9] which stated that, in the year 2014, most multimedia applications utilize a combination of real-time transport protocol (RTP) and User Datagram Protocol (UDP).

A. The Scientific Basement for ACS

Based on [8] proposal, this research work propose ACS, their proposed TCP/UDP router can be described as linear time-delay system model which augmenting stability conditions in the network with elementary router parameters by utilizing the 2-D Laplace-Z transform mechanism, by proposing parameter setting for TCP/UDP routers in order to improve the network congestion control based on 2-D local stability conditions. A serial of simulation results obtained by [8] verified that their proposed router can attain good level of network congestion control concerning the stability conditions considering performance comparison for different TCP/UDP network scenarios with the associated stability conditions. In [2], they were looking for incremental improvements to TCP to support real-time applications, such incremental improvements as TCP Reno and TCP/Vegas. TCP Vegas attempt to utilize the congestion information contained in packet round-trip time (RTT) samples. TCP Reno instead of TCP Vegas was used as part of this research work proposed ACS due to the sensitivity of VoIPv6 stream to packet loss.

The Real Time transport Protocol (RTP) specified in (RFC3550) was used as part of this research work proposed ACS in the switching process. This RTP was specified by (Microsoft Extensions, Thursday, May 15, 2014) which described the Real-Time Transport Protocol (RTP/RTCP) according to which RTPME is a set of extensions, as similar as this research work proposed ACS which is also considered as an extension or increment to TCP Reno/UDP in addition to the base Real-Time Transport Protocol (RTP) specified by the IETF in (RFC3550). RTP is a set of network transport functions suitable for applications transmitting real-time data, such as audio and video, across multimedia endpoints including VoIPv6.

[2] focused on delay-based congestion avoidance algorithms (DCA), like TCP/Vegas, which attempt to utilize the congestion information contained in packet round-trip time (RTT) samples in an attempt to look for performance improvement possibilities through measurement and simulation showing evidence suggesting that a single deployment of DCA, which is a TCP connection enhanced with a DCA algorithm, is not a viable enhancement to TCP over High-speed paths, this research work tried to overcome this by utilizing combination of Refined TCP, and RTP, and applying ACS between them.

[11] highlighted that some efforts combined the features of loss-based and delay-based algorithms, in this context this research work proposed (ACS) can be considered utilizing DCA to achieve fair bandwidth allocation and fairness among flows, and consequently improve VoIPv6 performance. A comparative analysis between different flavors of TCP congestion control namely Standard TCP congestion control (TCP Reno) which is the TCP flavor used in this research work proposed ACS, loss-based TCP congestion control (High-speed TCP, Scalable TCP, CUBIC TCP), delay-based TCP congestion control (TCP Vegas) and mixed loss-delay based TCP congestion control (Compound TCP). These TCP flavors were presented In terms of congestion window versus elapsed time after the connection is established. Work in [11] is the foundation for this research work proposed improvement method to enhance VoIPv6 performance (ACS), which can be considered as Delay Based TCP/UDP Congestion Control Algorithm with Selective Switching features.

[5] coined the term delay-based congestion avoidance which is used as part of ACS, while admitting that his proposed algorithm was not sufficient for a practical network, his research work provided the foundation for future research on DCA including ACS. Accordingly the work described in [15] and [16] which showed that TCP/Vegas can improve TCP throughput over the Internet by avoiding packet loss, comparatively these studies were based on Internet paths, or network infrastructure that existed in the early 1990s, which generally involved at least one-T-one speed link and consequently allows any given flow to consume a significant fraction of available bandwidth. These studies were on packet loss events and also did not isolate the impact of applying the congestion avoidance algorithm.

As part of VoIPv6 packets flow performance improvement and according to [20] most VoIP equipment's up to the mid of 2002 has dynamic jitter buffers. These jitter buffers compare

the receive time with the sent time stamp in the RTP header to calculate the jitter and automatically adjust the jitter buffer size. The installer has little control of the jitter buffer size and can only send the minimum and maximum size of data provided that the maximum jitter buffer size should be at least two times the voice sample collection time, butting in mind that the Jitter buffer size directly affects the latency in V0IPv6 experienced by the user. A jitter buffer cannot solve the problem of excessive packet jitter, and the Solutions for excessive packet jitter must be made in the network to reduce congestion and queuing time for voice packets, in this context the scope of this research work is limited to enhancement in the protocol structure purposely to improve VoIPv6 as a real-time application.

B. TCP-over-UDP and TCP Switching (between Packet and Circuit Switching) Compared to This Research Work Proposed ACS

[17] presented an interesting idea for VoIPv6 voice streaming improvement, which was TCP-over-UDP (ToU), a conditional instance of TCP on top of UDP, this idea is based on bringing TCP traffic to flow over UDP conditionally, instead of which this research work found it more practical to propose the switching idea (ACSP) between TCP and RTP conditionally. According to [18], there has been much discussion about the best way to combine the benefits of new optical circuit switching technology with the established packet switched Internet in order to explore how electronic and/or optical circuit switching might be introduced in an evolutionary manner.

[8] from the Institute of Information Science Beijing Jiaotong University of China, in their research work with the title (Congestion Control Algorithms for a new TCP/UDP router based on 2-D stability conditions), developed a congestion control algorithm for new TCP/UDP router based on 2-D stability conditions, they utilize the Dual Router idea to use both TCP and UDP, but not to switch between them conditionally as in this research work proposed ACS, and to use this to develop congestion control algorithms for the real time traffic based on switching focusing mainly on congestion avoidance as a step towards performance improvement, as in this research work an Automatic Conditional Switching (ACS) was developed to improve VoIPv6 performance. Research work by [8] can be viewed as an attempt to improve TCP performance which was used as part of ACS utilizing both RTP and Delay based TCP in order to provide smart switching capability to the VoIPv6 traffic.

According to [19], there was research effort concern in the year 2006 which had been raised about whether the increase of UDP traffic due to Skype type of applications will throttle down regular TCP users. This concern came from the fact that UDP connections are unresponsive to congestion, and accordingly UDP senders do not reduce their sending rates, which considered an incentive to investigate the fairness issue between VoIP flows and TCP flows under different network environment, considering both transport layer congestion control mechanism built-in TCP protocol stack and the self-adaptive-ness of VoIP users.

V. Conclusion and Future Works

ACS is a conditional switching between TCP Reno, and RTP based on the 150 milliseconds delay hit-limit standardized by the International Telecommunication Union (ITU) as in the G.114 standard recommendation (G.114 International Telecommunication Union, 1994). Without mentioning TCP due its associated delay problems, [9] stated that, currently (in the year 2014), most multimedia applications utilize a combination of real-time transport protocol (RTP) and User Datagram Protocol (UDP). In contrary this research work's novelty is a proposal to use flavored TCP Reno in one hand and Progressive RTP in the other hand, and applying Automatic Conditional Switching (ACS) between them, this ACS was tested, verified, and validated.

Developing mathematical equation and represent it with mathematical model using Laplace transform to describe the conditional switching in ACS is of concern as future work as it can be a platform for further improvements towards ACS practical implementation, and in this context a suggestion can be made to use Laplace transform to represent the switching mathematically.

Research work in [21] can be considered as base idea for future work to improve ACS, which is mainly to implement ACS in the real time, this is for a reason that they provided mathematical parallels between packet switching and information transmission. In order to conduct the tests in real time the steps presented in Fig. 1 must be followed carefully, and the following are needed:

- Analog to digital converter to produce the digitized voice in real time.
- Traffic shaper to shape the VoIPv6 traffic in IPv6.
- Perform the switching in real-time using ACS.

References

[1] G.114, One Way Transmission, International Telecommunication Union (ITU), 1994.

[2] J. Martin, A. Nilsson, and I. Rhee, "Delay Based Congestion Avoidance for TCP", ACM Tansactions on Networking, vol. 11(3), June 2003.

[3] Keixn Ma, Ravi R, Mazumdar, and Jun Luo, "On the performance of Primal Dual Schemes for Congestion Controls in Networks with Dynamic Flow", Proceedings of 27th IEEE International Conference on Computer Communications, pp. 326-330, April, 2008.

[4] C. Hansen, "Network Performance Measurement Framework for Real-time Multiplayer Mobile Games", IEEE Network and System Support for Games, December, 2013.

[5] R. Jain, "A Delay-based Approach for Congestion Avoidance in Interconnected Heterogeneous Computer Networks", Computer. Communication Revue, vol.19(5), pp. 56–71, October, 1989.

[6] ManasPratimSarma, "Performance Measurement of TCP and UDP Using different Queuing Algorithm in High Speed Local Area Networks", International Journal of Future Computer and Communication, vol. 2(6), December, 2013.

[7] Microsoft Extensions, Real-time Transport Protocol (RTP, RTCP), Technical Documentation. Microsoft published Open Specifications documentation, May 15, 2014.

[8] Yang Xiao, Lei Wang, Jun Niu, Seok Woo, and Kiseon Kim, "Congestion Control Algorithms for a New TCP/UDP Router Based on 2-D Stability Conditions", IEEE Press Piscataway, Proceedings of the 5th International Conference on Wireless Communications, Networking and Mobile Computing, pp. 4056-4060, 2009.

[9] W. Lei, W. Zhang, "Multipath Real-Time Transport Protocol Based on Application-Level Relay (MPRTP-AR)", Network Working Group of the Internet Engineering Task Forth (IETF), July 25, 2014.

[10] Asaad Abdallah Abusin, M D Jahangir Alam and Junaidi Abdullah. "Testing and analysis of VoIPv6, (Voice over Internet Protocol V6) performance using FreeBSD", International Journal of Computing, Networking, and System Sciences, May, 2012.

[11] Habibullah Jamal, and Kiran Sultan, "Performance Analysis of TCP Congestion Control Algorithms", International Journal of Computers and Communications, vol. 2(1). (2008),

[12] K. Egevang, RFC 1631, International Telecommunication Union (ITU). May, 1994.

[13] T. Hain, Microsoft, RFC 2993, Telecommunication Union (ITU). November, 2000.

[14] Visual Box www.virtualbox.org, https://www.virtualbox.org/ticket/5918, reported by (Clockwork). 6/11/2010.

[15] L. S. Brakmo, S. W. O'Malley, and Peterson, "TCP Vegas: New Techniques for Congestion Detection and Avoidance", Proceedings of ACM SIGCOMM, pp. 24–35, August, 1994.

[16] J. Ahn, P. Danzig, Z. Liu, and L. Yan, "Evaluation of TCP Vegas: Emulation and Experiment", ACM SIGCOMM, pp. 185–195, 1995.

[17] S. Baset, and H. Schulzrinne, "TCP-over-UDP", Transport Area Working Group of the IETF, June, 2009.

[18] Pablo Molinero Fernández, and Nick McKeown, "TCP Switching: Exposing Circuits to IP", Stanford University (PowerPoint Presentation). 2002.

[19] Tian Bua, Yong Liub and Don Towsleyc, "On the TCP-Friendliness of VoIP Traffic", IEEE INFOCOM, 2006.

[20] Gartner Consulting, "Convergence Challenges for Enterprise Networks", Gartner Consulting White Papers, May, 2002.

[21] Tony Lee T, "The Mathematical Parallels between Packet Switching and Information Transmission", IEEE Transactions on Information Technology, October, 2006.

[22] Mark Allman, "Internet Measurement", Internet Measurement Research Group IMRG, 6 August, 2012.

[23] Luo Junhai, Chengdu Fan Mingyu, and Ye Danksia, "Research on Topology Discovery for IPv6 Networks", International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel Distributed Computing. vol. 3, page 5, 1 August, 2007.

[24] International IPv6 Conference NAV6 Malaysia, Proceedings of NAV6, Jointly organized by the University Since Malaysia USM in collaboration with university of Malaya (UM) and the International University Malaysia-Wales (IUMW), 2013.

[25] Geoff Huston, "World IPv6 Day: A year in the Life", AMPST of Australia, 10 Jun 2013.

[26] http://www.technowledge.org, "A short detailed report that compares the real world performance of the most popular VOIP mobile phone", May 17, 2013.

[27] Internet drafts, TCP over UDP, IETF, June 2013.

[28] Muhand A. Mutaab, Nahla Abdul Jalil, and Jameelah H.Suad, "A Network Chatting and Sending Data", International Journal of Scientific & Engineering Research, December, 2014.

[29] Mandakini Yayade and Sanjeev Sharma, "Review of different TCP Variants in Ad-hoc Networks", International Journal of Engineering Science and Technology (IJEST), vol. 3(3), March 2011.

[30] B. Arunakumari and P. Chennareddy, "TCP Reno, Sack and Vegas Performance Analysis", International Journal on Cybernetics & Informatics vol. 4(2), April 2015.

Design of 1-bit Comparator using 2 Dot 1 Electron Quantum-Dot Cellular Automata

Angona Sarker
Dept. of Information and Communication Technology
Mawlana Bhashani Science and Technology University
Tangail, Bangladesh

Md. Badrul Alam Miah
Dept. of Information and Communication Technology
Mawlana Bhashani Science and Technology University
Tangail, Bangladesh

Ali Newaz Bahar
Dept. of Information and Communication Technology
Mawlana Bhashani Science and Technology University
Tangail, Bangladesh

***Abstract*—In nanotechnologies, quantum-dot cellular automata (QCA) offer promising and attractive features for nano-scale computing. QCA effectively overcomes the scaling shortfalls of CMOS technology. One of the variants of QCA is 4 Dot 2 Electron QCA which is well explored and researched. The main concentration of this study is on 2 Dot 1 Electron QCA, an emerging variant of QCA. A novel and efficient XOR gate based on 2 Dot 1 Electron QCA is designed. Moreover a comparator using the proposed novel XOR gate is presented in this present scope. The proposed architecture is justified using a well-accepted standard mathematical function based on Coulomb's law. Energy and power dissipation of the architecture are analyzed using different energy parameters. AS the compactness of proposed design is 76.4% the design met high degree of compactness and better efficiency.**

Keywords*—QCA; 2 Dot 1 Electron QCA; Comparator; Coulomb's principle

I. Introduction

Expansion of cost-effective, efficient nanotechnologies is conducting owing to obtain the performance that is not attainable by CMOS technology due to its scaling limits like off-state leakage current, degrading switching activity, etc. Quantum-dot cellular automata (QCA) technology is one of the up-and-coming replacements to overcome the scaling limits of CMOS [1]-[3]. QCA provides efficient properties including high- speed nano-scale designs at terahertz frequency range, long lifetime small feature size, together with low power consumption, ultra-low power dissipation [4]-[6].

One of the recent variant of QCA is 2 Dot 1 Electron QCA which offer all the advantages of QCA technology over conventional CMOS technology together with benefits over the 4 Dot 2 Electron QCA structure. The most advantageous aspect of 2 Dot 1 Electron QCA over 4 Dot 2 Electron QCA is the total number electron in the cell is halved, so energy requirement is reduced. There exits four ambiguous configurations in 4 Dot 2 Electron QCA, only two of them are valid. However, there is no ambiguous configuration in case of 2 Dot 1 Electron QCA [7]-[9]. Moreover in 2 Dot 1 Electron QCA wiring complexity is minimized as binary information can be transmitted from one cell to another using cell-to-cell interaction and obeying Coulomb's repulsion principle.

Comparator is fundamental digital devices and essential component for modern computing environments. There are diverse form of existing literature on comparator design using 4 Dot 2 Electron QCA as in [10]–[15]. Comparator design using 2 Dot 1 Electron QCA is unexplored till now. This article proposes an optimized comparator implemented using 2 Dot 1 Electron QCA cell.

The later part of the study is structured in the following way. Section II represents the overview of 2 Dot 1 Electron QCA. Section III offers design and scheme of a novel XOR gate. The design of proposed 1 bit comparators using the novel XOR gate is presented in section IV. Section V verifies the output energy states using standard mathematical procedure. Effective area analysis of the proposed architecture is provided in section VI. In section VII energy and power dissipation to run the proposed design is briefly analyzed. Finally Conclusion is demonstrated in section VIII.

II. Overview of 2 Dot 1 Electron QCA

The building block of Quantum dot Cellular Automata are cells and quantum dots or holes. Different formations of cells lead to different variants of QCA. Two dimensional 2 Dot 1 Electron QCA is the structural variation of QCA. As the name clearly suggests that a 2-Dot 1-Electron QCA cell consists of two logically interacting quantum dots and one electron. Single electron can move between the two quantum dots through the tunnel. Also, the single electron in the cell is able to represent binary information by the occupancy of electron in the quantum dots.

The cell structure of 2 Dot 1 Electron QCA is rectangular, either vertically or horizontally oriented along with two dots are at the two ends [9]. The structure of the 2-dot 1-electron QCA cells and their polarities are shown in Fig. 1(a) and 1(b). The position of the electron within a cell represents binary information.

In case of vertical cell when the electron positioned in the quantum dot below, and above, represent binary '0' and '1' respectively shown in Fig. 1(a) and 1(b). In the same way, in horizontal cell when the electron positioned in the right and left quantum dot represent binary '0' and '1' respectively.

However, like 4 Dot 2 Electron QCA 2 Dot 1 Electron QCA has fundamental design constructs such as binary wire, inverter by oppositely positioned cell between two same oriented cell of a binary wire, planar crossing of wires [7],

inverter by placing cell at corner and majority voter gate as depicted in Fig. 1 (c), 1 (d), 1(e), 1(f) and 2 respectively.

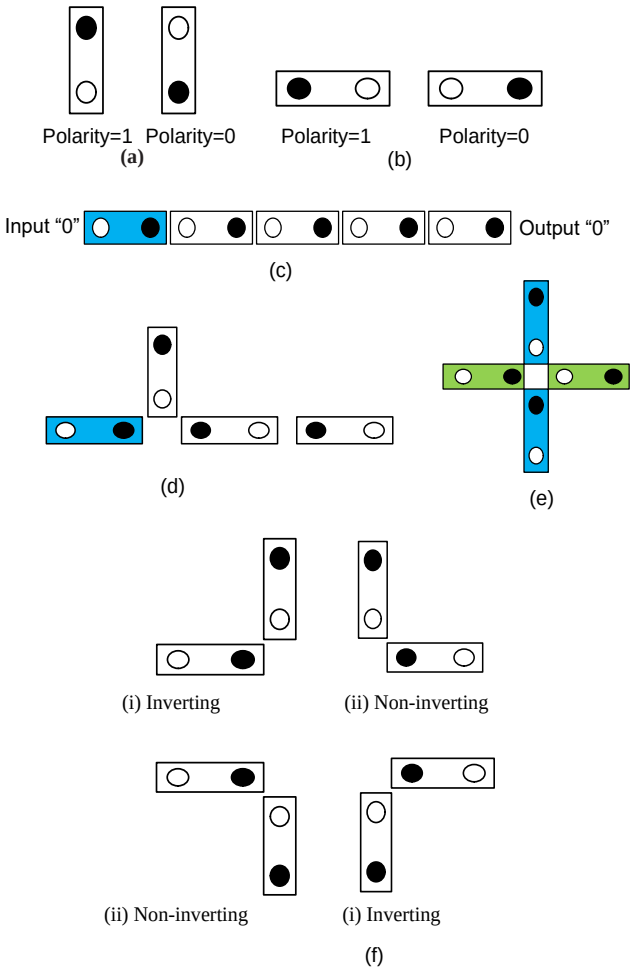


Fig. 1. The 2 Dot 1 Electron based(a) vertically aligned cells (b) horizontally aligned cells (c) binary wire (d) inverter (e) planar crossing of wires

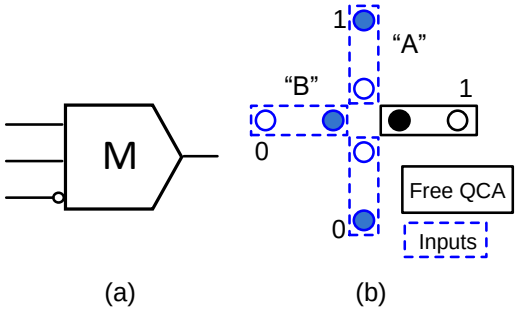


Fig. 2. Majority Voter (MV) gate (a) logical diagram and (b) 2 Dot 1 Electron QCA implementation

A. QCA Clocking

In CMOS technology clocking is used for the purpose of synchronization, whereas; in 2 Dot 1 Electron QCA architecture clocking fulfill two purposes. One is to control direction of data flow and other is to empower the weak input signals since they can traverse the whole circuit [8]. QCA clocking consists of four phases namely switch, hold, release, and relax.

One clock zone is out of phase with the subsequent clock zone by $\pi/2$ as depicted in Fig. 3(a) and several color codes indicate various clock zone as shown in Fig. 3(b).

In switch phase, initially electrons into the quantum dots contain minimum energy. The input signals are not adequate enough to empower the electrons. After that clock signal amplitude increases and potential energy of electrons begin to rise. Finally electron gains highest potential energy at the end of this phase.

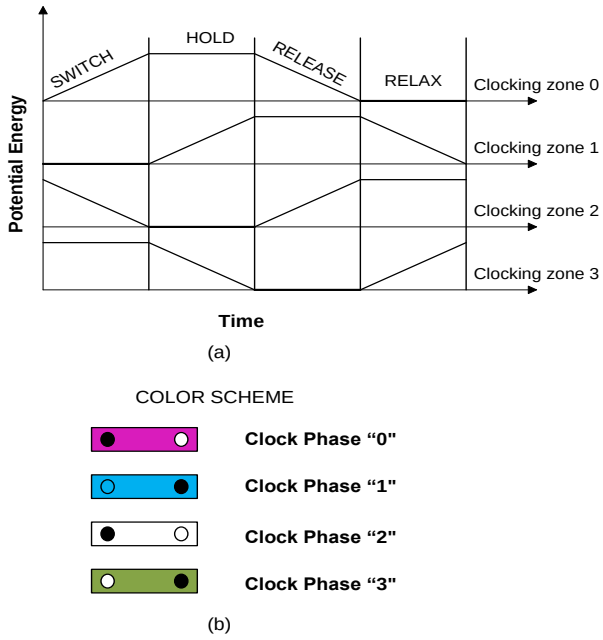


Fig. 3. (a) The 2 Dot 1 Electron clocking (b) color scheme of various clock zones

During the hold phase, the high phase, electrons become effectively energized to exceed the tunneling barrier. At this phase electrons loose polarity and get completely delocalized. The cells are stated to obtain null phase.

In the release phase, the high to low phase, actual computation is performed and electrons start to dissipate potential energy. They latch at the other quantum dots. The cells gradually begin to obtain a certain polarity.

In the relax phase, the low phase, electrons bear minimum energy and confined into the quantum dots. The cells in one clock zone will perform as input for the next clock zone.

III. The Novel Ex-Or Gate

A several form of XOR (exclusive-OR) gate has been designed until now [16], [17] and most of the cases utilized 3-input majority gate. In digital logic circuit designs XOR gate is an exigent element. Different XOR-based circuits have been designed yet, owing to the great essentiality of XOR gate [18]. Fig.4 illustrates the 2 Dot 1 Electron two input XOR Gate. However, the novel 2 Dot 1 Electron two input XOR Gate is not majority based rather it utilizes explicit interactions among QCA cells to determine the expected output.

IV. Proposed Design of 1 Bit Comparator

1 bit comparator as the name indicates that this digital logic circuit could be utilized to compare whether two 1 bit binary numbers are identical. Moreover it also identifies which number is larger. Therefore in logic circuits design 1 bit comparators are used for decision making. Fig.5 illustrates the schematic of 1-Bit Comparator using logic gates. Table I demonstrates the truth table of 1 Bit Comparator.

$$Y_{A>B} = A.\bar{B}$$
$$Y_{A<B} = \bar{A}.B$$
$$Y_{A=B} = \bar{A}.\bar{B}+A.B$$
$$= A \text{ XNOR } B$$

The proposed 1 bit Comparator has been implemented utilizing 2 Dot 1 Electron QCA. The design forms of two AND gates and one novel Ex-OR gate to determine the logic functionality of the 1 bit Comparator. In the proposed design there are two input A and B and three output A>B, A<B, A=B. the implementation of proposed 1 bit Comparator using 2 Dot 1 Electron QCA cell is depicted in Fig. 6.

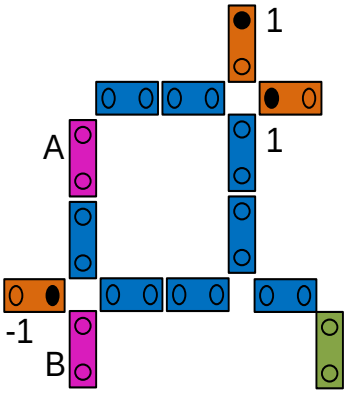


Fig. 4. (a) The 2 Dot 1 Electron two input XOR Gate

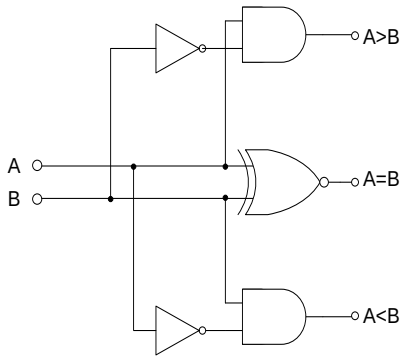


Fig. 5. Schematic of 1- Bit comparator using logic gate

TABLE. I. Truth Table of 1 Bit Comparator

Input		Output		
A	B	$Y_{A=B}$	$Y_{A<B}$	$Y_{A>B}$
0	0	1	0	0
0	1	0	1	0
1	0	0	0	1
1	1	1	0	0

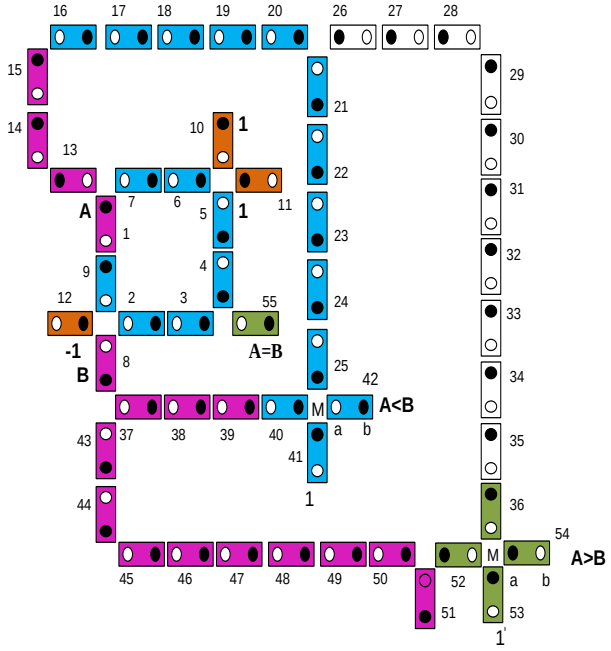


Fig. 6. 1 Bit comparator implemented using 2 Dot 1 Electron QCA cell

V. Estimation of Output Energy State of the Proposed Architecture

Since there is no available open source simulation software to simulate 2 Dot 1 Electron QCA circuit till now, like QCADesigner [19] for 4 Dot 2 Electron QCA architectures. Thus, this work utilizes a well-accepted standard mathematical function based on Coulomb's law to justify the proposed designs as available in [7] - [9], [20].

The calculation of potential energy between two electron charges is done using the following equations.

$$U = Kq_1q_2/r$$
$$Kq_1q_2 = 9 \times 10^{-9}(1.6)^2 \times 10^{-20}$$
$$U_T = \sum_{t=1}^{n} U_t \text{ .....(1)}$$

Where U indicates the potential energy between the two electron charges $q1$ and $q2$, K is the Boltzman constant and r denotes the distance between the two point electric charges. U_T denotes the total potential energy for a specific electron position because of the effects of all of its neighbor electrons. Electron and quantum dot contains negative charge and induced positive charge respectively. Electron, due to its characteristics always tends to reach at a position with least potential energy. Thus, to evaluate the output state of the proposed design, comparative analysis of the total potential energy is calculated for each of the allowable electron positions within a cell. Let, each of 2 Dot 1 Electron QCA cell has length = 13 nm and the space between two cell = 5 nm. Fig.6 represents the numbering of cell of the proposed comparator design using 2 Dot 1 electrons QCA and Table II presents respective potential energy evaluations.

VI. Effective Area Analysis

As mention before the structure of 2 Dot 1 electron QCA cells are rectangular. Consider the length and breadth of a 2 Dot 1 electron QCA cell is p and q respectively. Therefore area

of each cell is p×q nm^2. To form the proposed 1 bit Comparator 55 number of 2 Dot 1 electron QCA cell is required as shown in Fig. 6 Thus the effective area under the architecture is 55pq and the area covered by the architecture is 72pq. The ratio of area utilization of the proposed architecture is 55:72. The compactness is 76.4%. Hence the design met high degree of compactness.

VII. Energy and Power Dissipation Analysis

Various expressions are used as in [22], [23] to calculate energy parameters. E_m is the minimum energy to be provided to the scheme with *N* cells; E_{clock} is the energy applied by the clock to the circuit with *N* cells; E_{diss} is dissipation of energy from the circuit with *N* cells; $\boldsymbol{v}_2$ denotes frequency of energy dissipation; τ_2 is the time to dissipate to arrive to the relaxed state into the environment; $\boldsymbol{v}_1$ is the frequency of incident energy; τ_1 is the time needed to arrive from quantum level n_2 to the quantum level n. τ is the required time by the cells to switch from one to the subsequence polarization in a particular clock zone; t_p is propagation time to the entire circuit; $\boldsymbol{v}_1$-$\boldsymbol{v}_2$ indicates differential frequency. All of these parameter values for proposed 1 bit comparator are determined in Table III.

TABLE. II. Output State of 2-Dot I-Electron QCA Comparator

Cell	Position of electron	Cumulative potential energy	Comments
1, 8			Input cell A and B respectively
10, 12			Cell 10, 11 has fixed polarity "1".
11			Cell 11 has fixed polarity "-1".
4 – 7			Attain polarity from cell 10, 11 and A.
9			Gains polarity from input cell A
2, 3			Gain polarity from 8, 9 and 12.
55			Attains polarity from 3, 4.
37 – 40			Gain same polarity of input cell B.
13 – 15			Gain same polarity of input cell A.
16 – 25			Attain the inverse polarity of cell 15 (Fig.1(f) (iv))
41			Cell 41 has fixed polarity "1".
53			Cell 53 has fixed polarity "1".
37 – 40			Gain same polarity of input cell B.
42	a b	13.564×10−20 J 1.368 × 10−20 J	Electron will latch at position b due to less energy
26 – 36			Attains the inverse polarity of cell 21 (Fig.1 (f) (iv))
43 – 51			Gain same polarity of input cell B.
52			Attains the inverse polarity of cell 51 (Fig.1 (f) (iv))
54	a b	−3.329 × 10−20 J −0.537 × 10−20 J	Electron will latch at position a due to less energy

TABLE. III. Several Energy Parameter Values for Proposed 1 Bit Comparator

Parameters	Values
$E_{clock} = \frac{n^2\pi^2\hbar^2 N}{ma^2}$	3.9207×10^{-18} Joules
$E_{diss} = \frac{\pi^2\hbar^2}{ma^2}(n^2-1)N$	3.8815×10^{-18} Joules
$v_1 = \frac{\pi\hbar}{2ma^2}(n^2-n_2^2)N$	2.8407×10^{15} Hz
$v_2 = \frac{\pi\hbar}{2ma^2}(n^2-1)N$	2.92945×10^{15} Hz
$(v_1 - v_2) = \frac{\pi\hbar}{ma^2}(n^2-1)N$	8.875×10^{13} Hz
$\tau_1 = \frac{1}{v_1} = \frac{2ma^2}{\pi\hbar(n^2-n_2^2)N}$	3.5202×10^{-16} sec
$\tau_2 = \frac{1}{v_2} = \frac{2ma^2}{\pi\hbar(n^2-1)N}$	3.41361×10^{-16} sec
$\tau = \tau_1 + \tau_2$	6.93381×10^{-16} sec
$t_p = \tau + (k-1)\tau_2$	17.17464×10^{-16} sec

VIII. Conclusion

This present scope proposes an improved design methodology of 1 bit comparator using 2 Dot 1 Electron QCA which attains high degree of compactness and require lesser amount of energy to run. Since there is no open source simulation tool existing for 2 Dot 1 Electron. Therefore, potential energy calculation is applied in order to justify the proposed design. In addition effective area and stability is presented to analyze the acceptance level of the architecture. Several energy parameters and power dissipation are also analyzed for the proposed scheme.

References

[1] Lent, C. S., Tougaw, P. D., Porod, W., & Bernstein, G. H. (1993). Quantum cellular automata. *Nanotechnology*, *4*(1), 49.

[2] Liu, W., Lu, L., O'Neill, M., & Swartzlander, E. E. (2011, May). Design rules for quantum-dot cellular automata. In Circuits and Systems (ISCAS), 2011 IEEE International Symposium on (pp. 2361-2364). IEEE.

[3] Kim, K., Wu, K., & Karri, R. (2005, March). Towards designing robust QCA architectures in the presence of sneak noise paths. In Proceedings of the conference on Design, Automation and Test in Europe-Volume 2 (pp. 1214-1219). IEEE Computer Society.

[4] Bahar, A. N., Rahman, M. M., Nahid, N. M., & Hassan, M. K. (2017). Energy dissipation dataset for reversible logic gates in quantum dot-cellular automata. Data in Brief, 10, 557-560.

[5] Abdullah-Al-Shafi, M., & Bahar, A. N. (2016). Novel binary to gray code converters in QCA with power dissipation analysis. International Journal of Multimedia and Ubiquitous Engineering, 11(8), 379-396.

[6] Sheikhfaal, S., Angizi, S., Sarmadi, S., Moaiyeri, M. H., & Sayedsalehi, S. (2015). Designing efficient QCA logical circuits with power dissipation analysis. Microelectronics Journal, 46(6), 462-471.

[7] Ghosh, M., Mukhopadhyay, D., & Dutta, P. (2015). A 2 dot 1 electron quantum cellular automata based parallel memory. In Information systems design and intelligent applications (pp. 627-636). Springer India.

[8] Hook IV, L. R., & Lee, S. C. (2011). Design and simulation of 2-D 2-dot quantum-dot cellular automata logic. IEEE Transactions on Nanotechnology, 10(5), 996-1003.

[9] Ghosh, M., Mukhopadhyay, D., & Dutta, P. (2016). 2-Dimensional 2-Dot 1-Electron Quantum Cellular Automata-Based Dynamic Memory Design. In Proceedings of the 4th International Conference on Frontiers in Intelligent Computing: Theory and Applications (FICTA) 2015 (pp. 357-365). Springer India.

[10] Mukhopadhyay, D., Dinda, S., & Dutta, P. (2011). Designing and implementation of quantum cellular automata 2: 1 multiplexer circuit. International Journal of Computer Applications, 25(1), 21-24.

[11] Ke-ming, Q., & Yin-shui, X. (2007, October). Quantum-dots cellular automata comparator. In ASIC, 2007. ASICON'07. 7th International Conference on (pp. 1297-1300). IEEE.

[12] Lampreht, B., Stepancic, L., Vizec, I., Zankar, B., Mraz, M., Bajec, I. L., & Pecar, P. (2008, September). Quantum-dot cellular automata serial comparator. In Digital System Design Architectures, Methods and Tools, 2008. DSD'08. 11th EUROMICRO Conference on (pp. 447-452). IEEE.

[13] Wagh, M. D., Sun, Y., & Annampedu, V. (2008, June). Implementation of comparison function using quantum-dot cellular automata. In Proc. Nanotechnol. Conf. Trade Show (pp. 76-79).

[14] Xia, Y., & Qiu, K. (2009). Comparator design based on quantum-dot cellular automata. J. Electron. Inf. Technol, 31(6), 1517-1520.

[15] Abdullah-Al-Shafi, M., & Bahar, A. N. (2016). Optimized design and performance analysis of novel comparator and full adder in nanoscale. Cogent Engineering, 3(1), 1237864.

[16] Bahar, A. N., Uddin, M. S., Abdullah-Al-Shafi, M., Bhuiyan, M. M. R., & Ahmed, K. (2017). Designing efficient QCA even parity generator circuits with power dissipation analysis. Alexandria Engineering Journal.

[17] Bahar, A. N., Waheed, S., Hossain, N., & Asaduzzaman, M. (2017). A novel 3-input XOR function implementation in quantum dot-cellular automata with energy dissipation analysis. Alexandria Engineering Journal.

[18] Bahar, A. N., Waheed, S., & Hossain, N. (2015). A new approach of presenting reversible logic gate in nanoscale. SpringerPlus, 4(1), 153.

[19] Walus, K., Dysart, T. J., Jullien, G. A., & Budiman, R. A. (2004). QCADesigner: A rapid design and simulation tool for quantum-dot cellular automata. IEEE transactions on nanotechnology, 3(1), 26-31.

[20] Bahar, A. N., & Waheed, S. (2016). Design and implementation of an efficient single layer five input majority voter gate in quantum-dot cellular automata. SpringerPlus, 5(1), 636.

[21] Hashemi, S., Farazkish, R., & Navi, K. (2013). New quantum dot cellular automata cell arrangements. Journal of Computational and Theoretical Nanoscience, 10(4), 798-809.

[22] Mukhopadhyay, D., & Dutta, P. (2015). A study on energy optimized 4 dot 2 electron two dimensional quantum dot cellular automata logical reversible flip-flops. Microelectronics Journal, 46(6), 519-530.

[23] Angizi, S., Alkaldy, E., Bagherzadeh, N., & Navi, K. (2014). Novel robust single layer wire crossing approach for exclusive or sum of products logic design with quantum-dot cellular automata. Journal of Low Power Electronics, 10(2), 259-271.